Contents

1 Discrete choice modeling 1
  1.1 Luce Model/McFadden conditional logit model . . . . . . . . . 2
  1.2 Random Utility Models . . . . . . . . . . . . . . . . . . . . . 5
  1.3 Solving the forecasting problem . . . . . . . . . . . . . . . . 7
    1.3.1 Debreu criticism of Luce’s model . . . . . . . . . . . 8
  1.4 Verifying that a probability choice model is consistent with an
    RUM . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8
  1.5 Daly-Zachary-Williams Theorem . . . . . . . . . . . . . . . . 9
  1.6 Example: Choice of transportation mode . . . . . . . . . . . 12

2 Multinomial Probit Model 15
  2.1 MNP Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
  2.2 Identification . . . . . . . . . . . . . . . . . . . . . . . . . . 16
  2.3 How do we solve the forecasting problem? . . . . . . . . . . 21
  2.4 Application from Hausman and Wise (1978) . . . . . . . . . 22

3 Simulation Methods for MNP models 23
  3.1 Early simulation method - crude frequency method . . . . . . 24
  3.2 Simulated method of moments . . . . . . . . . . . . . . . . . 25

4 Nonrandom sampling 31
  4.1 Some examples of choice-based sampling . . . . . . . . . . . 31
  4.2 Choice-based sampling . . . . . . . . . . . . . . . . . . . . 32
    4.2.1 Manski and Lerman (1978) . . . . . . . . . . . . . . . . 34

5 Limited Dependent Variable Models 35
  5.1 Type I Tobit Model . . . . . . . . . . . . . . . . . . . . . . . 35
    5.1.1 Truncated Type I Tobit Model . . . . . . . . . . . . . 36
5.2 Different Ways of Estimating Tobit Models ........................................... 37
5.3 Type II Tobit Model .............................................................. 38
  5.3.1 The likelihood: ............................................................... 39
  5.3.2 Examples ................................................................. 41
5.4 Type III Tobit: Heckman’s Model (Incorporates Choice of H) 42
5.5 Type IV Tobit Model ............................................................ 43
5.6 Type V Tobit: Switching Regression Model ............................... 43
  5.6.1 Examples ................................................................. 44

6 The Roy Model ........................................................................ 45
  6.1 Model ................................................................. 45
  6.2 Mean earnings given self-selection ....................................... 47
  6.3 Estimation ................................................................. 50
  6.4 Constructing counterfactuals .............................................. 52
Chapter 1
Discrete choice modeling

Classical regression model:

\[ y = x\beta + \varepsilon \]
\[ E(\varepsilon|x) = 0 \]
\[ \varepsilon \sim N(0, \sigma^2) \]

This model is untenable if \( y \) is discrete. To model discrete choices, we need to think of the ingredients that give rise to choices.

For example, suppose we want to forecast the demand for a new good and we observe consumption data on old goods, \( x_1...x_I \). Each good might represent a transportation mode or a type of car.

Assume that people choose the good that yields the highest utility. When we have a new good, we need a way of putting it on a basis with the old.

The earliest literature on discrete choice was developed in psychometrics where researchers were concerned with modeling choice behavior.

Two predominant modeling approaches:
(1) Luce (1953) Model \( \leftrightarrow \) McFadden conditional logit model
(2) Thurstone (1929) \( \leftrightarrow \) Quandt multivariate probit/normal model
CHAPTER 1. DISCRETE CHOICE MODELING

The Luce-McFadden model

- widely used in economics
- easy to compute
- identifiability of parameters well understood
- very restrictive substitution possibilities among goods

The Thurstone-Quandt model

- very general substitution possibilities
- allows for more general forms of heterogeneity
- more difficult to compute
- identifiability less easily established

1.1 Luce Model/McFadden conditional logit model

References: Manski and Mcfadden, Chapter 5; Greene; Amemiya

Notation:

\[ X \] : universe of objects of choice
\[ S \] : universe of attributes of people
\[ B \] : feasible choice set

Behavior rule mapping attributes into choices:

\[ h(B, s) = x \]
We might assume that there is a distribution of choice rules because

- in observation we lose some information governing choices
- there can be random variation in choices due to unmeasured psychological factors

Define

\[ P(x|s, B) = \Pr(h \in H \text{ such that } h(s, B) = x) \]

This represents the probability that an individual drawn randomly from the population with attributes \( x \) and alternative set \( B \) chooses \( x \).

