Discrete Choice Models

Li Gan

September 2010

Many decision variables are discrete –enrolling in graduate school, getting married, having a child, etc. In the case, the observed dependent variable takes two values. For example, the observed y takes 0 or 1 values, which is a binary response model. Define:

$p(x)=\Pr(y=1|x)$, and $\Pr(y=0|x) = 1- p(x)$. Then we have:

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

1. ***Ad hoc* specifications:**

An *ad hoc* way to specify the model is to assume some functional form of $p(x)$.

Given that:

$E(y|x) = p(x) * 1 + (1-p(x))*0 = p(x)$,

It is natural to setup a regression model:

$y = p(x) + u$

There are several ways to specify p(x) function:

(1) Linear probability model for Binary response: $E(y|x) = p(x) = x\beta$.

$$\text{var}(y|x) = x\beta(1- x\beta)$$

Consider a simple regression on the model (Linear Probability Model, or LPM):

$$y = x\beta+u.$$

$\hat{\beta}_{OLS}$ is consistent but the error term u must be heteroscedastic. Therefore, the heteroscedastic-robust standard error has to be used.

$$Var\left(\hat{\beta}_{OLS}\right) = (x'x)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2 x_i' x_i\right)(x'x)^{-1}$$

Can we do GLS?

$$\hat{\Omega} = x\hat{\beta}_{OLS}\left(1 - x\hat{\beta}_{OLS}\right) \text{ or } \hat{\sigma}_i^2 = x_i\hat{\beta}_{OLS}\left(1 - x_i\hat{\beta}_{OLS}\right).$$

Weighted least square:

$$y_i / \hat{\sigma}_i = (x_i / \hat{\sigma}_i)\hat{\beta}_{OLS} + u_i / \hat{\sigma}_i$$

Or one can simply use White's robust standard error.

This all seems too simple. What is the problem? Note nothing guarantees that the predicted value $0 \le \hat{y}_i \le 1$. This creates the obvious problem.

Example: married woman labor force participation (mroz.raw).

Of the 753 women in the sample, 428 report working.

Using the LPM model: of 753 fitted probabilities, 33 are outside the unit interval. When $n \to \infty$, this should not happen often.

(2) Assume $p(x) = \Phi(x\beta)$. This suggests a nonlinear regression model:

$$\underset{\beta}{Min} \quad \sum_{i=1}^{N} (y_i - \Phi(x_i\beta))^2 \qquad (1)$$

This model would result in a consistent estimator of β. Note that this model is different from the usual Probit model in terms of its MLE estimation method. The usual MLE method would produce more efficient estimates than this with similar amount of assumptions.

One potential application of this method, however, is on aggregate data. For example, if we are interested in choice of different types of vehicles but we do not observe household-level data. Only the aggregate shares of different types of vehicles at city or state levels are available. In this case, y_{ij} is the share of jth-type of vehicles for ith state or city, x_{ij} would be the average characteristics of jth-type of vehicle for the ith state or city. This method has been used in Fullerton, Gan and Hattori (2008).

2. **The latent variable model:**

Consider the latent variable model.

$y^* = x\beta + e$; The observed y: $y = 1$ if $y^* > 0$, and $y = 0$ otherwise.

(1) Probit model, which assumes $p(x) = \Phi(x\beta)$.

(2) Logit model, which assumes:

$$p(x) = \frac{\exp(x\beta)}{1+\exp(x\beta)}$$

The density for y_i is given by:

$$f(y_i \mid x_i) = p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$$

The mode is estimated by maximum likelihood.

3. **McFadden's Random Utility Maximization model.**

A different way to motivate the problem is in terms of utility comparison:

$$\begin{cases} V_0 = x_0\beta_0 + \varepsilon_0 \\ V_1 = x_1\beta_1 + \varepsilon_1 \end{cases}$$

where V_0 indicates utility from choosing 0, and V_1 indicates utility from choosing 1. McFadden's Random Utility Maximization theory assumes that a choice is chosen if and only if that the choice yields the largest utility among all choices.

Choosing 1 over 0 if and only if $V_1 \geq V_0$, which leads to:

$$x_1\beta_1 + \varepsilon_1 \geq x_0\beta_0 + \varepsilon_0 \qquad (2)$$

Suppose ε_0 and ε_1 have densities f(ε_0) and f(ε_1) and are independent:

$$\begin{aligned} &\Pr(choice\ 1) = \Pr(V_{i1} > V_{i0}) \\ &= \Pr(x_{i1}\beta_1 + \varepsilon_{i1} > x_{i0}\beta_0 + \varepsilon_{i0}) \\ &= \Pr(\varepsilon_{i0} < x_{i1}\beta_1 - x_{i0}\beta_0 + \varepsilon_{i1}) \\ &= \int_{-\infty}^{\infty} f(\varepsilon_{i1})\left(\int_{-\infty}^{x_{i1}\beta_1 - x_{i0}\beta + \varepsilon_{i1}} f(\varepsilon_{i0})d\varepsilon_{i2}\right)d\varepsilon_{i1} \\ &= \int_{-\infty}^{\infty} f(\varepsilon_{i1})F(x_{i1}\beta_1 - x_{i0}\beta_0 + \varepsilon_{i1})d\varepsilon_{i1} \end{aligned}$$

If ε_0 and ε_1 have standard normal distributions and are independent, then:

$$\begin{aligned} &\Pr(choice\ 1) = \Pr(V_{i1} > V_{i0}) \\ &= \int_{-\infty}^{\infty} \phi(\varepsilon_{i1})\Phi(x_{i1}\beta_1 - x_{i0}\beta_0 + \varepsilon_{i1})d\varepsilon_{i1} \end{aligned} \qquad (3)$$

Alternatively, one can rewrite this inequality (3):

$$\varepsilon_0 - \varepsilon_1 \leq x_1\beta_1 - x_0\beta_0$$

Define $\varepsilon_0 - \varepsilon_1 = u$. If ε_0 and ε_1 have standard normal densities, then $u \sim N(0,2)$.

$$\begin{aligned} \Pr(choice\ 1) &= \Pr(\varepsilon_{i0} - \varepsilon_{i1} \leq x_{i1}\beta_1 - x_{i0}\beta_0) \\ &= \Pr(u \leq x_{i1}\beta_1 - x_{i0}\beta_0) \\ &= \Phi\left(\frac{x_{i1}\beta_1 - x_{i0}\beta_0}{\sqrt{2}}\right) \end{aligned} \tag{4}$$

The estimates from (3) and (4) should be similar.

However, we typically normalize u to be standard normal, in this case, we have a standard Probit model:

$$\Pr(choice\ 1) == \Phi(x_{i1}\beta_1 - x_{i0}\beta_0)$$

Further, if u (the difference between ε_1 and ε_0) has a density:

$$f(u) = \frac{e^{-u}}{(1+e^{-u})^2}\text{, and } F(u) = \frac{1}{1+e^{-u}}$$

Then:

$$\Pr(u > -x\beta) = 1 - F(-x\beta) = \frac{1}{1+\exp(x\beta)}$$

If ε_0 and ε_1 are extreme value distribution:

$F(x) = \Pr(e < x) = \exp(-e^{-x})$, and $f(x) = \exp(-x-e^{-x})$. Let $V_{ij} = \mu_{ij} + \varepsilon_{ij}$.

$$
\begin{aligned}
&\Pr(choice\ 1) = \Pr(V_{i1} > V_{i0}) \\
&= \Pr(\mu_{i1} + \varepsilon_{i1} > \mu_{i0} + \varepsilon_{i0}) \\
&= \Pr(\varepsilon_{i0} < \mu_{i1} - \mu_{i0} + \varepsilon_{i1}) \\
&= \int_{-\infty}^{\infty} f(\varepsilon_{i1}) \left(\int_{-\infty}^{\mu_{i1}-\mu_{i0}+\varepsilon_{i1}} f(\varepsilon_{i0}) d\varepsilon_{i2} \right) d\varepsilon_{i1} \\
&= \int_{-\infty}^{\infty} f(\varepsilon_{i1}) F(\mu_{i1} - \mu_{i0} + \varepsilon_{i1}) d\varepsilon_{i1} \\
&== \int_{-\infty}^{\infty} \exp\left(-\varepsilon_{i1} - e^{-\varepsilon_{i1}}\right) \exp\left(-e^{-(\mu_{i1}-\mu_{i0}+\varepsilon_{i1})}\right) d\varepsilon_{i1} \\
&= \int_{-\infty}^{\infty} \exp\left(-e^{-\varepsilon_{i1}}\right) \exp\left(-e^{-\varepsilon_{i1}} - e^{-(\mu_{i1}-\mu_{i0}+\varepsilon_{i1})}\right) de^{-\varepsilon_{i1}} \\
&= -\int_{-\infty}^{\infty} \exp\left(-e^{-\varepsilon_{i1}} - e^{-(\mu_{i1}-\mu_{i0})} e^{-\varepsilon_{i1}}\right) de^{-\varepsilon_{i1}} \\
&= -\int_{\infty}^{0} \exp\left(-x - e^{-(\mu_{i1}-\mu_{i0})} x\right) dx \\
&= \int_{0}^{\infty} \exp\left(-x\left(1 + e^{-(\mu_{i1}-\mu_{0})}\right)\right) dx \\
&= -\frac{1}{\left(1 + e^{-(\mu_{i1}-\mu_{0})}\right)} \exp\left(-x\left(1 + e^{-(\mu_{i1}-\mu_{0})}\right)\right)\Big|_{0}^{\infty} \\
&= \frac{1}{1 + e^{-(\mu_{i1}-\mu_{i0})}} \\
&= \frac{\exp(\mu_{i1})}{\exp(\mu_{i1}) + \exp(\mu_{i0})}
\end{aligned}
$$

The difference between the standard normal and the "McFadden-type" Probit is that coefficient differs by a scale.

Discussions:

(1) From previous discussions, it is important to notice that discrete choice models can only identify the difference $x_1\beta_1 - x_0\beta_0$ and parameters values up to a scale.

Example: choice between A&M and UT economics departments.

There are n students. Some would come to A&M ($y_i = 1$) while others would go to UT ($y_i = 0$). We observe student characteristics z_i, i = 1, 2, …, n, which does not change over choices, and school characteristics x_{ij}, j = 0 and 1.

$$\Pr(y_{ij} = 1) = \Phi(z_i\gamma + (x_{i1} - x_{i0})\beta)$$

Note the parameter γ measures the difference of parameters between UT and A&M. A positive γ indicates that z_i is relatively more important for A&M than for UT. $(x_{i1}- x_{i0})\beta$ obviously tell the effect of the difference in x's.

(2) In many cases, the latent variable model is simple and intuitive, and sufficient for analysis. However, when the dependent variable has more than two choices, it is difficult to construct the model with the latent variable setup. In this case, the McFadden random utility maximization model provides an easy and intuitive way to construct the model – the so-called multinomial logit or probit, and the nested logit model.

Estimation:

Previous models are often estimated by Maximum Likelihood. In general, suppose the density of y_i is given by:

$$f(y_i \mid x_i) = G(x_i\beta)^{y_i}(1 - G(x_i\beta))^{1-y_i}$$

The likelihood function is given by:

$$l(\beta) = y_i \log(G(x_i\beta)) + (1 - y_i)\log(1 - G(x_i\beta))$$

The score function (first order condition) is:

$$S_i(\beta) = \frac{g(x_i\beta)x_i'(y_i - G(x_i\beta))}{G(x_i\beta)(1 - G(x_i\beta))}$$

The Hessian is:

$$-E[H_i(\beta \mid x_i)] = \frac{g(x_i\beta)^2 x_i' x_i}{G(x_i\beta)(1 - G(x_i\beta))} = A(x_i\beta)$$

$$A\operatorname{var}(\hat{\beta}) = \sum_{i=1}^{n} \frac{g(x_i\hat{\beta})^2 x_i' x_i}{G(x_i\hat{\beta})(1 - G(x_i\hat{\beta}))} = \hat{V}$$

Testing

$\Pr(y=1|x,z) = G(x\beta + z\gamma)$

H_0: $\gamma = 0$ Likelihood ratio test

Comparison of the three models (LPM, Logit and Probit):

The biggest difference between the LPM and the Logit or Probit is that LPM assumes the constant marginal effect, while Logit and Probit do not make that assumption.

For example, one or more small children reduces labor force participation by about 0.262, regardless of how many young children the woman already has (and the regardless of the levels of other variables). We can contrast this by finding with the estimated marginal effect from probit. For concreteness, take a woman with nwifeinc = 20.13, educ = 12,3, exper = 10,6, age = 42.5, roughly the sample averages – and kidsage6 =1. To calculate the effect of having one kid on the labor force participation, $\Phi(\bar{x}\beta) - \Phi(\bar{x}\beta - .868) = -.334$. If kids from one to two, the additional effect on labor force participation is given by: $\Phi(\bar{x}\beta - .868) - \Phi(\bar{x}\beta - 2 * .868) = -.256$.