**Luce Axioms:**

Maintain some restrictions on \( P(x|s, B) \) and derive implications for the functional form of \( P \).

**Axiom #1: Independence of Irrelevant Alternatives (IIA)**

\[
\frac{P(x|s, \{xy\})}{P(y|s, \{xy\})} = \frac{P(x|s, B)}{P(y|s, B)}
\]

where \( B \) is a larger choice set

This means that other choices are irrelevant for the relative probability ratio. This is not necessarily a good assumption. For example, suppose the choice is a career decision and the possible choices are economist (E), policeman (P) and fireman (F). IIA implies

\[
\frac{P(E|s, \{EF\})}{P(F|s, \{EF\})} = \frac{P(E|s, \{EFP\})}{P(F|s, \{EFP\})}
\]

(example due to Stern)

This problem is known as the *Red Bus-Blue Bus problem*, which was first noted by Debreu. Suppose choices are \{car, red bus, blue bus\}

\[
\frac{P(c|s, \{c, RB\})}{P(RB|s, \{c, RB\})} = \frac{P(c|s, \{c, RB, BB\})}{P(RB|s, \{c, RB, BB\})}
\]
CHAPTER 1. DISCRETE CHOICE MODELING

Axiom 2: Eliminate 0 probability choices

\[ \Pr(y|s, B) > 0 \] for all \( y \in B \)

Let’s analyze the implications of the above axioms:

Define \( P_{xx} = P(x|x, \{xx\}) \) and assume \( P_{xx} = 1/2 \)

By IIA

\[
P(y|s, B) = \frac{P_{yx}}{P_{xy}} P(x|s, B)
\]

\[
\sum_{y \in B} P(y|s, B) = 1 \implies P(x|s, B) = \frac{1}{\sum_{y \in B} \frac{P_{yx}}{P_{xy}}}
\]

Furthermore, we have from IIA

\[
P(y|s, B) = \frac{P_{yz}}{P_{zy}} P(z|s, B)
\]

\[
P(x|s, B) = \frac{P_{xz}}{P_{zx}} P(z|s, B)
\]

\[
P(y|s, B) = \frac{P_{yx}}{P_{xy}} P(x|s, B)
\]

which implies

\[
\frac{P_{yx}}{P_{xy}} = \frac{P(y|s, B)}{P(x|s, B)} = \frac{P_{yx}}{P_{xy}} \frac{P_{xz}}{P_{zx}}
\]

Define \( \tilde{V}(s, x, z) = \ln \left( \frac{P_{xz}}{P_{zx}} \right) \) and \( \tilde{V}(s, y, z) = \ln \left( \frac{P_{yz}}{P_{zy}} \right) \). Then we can write

\[
\frac{P_{yx}}{P_{xy}} = \frac{e^{\tilde{V}(s,y,z)}}{e^{\tilde{V}(s,x,z)}}
\]

Axiom 3: Separability assumption

Assume that \( \tilde{V}(s, x, z) = v(s, x) - v(s, z) \).

\( \tilde{V} \) can be interpreted as a utility indicator of relative tastes.
1.2. RANDOM UTILITY MODELS

Then,

\[
P(x|s, B) = \frac{1}{\sum_{y \in B} \frac{P_{xs}}{P_{xy}}} = \frac{1}{\sum_{y \in B} \frac{e^{v(s,y) - v(s,x)}}{e^{v(s,x) - v(s,x)}}} = \frac{1}{\sum_{y \in B} e^{v(s,y)}}
\]

Thus, you can derive logistic form from the Luce Axioms.

1.2 Random Utility Models

Now, link to more familiar models in economics. Marshak (1959) established the link between the Luce model and so-called random utility models (RUMs). These kinds of models were proposed by Thurstone (1927, 1930’s)

Assume utility from choosing alternative \( j \) is

\[
u(s, x_j) = v(s, x_j) + \varepsilon(s, x_j)
\]

The second component \( \varepsilon(s, x_j) \) reflects stochastic, idiosyncratic tastes

\[
\Pr(j \text{ is maximal in set } B) = \Pr(u(s, x_j) \geq u(s, x_l) \ \forall l \neq j)
\]

\[
= \Pr(v(s, x_j) + \varepsilon(s, x_j) \geq v(s, x_l) + \varepsilon(s, x_l) \ \forall l \neq j)
\]

Specify a cdf \( F(\varepsilon_1, ..., \varepsilon_j) \). Then,

\[
\Pr(v_j - v_l \geq \varepsilon_l - \varepsilon_j \ \forall l \neq j)
\]

\[
= \Pr(v_j - v_l + \varepsilon_j \geq \varepsilon_l \ \forall l \neq j)
\]

\[
= \int_{-\infty}^{\infty} F_j(v_j - v_1 + \varepsilon_j, ..., v_j - v_{j-1} + \varepsilon_j, ..., v_j - v_J + \varepsilon_j) d\varepsilon_j
\]
Suppose that $x_1$ and $x_2$ are normal. Then
\[
\Pr(x_1 > x_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{x_1} f(x_1, x_2) dx_1 dx_2 = \int_{-\infty}^{\infty} F_1(x_1, x_1) dx_1
\]

If $\varepsilon$ is iid, then
\[
F(\varepsilon_1, \ldots, \varepsilon_J) = \prod_{i=1}^{n} F_i(\varepsilon_i)
\]
so the above expression equals
\[
\int_{-\infty}^{\infty} \left[ \prod_{l=1}^{J, l\neq j} F(v_j - v_l + \varepsilon_j) \right] F_i(\varepsilon_j) d\varepsilon_j
\]

**Binary Example ($J=2$)**

\[
\Pr(\text{choose 1}|x, B) = P(v_1 + \varepsilon_1 > v_2 + \varepsilon_2)
= P(\varepsilon_2 - \varepsilon_1 < v_1 - v_2)
= P(\varepsilon_2 < v_1 - v_2 + \varepsilon_1)
= \int_{-\infty}^{\infty} F_1(\varepsilon_1, v_1 - v_2 + \varepsilon_1) d\varepsilon_1
\]

If $\varepsilon_1, \varepsilon_2$ normal, then $\varepsilon_2 - \varepsilon_1$ normal, so $\Pr(\varepsilon_2 - \varepsilon_1 < v_1 - v_2)$ normal. If $\varepsilon_1, \varepsilon_2$ distributed type I extreme value (also called log Weibull) then $\varepsilon_2 - \varepsilon_1$ logistic.

\[\Rightarrow P(\varepsilon < c) = e^{-e^{-c+\alpha}}\]

$\varepsilon^{-}$ type I extreme value then

\[
\Pr(V_1 + \varepsilon_1 > V_2 + \varepsilon_2) = \Pr(\varepsilon_2 - \varepsilon_1 < V_1 - V_2)
= \Lambda(V_1 - V_2)
= \frac{e^{V_1 - V_2}}{1 + e^{V_1 - V_2}}
= \frac{e^{V_1}}{e^{V_1} + e^{V_2}}.
\]
• Assuming that errors follow a Weibull distribution yields same logit model as derived from the Luce Axioms (due to Marshak, 1959).

• It turns out that type I extreme value is sufficient but not necessary, some other distributions also generate a logit, but Yellot (1977) showed that if we require “invariance under uniform expansions of the choice set” then only the double exponential gives a logit. example: {coffee, tea, milk}

invariance requires that the probabilities stay the same if we double the choice set.

1.3 Solving the forecasting problem

Let $x_j$ be the set of characteristics associated with choice $j$.
Usually, we assume $V(s, x_j) = \theta(s)'x_j$, where dependence on $s$ denotes that individuals may differ in how they value characteristics of goods (e.g. cars).

$$
\Pr(j|s, B) = \frac{e^{\theta(s)'x_j}}{\sum_{l=1}^{N} e^{\theta(s)'x_l}}
$$

Solve for the parameters by Maximum likelihood:

$$
D_{ij} = 1 \text{ if individual } i \text{ chooses } j, = 0 \text{ else.}
$$

$$
\max_{\theta(s)} \prod_{i=1}^{N} \left[ \frac{e^{\theta(s)'x_{i1}}}{\sum_{l=1}^{J} e^{\theta(s)'x_{il}}} \right]^{D_{i1}} \cdots \left[ \frac{e^{\theta(s)'x_{iJ}}}{\sum_{l=1}^{J} e^{\theta(s)'x_{il}}} \right]^{D_{iJ}}
$$