Table 15.1
LPM, Logit, and Probit Estimates of Labor Force Participation

Dependent Variable: *inlf*			
Independent Variable	LPM (OLS)	Logit (MLE)	Probit (MLE)
nwifeinc	−.0034 (.0015)	−.021 (.008)	−.012 (.005)
educ	.038 (.007)	.221 (.043)	.131 (.025)
exper	.039 (.006)	.206 (.032)	.123 (.019)
$exper^2$	−.00060 (.00019)	−.0032 (.0010)	−.0019 (.0006)
age	−.016 (.002)	−.088 (.015)	−.053 (.008)
kidslt6	−.262 (.032)	−1.443 (0.204)	−.868 (.119)
kidsge6	.013 (.013)	.060 (.075)	.036 (.043)
constant	.586 (.151)	.425 (.860)	.270 (.509)
Number of observations	753	753	753
Percent correctly predicted	73.4	73.6	73.4
Log-likelihood value	—	−401.77	−401.30
Pseudo *R*-squared	.264	.220	.221

Multinomial Probit and Logit

Consider the McFadden's random utility maximization model,

$$V_{ij} = \mu_{ij} + \varepsilon_{ij}. \quad \Pr(j \text{ is chosen}) = \Pr(V_{ij} = \max(V_{i1}, V_{i2}, \ldots, V_{ik}))$$

Suppose three choices: 1, 2, 3.

$$\Pr(choice\ 1) = \Pr(V_{i1} > V_{i2}, V_{i1} > V_{i3})$$
$$= \Pr(\mu_{i1} + \varepsilon_{i1} > \mu_{i2} + \varepsilon_{i2}, \mu_{i1} + \varepsilon_{i1} > \mu_{i3} + \varepsilon_{i3})$$
$$= \Pr(\varepsilon_{i2} < \mu_{i1} - \mu_{i2} + \varepsilon_{i1}, \varepsilon_{i3} < \mu_{i1} - \mu_{i3} + \varepsilon_{i1})$$
$$= \int_{-\infty}^{\infty} f(\varepsilon_{i1}) \left(\int_{-\infty}^{\mu_{i1}-\mu_{i1}+\varepsilon_{i1}} f(\varepsilon_{i2}) d\varepsilon_{i2} \int_{-\infty}^{\mu_{i3}-\mu_{i1}+\varepsilon_{i1}} f(\varepsilon_{i3}) d\varepsilon_{i3} \right) d\varepsilon_{i1}$$
$$= \int_{-\infty}^{\infty} f(\varepsilon_{i1}) (F(\mu_{i1} - \mu_{i2} + \varepsilon_{i1}) F(\mu_{i1} - \mu_{i3} + \varepsilon_{i1})) d\varepsilon_{i1}$$

If ε_{ij} is normal,

$$\Pr(choice\ 1) = \int_{-\infty}^{\infty} \phi(\varepsilon_{i1}) \Phi(\mu_{i1} - \mu_{i2} + \varepsilon_{i1}) \Phi(\mu_{i1} - \mu_{i3} + \varepsilon_{i1}) d\varepsilon_{i1}$$

If ε_{ij} is Type I extreme value distribution:

$$f(\varepsilon_j) = \exp(-\varepsilon_j) \exp(-\exp(-\varepsilon_j))$$
$$F(\varepsilon_j) = \exp(-\exp(-\varepsilon_j))$$

Then:

$$\Pr(choice\ 1)$$
$$= \int_{-\infty}^{\infty} \exp(-\varepsilon_{i1} - e^{-\varepsilon_{i1}}) e^{-\exp(\mu_{i1}-\mu_{i2}+\varepsilon_{i1})} e^{-\exp(\mu_{i1}-\mu_{i3}+\varepsilon_{i1})} d\varepsilon_{i1}$$
$$= \int_{-\infty}^{\infty} \exp(-e^{-\varepsilon_{i1}}) e^{-\exp(\mu_{i1}-\mu_{i2}+\varepsilon_{i1})} e^{-\exp(\mu_{i1}-\mu_{i3}+\varepsilon_{i1})} de^{-\varepsilon_{i1}}$$
$$= \int_{-\infty}^{\infty} \exp(-e^{-\varepsilon_{i1}} - e^{-(\mu_{i1}-\mu_{i2})} e^{-\varepsilon_{i1}} - e^{-(\mu_{i1}-\mu_{i3})} e^{-\varepsilon_{i1}}) de^{-\varepsilon_{i1}}$$
$$= \int_{-\infty}^{\infty} \exp(-x(1 + e^{-(\mu_{i1}-\mu_{i2})} + e^{-(\mu_{i1}-\mu_{i3})})) dx$$
$$= \frac{1}{1 + e^{-(\mu_{i1}-\mu_{i2})} + e^{-(\mu_{i1}-\mu_{i3})}}$$
$$= \frac{\exp(\mu_{i1})}{\exp(\mu_{i1}) + \exp(\mu_{i2}) + \exp(\mu_{i3})}$$

Independent of Irrelevant Alternatives (IIA):

First, three choices: Car, Blue Bus, Subway

Multinomial logit:

Suppose: $$\frac{\Pr(Car)}{\Pr(Blue\ Bus)} = \frac{\exp(\mu_{Car})}{\exp(\mu_{BlueBus})} = 1 \text{ and} \tag{5}$$

$$\frac{\Pr(Car)}{\Pr(Subway)} = \frac{\exp(\mu_{Car})}{\exp(\mu_{Subway})} = 1 \tag{6}$$

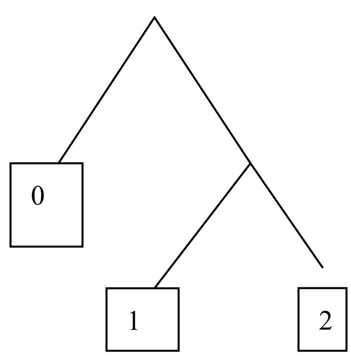
This is called Independent of Irrelevant Alternatives (IIA). Previous example implies Pr(Car) = Pr(Subway) = Pr(Blue Bus) = 1/3.

Now we introduce another choice: Red Bus. Within the multinomial logit framework, we should still have the relationship as in (5) and (6), which implies that:

Pr(Car) = Pr(Subway) = Pr(Blue Bus) = Pr(Red Bus) = ¼,

This clearly is wrong.

The IIA assumption is clearly wrong. What is wrong is the assumption of independence. In the example, the errors of Blue Bus and Red Bus are clearly not independent.



Nested Logit Model

$$f(\varepsilon_1, \varepsilon_2) = \exp\left(-\left(\exp\left(-\frac{\varepsilon_1}{\rho}\right) + \exp\left(-\frac{\varepsilon_2}{\rho}\right)\right)^{\rho}\right)$$

This is Gumblel's Type B bivariate extreme-value distribution. The correlation coefficient can be shown to be $(1-\rho^2)$.

$$
\begin{aligned}
&\Pr(y=0)=\Pr(V_0>V_1,V_0>V_2)\\
&=\Pr(\mu_0+\varepsilon_0>\mu_1+\varepsilon_1,\mu_0+\varepsilon_0>\mu_2+\varepsilon_2)\\
&=\frac{\exp(\mu_0)}{\exp(\mu_0)+\left(\exp\left(\frac{\mu_1}{\rho}\right)+\exp\left(\frac{\mu_2}{\rho}\right)\right)^{\rho}}
\end{aligned}
$$

$$
\begin{aligned}
&\Pr(y=1)=\Pr(V_1>V_0,V_1>V_2)\\
&=\Pr(y=1\,\&\,2)\cdot\Pr(y=1\mid y=1\,\&\,2)\\
&=\frac{\left(\exp\left(\frac{\mu_1}{\rho}\right)+\exp\left(\frac{\mu_2}{\rho}\right)\right)^{\rho}}{\exp(\mu_0)+\left(\exp\left(\frac{\mu_1}{\rho}\right)+\exp\left(\frac{\mu_2}{\rho}\right)\right)^{\rho}}\bullet\frac{\exp\left(\frac{\mu_1}{\rho}\right)}{\exp\left(\frac{\mu_1}{\rho}\right)+\exp\left(\frac{\mu_2}{\rho}\right)}
\end{aligned}
$$

Although the previous equation looks complicated, it is easy to follow – probability of a choice can be written into the probability of the nest that the choice belongs, multiplied by the probability of the choice within this nest.

Three-level nested logit:

$$
\begin{aligned}
&F(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n)\\
&=\exp\left(-\sum_u b_u\left\{\sum_{s\in C_u}a_s\left[\sum_{j\in B_s}\exp\left(-\frac{\varepsilon_j}{\rho_s}\right)\right]^{\rho_s/\sigma_n}\right\}^{\sigma_u}\right)
\end{aligned}
$$

$$
\Pr(y=j\mid s)=\frac{\exp(\mu_j/\rho_s)}{\sum_{k\in B_S}\exp(\mu_k/\rho_s)}\qquad\text{The third level}
$$

$$
\sum_{j\in B_s}\Pr(y=j\mid u)=\frac{a_s\left[\sum_{j\in B_s}\exp(\mu_j/\rho_s)\right]^{\rho_s/\sigma_u}}{\sum_{\tau\in C_u}a_\tau\left[\sum_{j\in B_\tau}\exp(\mu_j/\rho_\tau)\right]^{\rho_\tau/\sigma_u}}\qquad\text{The second level}
$$

$$\sum_{j \in C_u} \Pr(y = j) = \frac{b_u \sum_{\tau \in C_u} a_\tau \left[\sum_{j \in B_\tau} \exp(\mu_j / \rho_\tau) \right]^{\rho_\tau / \sigma_u}}{\sum_w b_w \left\{ \sum_{\tau \in C_w} a_\tau \left[\sum_{j \in B_\tau} \exp(\mu_j / \rho_\tau) \right]^{\rho_\tau / \sigma_u} \right\}^{\sigma_w}} \qquad \text{The third level.}$$

Generalized extreme-value model

$$F(\varepsilon_1, \varepsilon_2, ..., \varepsilon_m) \\ = \exp(-G((\exp(-\varepsilon_1), (\exp(-\varepsilon_2), ...(\exp(-\varepsilon_m)))$$

where G satisfying the condition

(i) $G(\mu_1,\dots, \mu_m) \geq 0 \quad \mu_1,\dots, \mu_m \geq 0$

(ii) $G(\alpha\mu_1,\dots, \alpha\mu_m) = \alpha G(\mu_1,\dots, \mu_m)$

(iii) $\dfrac{\partial^k G}{\partial\mu_1 \partial\mu_2 ... \partial\mu_k} \begin{cases} \geq 0 & \text{if } k \text{ is odd} \\ \leq 0 & \text{if } k \text{ is even} \end{cases}$

If $v_j = \mu_j + \varepsilon_j$, then:

$$P_j = \frac{\exp(\mu_j) G_j (\exp(\mu_1), \exp(\mu_2)..., \exp(\mu_m))}{G(\exp(\mu_1), \exp(\mu_2)..., \exp(\mu_m))}$$

Example:

Consider a family who choosing vehicles. One way to consider the nesting structure is as follows, they make decisions about whether they are going to have a one-vehicle bundle, a two-vehicle bundle, and no-vehicle bundle. For the one vehicle bundle, they decide to go for a SUV or a car; for the two-vehicle bundle, they decide among (Car, Car), (Car, SUV), and (SUV, SUV) bundles.

Let the discrete choice variable be y_i. It is easy to write out the likelihood function. Given the following graph:

$$\Pr(y_i = 1) = \Pr(y_i = 1 \mid Nest\ 1)\Pr(Nest\ 1)$$

$$= \frac{\exp(V_1/\rho_1)}{\exp(V_{Nest\,1})} \cdot \frac{\exp(V_{Nest\,1})}{\exp(V_{Nest\,1}) + \exp(V_{Nest\,2}) + \exp(V_{Nest\,no})}$$

$$= \frac{\exp\left(\frac{V_1}{\rho_1}\right)}{\exp\left(\frac{V_1}{\rho_1}\right) + \exp\left(\frac{V_2}{\rho_1}\right)} \cdot \frac{\left(\exp\left(\frac{V_1}{\rho_1}\right) + \exp\left(\frac{V_2}{\rho_1}\right)\right)^{\rho_1}}{\left(\exp\left(\frac{V_1}{\rho_1}\right) + \exp\left(\frac{V_2}{\rho_1}\right)\right)^{\rho_1} + \left(\exp\left(\frac{V_3}{\rho_2}\right) + \exp\left(\frac{V_4}{\rho_2}\right) + \exp\left(\frac{V_6}{\rho_2}\right)\right)^{\rho_2} + \exp(V_6)}$$