Let $x_{J+1}$ be the vector of characteristics for a new good, and $B' = \{B, J + 1\}$ the augmented choice set.

$$
\Pr(J + 1|s, B') = \frac{e^{\hat{\theta}(s)'x_{j+1}}}{\sum_{l=1}^{J+1} e^{\hat{\theta}(s)'x_{l}}},
$$

where $\hat{\theta}(s)$ are the MLE estimates.
1.3.1 Debreu criticism of Luce’s model

Suppose the J+1th alternative is identical to the first alternative. Then

\[
\Pr(\text{choose 1 or } J + 1 | s, B') = \frac{2e^{\theta(s)'x_{N+1}}}{\sum_{l=1}^{J+1} e^{\theta(s)'x_l}}
\]

- This means that introducing an identical good into the choice set changes the probability, which is not an attractive result.
- The result comes from making an iid assumption on the error term associated with the new alternative

\[
\begin{align*}
U_1 + \varepsilon_1 \\
U_{J+1} + \varepsilon_{J+1}
\end{align*}
\]

If the goods are identical, then we expect the error terms to be perfectly correlated, not iid.

1.4 Verifying that a probability choice model is consistent with an RUM

What are criteria for a good probabilistic choice system?

1. has flexible functional form
2. is computationally practical
3. allows for flexibility in representing substitution patterns among choices
4. is consistent with a RUM (has a structural interpretation)

How do you verify whether consistent with a RUM?

• start with a RUM and derive the PCS

\[
U_i = V(s, x^i) + \varepsilon(s, x^i)
\]

and solve integral for

\[
\Pr(U_i > U_l \text{ for all } l \neq i) = \Pr(V^i + \varepsilon^i > V^l + \varepsilon^l \text{ for all } l \neq i)
\]

• start with PCS and verify that it is consistent with a RUM. McFadden provides sufficient conditions for this
1.5 Daly-Zachary-Williams Theorem

Daly-Zachary (1976), Williams (1977)

Provides a set of conditions that make it easy to derive a PCS from a RUM within a class of models called generalized extreme value (GEV) models.

Define $G = G(Y_1, ..., Y_J)$

If $G$ satisfies the following conditions:

1. nonnegative defined on $Y_1, ..., Y_J \geq 0$
2. homogeneous of degree one in its arguments
3. $\lim_{Y_i \to \infty} G(Y_1, ..., Y_i, ..., Y_J) \to \infty$, for all $i = 1 .. J$
4. $\frac{\partial^k G}{\partial Y_1 \ldots \partial Y_k}$ is nonnegative if $k$ is odd, and nonpositive if $k$ is even

Then, for a RUM with $U_i = V_i + \varepsilon_i$ and $F(\varepsilon_1, ..., \varepsilon_J) = \exp\{-G(e^{-\varepsilon_1}, ..., e^{-\varepsilon_J})\}$ the PCS is given by (with $Y_i = e^{V_i}$)

$$P_i = \frac{\partial \ln G}{\partial \ln Y_i} = \frac{e^{V_i} G_i(e^{V_1}, ..., e^{V_J})}{G(e^{V_1}, ..., e^{V_J})}$$

Remark: McFadden shows that under certain conditions on the form of $V_i$ (satisfies AIRUM form), the DZW theorem can be seen as a form of Roy’s identity.

Let’s apply the result:

(a) Multinomial logit model

$$F(\varepsilon_1, ..., \varepsilon_J) = e^{-\varepsilon_1} ... e^{-\varepsilon_J} = e^{-\sum_{j=1}^{J} \varepsilon_j}$$

Here,

$$G(e^{V_1}, ..., e^{V_J}) = \sum_{j=1}^{J} e^{V_j}$$

satisfies DZW conditions

Applying DZW, get the MNL model:

$$P_i = \frac{\partial \ln G}{\partial V_i} = \frac{e^{V_i}}{\sum_{j=1}^{J} e^{V_j}}$$
(b) Nested logit model (addresses to a limited extent the IIA criticism)

\[ G(e^{V_1}, \ldots, e^{V_J}) = \sum_{m=1}^{M} a_m \left[ \sum_{i \in B_m} e^{V_i} \right]^{1-\sigma_m} \]

Idea is to first divide goods into \( M \) branches. First choose a branch and then choose a good within a branch. Will allow for correlation between the errors within a given branch.