Similarly, for family *i*,

$$\Pr(y_i = 3) = \Pr(y_i = 3 \mid Nest\ 2)\Pr(Nest\ 2)$$

$$= \frac{\exp(V_3/\rho_2)}{\exp(V_{Nest\,2})} \cdot \frac{\exp(V_{Nest\,2})}{\exp(V_{Nest\,1}) + \exp(V_{Nest\,2}) + \exp(V_{Nest\,no})}$$

$$= \frac{\exp\left(\frac{V_3}{\rho_2}\right)}{\exp\left(\frac{V_3}{\rho_2}\right) + \exp\left(\frac{V_4}{\rho_2}\right) + \exp\left(\frac{V_6}{\rho_2}\right)} \cdot \frac{\left(\exp\left(\frac{V_3}{\rho_2}\right) + \exp\left(\frac{V_4}{\rho_2}\right) + \exp\left(\frac{V_6}{\rho_2}\right)\right)^{\rho_2}}{\left(\exp\left(\frac{V_1}{\rho_1}\right) + \exp\left(\frac{V_2}{\rho_1}\right)\right)^{\rho_1} + \left(\exp\left(\frac{V_3}{\rho_2}\right) + \exp\left(\frac{V_4}{\rho_2}\right) + \exp\left(\frac{V_6}{\rho_2}\right)\right)^{\rho_2} + \exp(V_6)}$$

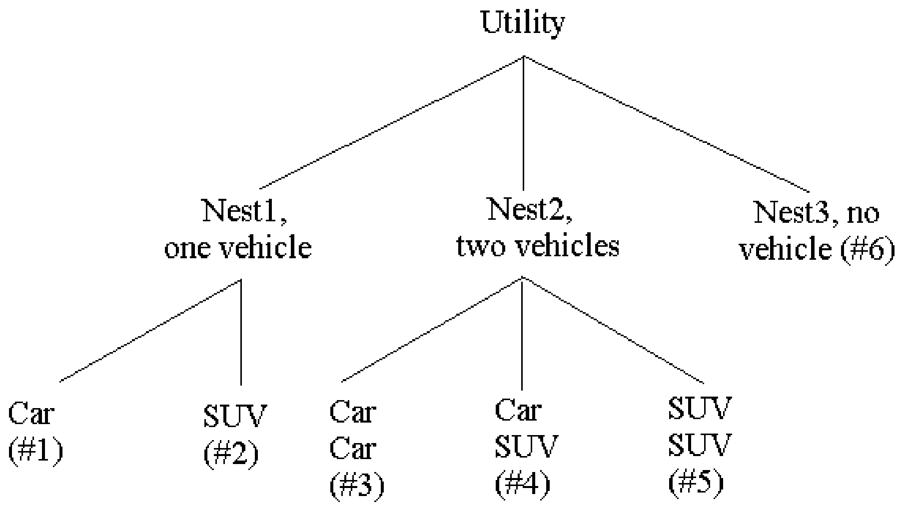
and,

$$\Pr(y_i = 6) = \frac{\exp(V_6)}{\left(\exp\left(\frac{V_1}{\rho_1}\right) + \exp\left(\frac{V_2}{\rho_1}\right)\right)^{\rho_1} + \left(\exp\left(\frac{V_3}{\rho_2}\right) + \exp\left(\frac{V_4}{\rho_2}\right) + \exp\left(\frac{V_6}{\rho_2}\right)\right)^{\rho_2} + \exp(V_6)}$$

Note $V_j = X_j\beta + Z\gamma_j$, $j = 1, \ldots, 6$.

The likelihood function:

$$l = \sum_{i=1}^{N} \sum_{j=1}^{6} (y_i = j) \ln \Pr(y_i = j)$$



<u>Ordered response</u>

Example: satisfaction survey. Consider a survey on job satisfaction with an answer of 1 (indicates not satisfactory), 2 is somewhat satisfied, and 3 is very satisfied.

$$
\begin{array}{ll}
y^* = x\beta + e & \text{with } e|x \sim \mathrm{N}(0,1) \\
y = 0 & \text{if } y^* \leq a_1 \\
y = 1 & \text{if } a_1 < y^* \leq a_2 \\
\ldots & \\
y = J & \text{if } a_J < y^*
\end{array}
$$

The density for y is given by:

$\Pr(y=0|x) = \Pr(y^* \leq a_1|x) = \Phi(a_1 - x\beta)$

$\Pr(y=1|x) = \Pr(a_1 < y^* \leq a_2|x) = \Phi(a_2 - x\beta) - \Phi(a_1 - x\beta)$

STATA Deviation:

(1) If error term is assumed to be normal $\mathrm{N}(0, \sigma^2)$, it is a ordered probit model:
STATA: `oprobit y x`

(2) If error term is assumed to be logistic, it is a ordered probit model:
STATA: `ologit y x`

Specification issue:

Consider a model: $\Pr(y = 1|x, c) = \Phi(x\beta+\gamma c)$, or, $y^* = x\beta+\gamma c$, and $y=1(y^*>0)$.

If $c \sim N(0,\tau^2)$, and independent of x, but c is unobserved. In linear model, this would not create any problems in consistency. However, here it will.

$$\Pr(y = 1 |x) = \Pr(\gamma c+e>- x\beta|x) = \Phi(x\beta/\sigma), \text{ where } \sigma^2 = 1+\gamma^2\tau^2$$

This is the so-called attenuation bias in estimating β.

However, partial effect $\dfrac{\partial \Pr(y=1 \mid x,c)}{\partial x_j} = \beta_j \phi(x\beta + \gamma c)$

Evaluate c at $E(c) = 0$, the partial effect is given by: $\beta_j\varphi(x\beta)$. Compare this with partial effect based on the consistently estimated coefficients, which is: $\beta_j/\sigma\varphi(x\beta/\sigma)$. It is not clear which value is larger or smaller. So the attenuation effect (for the partial effect model) does not necessarily exist. The bias still exists, though.

Endogeneity in Discrete Choice Models

***Example* 1**: A model with continuous endogenous explanatory variables:

Consider a model of female labor force participation, which depends on some exogenous variables z_1, and, more importantly, number of kids, denoted as *Numkids*. The problem is that the number of kids is often endogenous. How to deal with this problem? In general,

Consider a model:

$$\begin{aligned} y_1^* &= z_1\delta_1 + \alpha_1 Numkids + u_1 \\ Numkids &= z_1\delta_{21} + z_2\delta_{22} + v_2 = z\delta_2 + v_2 \qquad (7) \\ y_1 &= 1[y_1^* > 0] \end{aligned}$$

where (u_1, v_2) has zero means, bivariate normal and is independent of z. The key issue here is that y_2 is endogenous if u_1 and v_2 are correlated. If u_1 and v_2 are independent, there is no endogeneity problem. In our example, y_1^* governs if the person participates the labor market, and *Numkids* indicates the number of kids.

The instrument variables z_2 may include age of husband, age difference between husband and wife, a dummy variable indicating if the first two kids have the same sex, etc.

Rivers and Vuong (1998) two-stage procedure:

We let u_1 and v_2 be normally distributed. var(u_1) = 1. let:

$$u_1 = \theta_1 v_2 + e_1$$

If we let var(v_2) = τ_2^2. Assume e_1 to be independent of z and v_2, and therefore of *Numkids*. $E(e_1) = 0$, and $Cov(e_1, Numkids) = 0$.

➔ $y_1^* = z_1\delta_1 + \alpha_1 Numkids + \theta_1 v_2 + e_1$,

One way to test if *Numkids* are exogenous is as follows:

Step 1: run regression *Numkids* = $z\delta_2 + v_2$, obtain residual: $\hat{v}_2$.

Step 2: run probit model (if e_1 is normal):

$$y_1 = 1(z_1\delta_1 + \alpha_1 Numkids + \theta_1 \hat{v}_2 + e_1 > 0) \tag{8}$$

or, $\Pr(y_1 = 1) = \Phi(z_1\delta_1 + \alpha_1 Numkids + \theta_1 \hat{v}_2)$

H_0: $\theta_1 = 0$.

If H_0 is true, then *Numkids* is exogenous; otherwise, *Numkids* is endogenous.

A nice feature of the procedure is that the usual probit *t*-statistic on $\hat{v}_2$ is a valid test of the null hypothesis that y_2 is exogenous, that is, H_0: $\theta_1 = 0$. The test of exogeneity is valid without assuming normality or homoskedasticity of v_2, and it can be applied very broadly.

Discussions:

(1) Note one cannot use the previous two-step procedure to obtain parameter estimates. Note In equation (8), $Var(e_1) = 1 - \theta_1^2\tau_2^2 < 1$. So a simple Probit of (7) would not give us the correct coefficient even if v_2 is observed. Instead, the parameter estimates based on equation (8) are scaled by $(1 - \theta_1^2\tau_2^2)^{1/2}$.

Also note that when $\theta_1 \neq 0$, the usual Probit standard error and test statistics are not strictly valid, and we have only estimated δ_1 and θ_1 up to scale. The asymptotic variance of the two-step estimator can be derived using the M-estimator results.

(2) Consider the following two-step regression:

Step 1: A simple OLS of $Numkids = z\delta_2 + v_2$, obtain predicted $\hat{Numkids}$.

Step 2: Probit $y_1 = 1\left(z_1\delta_1 + \alpha_1 \hat{Numkids} + u_1 > 0 \right)$

This two-step procedure is wrong because:

$$E(E(y_1 \mid Numkids))$$
$$= E(\Phi(z_1\delta_1 + \alpha_1 Numkids)) \neq \Phi(z_1\delta_1 + \alpha_1 E(Numkids))$$

Note: $E(Numkids \mid z) = \hat{Numkids}$

Interestingly, at step 2, instead of using Probit, one can use the Linear Probability Model, which would yield consistent estimates of δ_1 and α_1.

Maximum Likelihood Estimation:

Alternatively, the model can be estimated using conditional MLE. We are interested in obtaining joint density:

$$f(y_1, Numkids \mid z) = f(y_1 \mid Numkids, z)\, f(Numkids \mid z).$$

First assume joint normality of (u_1, v_2), then e_1 is also normally distributed.

$E(e_1) = 0$, $\operatorname{var}(e_1) = 1 - \theta_1^2\tau_2^2$

➔ $y_1^* = z_1\delta_1 + \alpha_1 Numkids + \theta_1 v_2 + e_1$,

where $e_1 \mid z, Numkids, v_2 \sim N(0, 1 - \theta_1^2\tau_2^2)$. A standard calculation shows that:

$$\Pr(y_1 = 1 \mid z, Numkids, v_2) = \Pr(z_1\delta_1 + \alpha_1 Numkids + \theta_1 v_2 + e > 0 \mid z, Numkids, v_2)$$
$$= \Phi\left(\frac{z_1\delta_1 + \alpha_1 Numkids + \theta_1 v_2}{\sqrt{1 - \theta_1^2\tau_2^2}} \right)$$

Since we have:

$$\Pr(y_1 = 1 \mid z, y_2) = \Phi\left(\frac{z_1\delta_1 + \alpha_1 Numkids + \theta_1(Numkids - z\delta_2)}{\sqrt{1-\theta_1^2\tau_2^2}}\right)$$

Here we apply the fact that $v_2 = Numkids - z\delta_2$. Note the coefficients α_1, θ_1, and δ_2 are not separately identifiable if we only use y_1 part, or the $f(y_1|Numkids, z)$ part. However, the model is identifiable if we include the continuous part, $f(Numkids|z)$.

Therefore, the joint density is given by:

$$f(y_1, Numkids \mid z) = f(y_1 \mid Numkids, z) f(Numkids \mid z)$$

$$= \left\{\Phi\left(\frac{z_1\delta_1 + \alpha_1 Numkids + \theta_1(Numkids - z\delta_2)}{\sqrt{1-\theta_1^2\tau_2^2}}\right)\right\}^{y_1=1}$$

$$\left\{1-\Phi\left(\frac{z_1\delta_1 + \alpha_1 Numkids + \theta_1(Numkids - z\delta_2)}{\sqrt{1-\theta_1^2\tau_2^2}}\right)\right\}^{y_1=0} \frac{1}{\tau_2}\phi\left(\frac{Numkids - z\delta_2}{\tau_2}\right)$$

Advantages: (1) MLE is more efficient (2) direct estimate of δ_2 and α_1, and (3) testing that y_2 is exogenous is easy. Once the MLE has been obtained H_0: $\theta_1 = 0$.