Define

\[ B_m \subseteq \{1, \ldots, J_m\} \]
\[ \bigcup_{m=1}^{M} B_m = B \]

Calculate the equation for \( P_i \) using the DZW result:

\[ P_i = \frac{\partial \ln G_i}{\partial V_i} = \frac{\sum_{m \text{ s.t. } i \in B_m} a_m \left( e^{\frac{V_i}{1-\sigma_m}} \right) \left[ \sum_{i \in B_m} e^{\frac{V_i}{1-\sigma_m}} \right]^{-\sigma_m}}{\sum_{m=1}^{M} a_m \left[ \sum_{i \in B_m} e^{\frac{V_i}{1-\sigma_m}} \right]^{1-\sigma_m}} \]

If we multiply by

\[ \left[ \sum_{i \in B_m} e^{\frac{V_i}{1-\sigma_m}} \right]^{-1} \left[ \sum_{i \in B_m} e^{\frac{V_i}{1-\sigma_m}} \right], \]

get

\[ = \sum_{m=1}^{M} \Pr(i|B_m) \Pr(B_m), \]

where

\[ \Pr(i|B_m) = \frac{\left( e^{\frac{V_i}{1-\sigma_m}} \right)}{\sum_{i \in B_m} e^{\frac{V_i}{1-\sigma_m}}} \text{ if } i \in B_m, = 0 \text{ otherwise} \]
\[ \Pr(B_m) = \frac{a_m \left[ \sum_{i \in B_m} e^{\frac{V_i}{1-\sigma_m}} \right]^{1-\sigma_m}}{\sum_{m=1}^{M} a_m \left[ \sum_{i \in B_m} e^{\frac{V_i}{1-\sigma_m}} \right]^{1-\sigma_m}} \]
How does the nested logit model accommodate the Red bus/Blue bus problem?

Suppose

\[ G = Y_1 + \left[ \frac{Y_2}{1 - \sigma} + \frac{Y_3}{1 - \sigma} \right]^{1 - \sigma} \]

\[ Y_i = e^{V_i} \]

\[ \frac{\partial \ln G}{\partial V_1} = \frac{e^{V_i}}{e^{V_1} + \left[ e^{V_2/(1 - \sigma)} + e^{V_3/(1 - \sigma)} \right]^{1 - \sigma}} \]

\[ \frac{\partial \ln G}{\partial V_2} = \frac{e^{V_i}}{e^{V_1} + \left[ e^{V_2/(1 - \sigma)} + e^{V_3/(1 - \sigma)} \right]^{1 - \sigma}} \]

\[ \frac{\partial \ln G}{\partial V_3} = \frac{e^{V_i}}{e^{V_1} + \left[ e^{V_2/(1 - \sigma)} + e^{V_3/(1 - \sigma)} \right]^{1 - \sigma}} \]

As \( V_3 \to -\infty \) get logistic

\[ \frac{e^{V_2}}{e^{V_1} + e^{V_2}} \]

As \( V_1 \to -\infty \)

\[ \Pr(2|\{123\}) = \frac{e^{V_2/(1 - \sigma)}}{e^{V_2/(1 - \sigma)} + e^{V_3/(1 - \sigma)}} \]

What role does \( \sigma \) play? \( \sigma \) is the degree of substitutability parameter.

Recall that

\[ F(\varepsilon_1, \varepsilon_2, \varepsilon_3) = \exp\{ -G(e^{-\varepsilon_1}, e^{-\varepsilon_2}, e^{-\varepsilon_3}) \} \]

Here,

\[ \sigma = \frac{\text{cov}(\varepsilon_2, \varepsilon_3)}{\sqrt{\text{var}(\varepsilon_2)\text{var}(\varepsilon_3)}} \]

Thus,

\[ -1 \leq \sigma \leq 1 \text{ (correlation coefficient)} \]

It turns out that we also require \( \sigma > 0 \) for the DZW conditions to be satisfied, which is unfortunate because we cannot allow the \( \varepsilon \)'s to be negatively correlated.