Comparing the Rivers-Vuong approach to the MLE show that the former is a limited information procedure. Essentially, Rivers-Vuong focus on $f(y_1|Numkids,,z)$ where they replace the unknown δ_2 with OLS estimator $\hat{\delta}_2$. MLE uses the information in $f(y_1|Numkids, z)$ and $f(Numkids|z)$ simultaneously. For the initial test of whether *Numkids* is exogenous, the Rivers-- Vuoug approach has significant computation advantages.

***Example 2*: Ordered Probit with an endogenous but continuous explanatory variable**

Consider the problem of class rating. We are interested if instructor "good-looking" would affect his/her ratings. For simplicity, rating is given by r_i = 1(worst, 2, 3 (best)

Define a latent variable:

$$r_i^* = x_i\beta + \alpha g_i + u_i$$

The problem here is that g_i is measured with error, creating an endogeneity problem of g_i. There are two ways to solve this problem. First, one can obtain several measures of g_i, and then use the average g_i. Another method is to obtain a second measure of g_i, denoted as g_{2i}, and use the second measure as IV for the g_i.

Therefore, we have another model:

$$g_i = g_{2i}\gamma + v_i ,$$

and u_i and v_i are correlated:

$$u_i = \rho v_i + \varepsilon_i.$$

Therefore, the original latent variable model becomes:

$$\begin{aligned} r_i^* &= x_i\beta + \alpha g_i + \rho v_i + \varepsilon_i \\ &= x_i\beta + \alpha g_i + \rho(g_i - g_{2i}\gamma) + \varepsilon_i \end{aligned}$$

The likelihood is obtained by considering the joint density of g_i and r_i:

$$\begin{aligned} f(r_i, g_i) &= f(r_i \mid g_i) f_v(v_i) \\ &= \left(\prod_{j=1}^{3} \Pr(r_i = j)^{1(r_i=j)}\right) f_v(g_i - g_{2i}\gamma) \\ &= \left[\left(\Pr(r_i^* \le a_1)\right)^{1(r_i=1)} \left(\Pr(a_1 < r_i^* \le a_2)\right)^{1(r_i=2)} \left(\Pr(r_i^* > a_2)\right)^{1(r_i=3)}\right] f_v(g_i - g_{2i}\gamma) \end{aligned}$$

If $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and $v_i \sim N(0, \sigma_v^2)$, then:

$$\begin{aligned} \Pr(r_i = 1) &= \Pr(r_i^* \le a_1) = \Phi(a_1 - x_i\beta - \alpha g_i - \rho(g_i - g_{2i}\gamma)) \\ \Pr(r_i = 2) &= \Phi(a_2 - x_i\beta - \alpha g_i - \rho(g_i - g_{2i}\gamma)) - \Phi(a_1 - x_i\beta - \alpha g_i - \rho(g_i - g_{2i}\gamma)) \\ \Pr(r_i = 3) &= 1 - \Phi(a_2 - x_i\beta - \alpha g_i - \rho(g_i - g_{2i}\gamma)) \end{aligned}$$

Note that without the term $f_v(g_i - g_{2i}\gamma)$ in the likelihood, we cannot separately identify α, ρ and γ. Instead, we can only identify $\alpha + \rho$ and $\rho\gamma$ without the density for v.

***Example* 3: A binary endogenous explanatory variable.**

Evans and Schwab (1995): "Finishing High School and Starting College: Do Catholic Schools Make a Difference?" *Quarterly Journal of Economics*, vol 110, 941-974

The issue here is whether enrolling in catholic school helps one finishing school and starting college.

$$y_1 = 1[z_1\delta_1 + \alpha_1 y_2 + u_1 > 0]$$
$$y_2 = 1[z\delta_2 + v_2 > 0]$$

where var(u_1) = 1, and var(v_2) = 1. Let y_1 be finishing high school, or enrolling in college, and y_2 is enrolling in Catholic schools. It is possible that those who enroll in Catholic schools are different from other people – they may be more likely to finish high schools and to start college.

Statistically, if cov(u_1, v_2)=0, then the model can be consistently estimated by only using first equation. However, if cov(u_1, v_2)≠0, let: $u_1 = \rho v_2 + \varepsilon$. Therefore, we have: var(ε) = $1 - \rho^2$.

Plug this equation into the y_1 equation:

$$y_1 = 1[z_1\delta_1 + \alpha_1 y_2 + \rho v_2 + \varepsilon > 0]$$

The joint density of y_1 and y_2 are given by:

$$\begin{aligned}
&\Pr(y_1 = 1, y_2 = 1 \mid z) \\
&= \Pr(z_1\delta_1 + \alpha_1 y_2 + \rho v_2 + \varepsilon > 0,\ z\delta_2 + v_2 > 0 \mid z) \\
&= \Pr(\varepsilon > -z_1\delta_1 - \alpha_1 y_2 - \rho v_2,\ v_2 > -z\delta_2 \mid z) \\
&= \int_{-z\delta_2}^{\infty} \Phi\left(\frac{z_1\delta_1 + \alpha_1 y_2 + \rho v_2}{\sqrt{1-\rho^2}}\right)\phi(v_2)dv_2
\end{aligned}$$

$$\begin{aligned}
&\Pr(y_1 = 0, y_2 = 1 \mid z) \\
&== \Pr(z_1\delta_1 + \alpha_1 y_2 + \rho v_2 + \varepsilon < 0,\ z\delta_2 + v_2 > 0 \mid z) \\
&= \int_{-z\delta_2}^{\infty} \left(1 - \Phi\left(\frac{z_1\delta_1 + \alpha_1 y_2 + \rho v_2}{\sqrt{1-\rho^2}}\right)\right)\phi(v_2)dv_2
\end{aligned}$$

$$\Pr(y_1 = 1, y_2 = 0 \mid z) = \int_{-\infty}^{-z\delta_2} \Phi\left(\frac{z_1\delta_1 + \alpha_1 y_2 + \rho v_2}{\sqrt{1-\rho^2}}\right)\phi(v_2)dv_2$$

$$\Pr(y_1 = 0, y_2 = 0 \mid z) = \int_{-\infty}^{-z\delta_2} \left(1 - \Phi\left(\frac{z_1\delta_1 + \alpha_1 y_2 + \rho v_2}{\sqrt{1-\rho^2}}\right)\right)\phi(v_2)dv_2$$

Discussions:

(1) It is important to note that one cannot apply "two-stage" regression. For example, it is tempting to apply the following two-stage regression:

Stage 1: estimate a Probit model of $y_2 = 1[z\delta_2 + v_2 > 0]$. Let the estimates be $\hat{\delta}_2$.

Given that $E(y_2) = \Phi(z\delta_2)$,

Stage 2: estimate another Probit model of:

$$y_1 = 1[z_1\delta_1 + \alpha_1\Phi(z\hat{\delta}_2) + u_1 > 0]$$

This "two-stage" model is WRONG – because of nonlinearity. This is one type of "forbidden regressions." The reason, as discussed before, is that:

$$E(y_1 \mid z) = E[1(z_1\delta_1 + \alpha_1 y_2 + u_1 > 0)] \neq \Phi(z_1\delta_1 + \alpha_1\Phi(z\delta_2))$$

(2) However, it is possible to test the endogeneity by the following Rivers-Vuong procedure:

Stage 1: estimate a Probit model of $y_2 = 1[z\delta_2 + v_2 > 0]$. Let the estimates be $\hat{\delta}_2$.

Obtain residual $\hat{v}_2 = y_2 - \Phi(z\hat{\delta}_2)$

Stage 2: run Probit:

$$y_1 = 1[z_1\delta_1 + \alpha_1 y_2 + \rho\hat{v}_2 + \varepsilon > 0]$$

The test of endogeneity is a test of: $\rho = 0$

(3) Moreover, running Probit on the following equation does not give you consistent estimates:

$$y_1 = 1[z_1\delta_1 + \alpha_1 y_2 + \rho\hat{v}_2 + \varepsilon > 0]$$

The reason for that is clear: var(ε) = $1 - \rho^2 < 0$. Therefore, any probit estimates based previous equation would be biased.

***Example* 4: A censored endogenous explanatory variable.**

Gan and Sabarwal (2007): "A Simple Test of Adverse Events and Strategic Timing Theories of Consumer Bankruptcy"

For personal bankruptcy, a key question of interest is: a person files for bankruptcy. Is it because he encounters adverse events (the adverse event model), or is it because he intentionally borrows money (for consumption) and then files for bankruptcy to avoid paying back (the strategic timing model)?

According to Gan and Sabarwal, if the strategic timing model is correct, the bankruptcy decision is going to affect the amount of debt. As a consequence, the debt is endogenous. Therefore, the key of testing which model is correct is to test if endogeneity of the financial benefits (or debt) from filing for bankruptcy. If financial benefit is endogenous, then the strategic timing model is correct. Otherwise, the adverse events model is correct.

One difficulty, however, is that the financial benefit, denoted as *FB*, is censored at zero. One cannot have negative financial benefit from filing for bankruptcy. Here is the model:

$$Bank = 1[z_1\delta_1 + \alpha_1 FB + u_1 > 0]$$
$$FB = \begin{cases} z_2\delta_2 + \beta AE + v_2 & if\ FB > 0 \\ 0 & if\ FB \le 0 \end{cases}$$

We let var(u_1) = 1, and var(v_2) = σ^2. The adverse events variables, *AE*, represents events such as unemployment, heath shocks, accidents, divorce, etc.

Statistically, if cov(u_1, v_2) = 0, then the model can be consistently estimated by only using first equation. However, if cov(u_1, v_2)≠ 0, let: $u_1 = \rho v_2 + \varepsilon$, where var($\varepsilon$) = 1 – $\rho^2\sigma^2$.

Plug this equation into the y_1 equation:

$$Bank = 1[z_1\delta_1 + \alpha_1 FB + \rho v_2 + \varepsilon > 0]$$

The joint density of *Bank* and *FB* are given by:

$$g\left(Bank=1, FB, FB>0 \mid z\right)$$
$$=\Pr\left(z_1\delta_1+\alpha_1 FB+\rho v_2+\varepsilon>0,\ FB;\ z_2\delta_2+\beta AE+v_2>0 \mid z\right)$$
$$=\Pr\left(z_1\delta_1+\alpha_1 FB+\rho v_2+\varepsilon>0, z_2\delta_2+\beta AE+v_2>0 \mid FB, z\right)f\left(FB\right)$$
$$=\int_{-z_2\delta_2-\beta AE}^{\infty}\Phi\left(\frac{z_1\delta_1+\alpha_1 FB+\rho v_2}{\sqrt{1-\rho^2\sigma^2}}\right)\frac{1}{\sigma}\phi\left(\frac{v_2}{\sigma}\right)dv_2\ \frac{1}{\sigma}\phi\left(\frac{FB-z_2\delta_2-\beta AE}{\sigma}\right)$$

$$g\left(Bank=0, FB; FB>0 \mid z\right)$$
$$=\Pr\left(z_1\delta_1+\alpha_1 FB+\rho v_2+\varepsilon<0, FB;\ z_2\delta_2+\beta AE+v_2>0\right)$$
$$=\int_{-z_2\delta_2-\beta AE}^{\infty}\left(1-\Phi\left(\frac{z_1\delta_1+\alpha_1 FB+\rho v_2}{\sqrt{1-\rho^2\sigma^2}}\right)\right)\frac{1}{\sigma}\phi\left(\frac{v_2}{\sigma}\right)dv_2\ \frac{1}{\sigma}\phi\left(\frac{FB-z_2\delta_2-\beta AE}{\sigma}\right)$$

$$g\left(Bank=1, FB=0 \mid z\right)$$
$$=\Pr\left(z_1\delta_1+\rho v_2+\varepsilon>0, z_2\delta_2+\beta AE+v_2<0\right)$$
$$=\int_{-\infty}^{-z_2\delta_2-\beta AE}\Phi\left(\frac{z_1\delta_1+\rho v_2}{\sqrt{1-\rho^2\sigma^2}}\right)\frac{1}{\sigma}\phi\left(\frac{v_2}{\sigma}\right)dv_2$$

$$g\left(Bank=0, FB=0 \mid z\right)$$
$$=\Pr\left(z_1\delta_1+\rho v_2+\varepsilon<0, z_2\delta_2+\beta AE+v_2<0\right)$$
$$=\int_{-\infty}^{-z_2\delta_2-\beta AE}\left(1-\Phi\left(\frac{z_1\delta_1+\rho v_2}{\sqrt{1-\rho^2\sigma^2}}\right)\right)\frac{1}{\sigma}\phi\left(\frac{v_2}{\sigma}\right)dv_2$$