Can show that

\[ \lim_{\sigma \to 1} \Pr(1|\{123\}) = \frac{e^{V_i}}{e^{V_1} + \max(e^{V_2}, e^{V_3})}. \]
What happens if $V_2 = V_3$?

Then,

$$\Pr(2|\{123\}) = \frac{e^{V_2/(1-\sigma)}[2e^{V_2/(1-\sigma)}]^{1-\sigma}}{e^{V_1} + [2e^{V_2/(1-\sigma)}]^{1-\sigma}}$$

$$= \frac{2^{1-\sigma} e^{V_2}}{e^{V_1} + 2^{1-\sigma} e^{V_2}}$$

$$\lim_{\sigma \to 1} = \frac{1}{2} \frac{e^{V_2}}{e^{V_1} + e^{V_2}}$$

which means you introduce an identical alternative and cut the probability of choosing good 2 by one half. This solves the Red-Bus-Blue-Bus problem.

Remark:

Can expand the nested logit to accommodate multiple levels. For example, three levels:

$$G = \sum_{q=1}^{Q} a_q \left\{ \sum_{m \in Q_q} a_m \left[ \sum_{i \in B_m} Y_i^{1/(1-\sigma_m)} \right]^{1-\sigma_m} \right\}$$

### 1.6 Example: Choice of transportation mode

Neighborhood $m$

Transportation mode $t$

$P(m)$ : probability of residing in neighborhood $m$

$P(i|B_m)$ : probability of choosing $i$th mode, given live in neighborhood $m$
1.6. EXAMPLE: CHOICE OF TRANSPORTATION MODE

Note all modes available in each neighborhood

\[
P_{m,t} = \frac{\sum_{t=1}^{T_m} e^{v(m,t) - \sigma_m}}{\sum_j \sum_{t=1}^{T_j} e^{v(j,t) - \sigma_m}}
\]

\[
P_{t|m} = \frac{\sum_t e^{v(m,t) - \sigma_m}}{\sum_t e^{v(m,t) - \sigma_m}}
\]

\[
P_m = \frac{\left[ \sum_{t=1}^{T_m} e^{v(m,t) - \sigma_m} \right]^{1-\sigma_m}}{\sum_j \sum_{t=1}^{T_j} e^{v(j,t) - \sigma_m}^{1-\sigma_m}}
\]

Standard type of utility function that people might use:

\[
z_t : \text{trans. mode characteristics}
\]

\[
x_i : \text{person characteristics}
\]

\[
y_m : \text{neighborhood characteristics}
\]

\[
V(m, t) = z'_t \gamma + x'_i \beta_{mt} + y'_m \alpha
\]

Could also include interaction terms between \(x\) and \(z\) and \(y\).

Estimate by MLE

\[
\Pi_{i=1}^n P_{tm}
\]
CHAPTER 1. DISCRETE CHOICE MODELING
Chapter 2

Multinomial Probit Model


2.1 MNP Model

Under the random utility model where $j$ denotes the alternative,

\[ U_j = Z_j \beta + \eta_j \]

\[ = Z_j \bar{\beta} + \{ Z_j (\beta - \bar{\beta}) + \eta_j \} \]

This is an example of a random coefficients model, because the parameter $\beta$ (which captures how people vary characteristics of the goods) is permitted to vary across people.

Assume there are $J$ alternatives, $K$ characteristics of the alternatives and that $\beta \sim N(\bar{\beta}, \Sigma_{\beta})$.

Define

\[
Z_{J \times K} \bar{\beta}_{K \times 1} = \left( \begin{array}{c} z'_1 \bar{\beta} \\ \vdots \\ z'_J \bar{\beta} \end{array} \right), \quad \tilde{\eta} = \left( \begin{array}{c} \eta_1 \\ \vdots \\ \eta_J \end{array} \right)
\]

Also, we will assume that $\tilde{\eta} \sim N(0, \Sigma_{\eta})$. 

15
Then,
\[
\Pr(\text{alternative } j \text{ selected}) = \Pr(U_j > U_i \text{ for all } i \neq j) = \int_{U_j = -\infty}^{\infty} \int_{U_1 = -\infty}^{U_j} \ldots \int_{U_J = -\infty}^{U_j} \phi(U|V_\mu, \Sigma_\mu) dU_J \ldots dU_1 dU_j
\]

Remarks:

- Here, \( \phi \) is a J-dimensional multivariate normal density with mean \( V_\mu \) and variance \( \Sigma_\mu \).

- Unlike the MVL, no closed form expression exists for the integral in the normal case. The integrals are therefore usually evaluated using simulation methods.

How many parameters are there?

\[
\begin{align*}
\tilde{\beta} & : \quad K \text{ parameters} \\
\Sigma_\beta & : \quad K \times K \text{ symmetric matrix, } \frac{K^2 - K}{2} + K = \frac{K(K + 1)}{2} \text{ parameters} \\
\Sigma_\eta & : \quad \frac{J(J + 1)}{2} \text{ parameters}
\end{align*}
\]

When a person chooses alternative \( j \), all we know is relative utility, not absolute utility. This suggests that not all the parameters in the model will be identified. We will therefore require some normalizations.

### 2.2 Identification

What does it mean to say a parameter is identified in a model? If a parameter or set of parameters is not identified, then a model with one parameterization is observationally equivalent to a model with a different parameterization.
Example: Binary probit model (with fixed $\beta$)

Let $X_i$ denote characteristics of the individual.

\[
\Pr(d = 1|X_i) = \Pr(V_1 + \varepsilon_1 > V_2 + \varepsilon_2) = \Pr(X_i\beta_1 + \varepsilon_1 > X_i\beta_2 + \varepsilon_2) = \Pr(X_i(\beta_1 - \beta_2) > \varepsilon_2 - \varepsilon_1) = \Pr\left(\frac{X_i(\beta_1 - \beta_2)}{\sigma} > \frac{\varepsilon_2 - \varepsilon_1}{\sigma}\right) = \Phi\left(\frac{X_i\tilde{\beta}}{\sigma}\right), \quad \text{where } \tilde{\beta} = \beta_1 - \beta_2
\]

$\Phi\left(\frac{X_i\tilde{\beta}}{\sigma}\right)$ is observationally equivalent to $\Phi\left(\frac{X_i\beta^*}{\sigma^*}\right)$ for $\frac{\tilde{\beta}}{\sigma} = \frac{\beta^*}{\sigma^*}$, so $\beta$ is not separately identified relative to $\sigma$.

But the ratio is identified. Suppose

\[
\Phi\left(\frac{\tilde{X}_i\beta}{\sigma}\right) = \Phi\left(\frac{\tilde{X}_i\beta^*}{\sigma^*}\right) = \Phi^{-1}\Phi\left(\frac{\tilde{X}_i\beta^*}{\sigma^*}\right) = \Phi^{-1}\Phi\left(\frac{\tilde{X}_i\beta}{\sigma}\right)
\]

\[
\sum_{i=1}^{n} \tilde{X}_i'\tilde{X}_i\left(\frac{\beta}{\sigma}\right) = \sum_{i=1}^{n} \tilde{X}_i'\tilde{X}_i\left(\frac{\beta^*}{\sigma^*}\right)
\]

which implies

\[
\frac{\beta}{\sigma} = \frac{\beta^*}{\sigma^*}
\]

The set

\[
\{b : b = \beta\gamma, \gamma \text{ any positive scalar}\}
\]

is identified

We would say that ”$\beta$ is identified up to scale and that the sign is identified” (because $\sigma$ must be positive).

**Definition of Identification**

For a parameter space $B$ and a random vector $Z$, identification is selecting a subset in $B$ that characterizes some aspect of the probability distribution of $Z$. 
Now, return to identification in the MVP model.

\[
\text{Pr}(j \text{ selected} \mid V_\mu, \Sigma_\mu) = \text{Pr}(U_i - U_j < 0 \text{ for all } i \neq j)
\]

Define the contrast matrix

\[
\Delta_j = \begin{pmatrix}
1 & 0 & \ldots & -1 & \ldots & 0 \\
-1 & 0 & \vdots & \vdots \\
0 & -1 & 0 & 1
\end{pmatrix}_{(J-1) \times J}
\]

\[
\Delta_j \tilde{U} = \begin{pmatrix}
U_1 - U_j \\
\vdots \\
U_J - U_j
\end{pmatrix}
\]

\[
\text{Pr}(j \text{ selected} \mid V_\mu, \Sigma_\mu) = \text{Pr}(\Delta_j \tilde{U} < 0 \mid V_\mu, \Sigma_\mu) = \Phi(0 \mid V_Z, \Sigma_Z)
\]

where

- \(V_Z\) is the mean of \(\Delta_j \tilde{U}\), which is \(\Delta_j \tilde{Z} \beta\)
- \(\Sigma_Z\) is the variance, which is \(\Delta_j \tilde{Z} \Sigma_\beta \tilde{Z}' \Delta_j' + \Delta_j \Sigma_\eta \Delta_j'\)

\(V_Z\) is \((J - 1) \times 1\)

\(\Sigma_Z\) is \((J - 1) \times (J - 1)\)

Here, we reduced the dimension of the integration problem by one. This representation shows that all the information is in the contrasts and we therefore cannot identify all of the components of \(V_\mu\) and \(\Sigma_\mu\).

Now define \(\tilde{\Delta}_j\) as \(\Delta_j\) with the ith column removed. Choose \(J\) as the "reference alternative" with corresponding \(\tilde{\Delta}_J\). We can verify that

\[
\Delta_j = \tilde{\Delta}_j \Delta_j
\]
2.2. IDENTIFICATION

For example, with three goods

\[
\begin{pmatrix}
1 & -1 \\
0 & -1
\end{pmatrix}
\begin{pmatrix}
1 & 0 & -1 \\
0 & 1 & -1
\end{pmatrix} =
\begin{pmatrix}
1 & -1 & 0 \\
0 & -1 & 1
\end{pmatrix}
\]

\[\tilde{\Delta}_2 \Delta_3 = \Delta_2\]

Therefore, we can write:

\[
V_Z = \Delta_j \tilde{Z} \tilde{\beta} \\
\Sigma_Z = \Delta_j \tilde{Z} \Sigma_{\beta} \tilde{Z}' \Delta'_j + \tilde{\Delta}_j \Delta_j \Sigma_{\eta} \Delta'_j \tilde{\Delta}'_j
\]

The last expression, \(C_j = \tilde{\Delta}_j \Delta_j \Sigma_{\eta} \Delta'_j \tilde{\Delta}'_j\), is \((J - 1) \times (J - 1)\), with \(\frac{J(J-1)}{2}\) total parameters.

- Because the original model can always be expressed in terms of a model with \((\beta, \Sigma_{\beta}, C_j)\), it follows that some of the parameters in the original model are not identified.

- How many parameters are identified?

Original model: \(K + \frac{K(K+1)}{2} + \frac{J(J+1)}{2}\)

Now: \(K + \frac{K(K+1)}{2} + \frac{J(J-1)}{2}\)

So, \(\frac{J^2 - J - (J^2 - J)}{2} = J\) not identified.

- It turns out that one additional parameter is not identified, because evaluation of

\[\Phi(0|kV_Z, k^2\Sigma_Z)\]

gives same result as evaluating

\[\Phi(0|V_Z, \Sigma_Z),\]

so can eliminate one more parameter through a suitable choice of \(k\). Thus, \(J + 1\) parameters from the original model are not identified.
CHAPTER 2. MULTINOMIAL PROBIT MODEL

Illustration of how to impose normalization

\[ J = 3 \]
\[ \Sigma_\eta = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \]

Set 3 as the reference alternative:

\[ C_3 = \Delta_3 \Sigma_\eta \Delta_3' \]
\[ = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \Sigma_\eta \begin{pmatrix} 1 & 0 & -1 \end{pmatrix}' \]
\[ = \begin{pmatrix} \sigma_{11} - 2\sigma_{31} + \sigma_{33} & \sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{33} \\ \sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{33} & \sigma_{22} - 2\sigma_{32} + \sigma_{33} \end{pmatrix} \]

Only combinations of parameters are identified. For example,

\[ \tilde{\Delta}_2 \Delta_3 \Sigma_\eta \Delta_3' \tilde{\Delta}_2' \]
\[ = \begin{pmatrix} 1 & -1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sigma_{11} - 2\sigma_{31} + \sigma_{33} & \sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{33} \\ \sigma_{21} - \sigma_{31} - \sigma_{32} + \sigma_{33} & \sigma_{22} - 2\sigma_{32} + \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix} \]

Normalization approach of Albreit, Lerman and Manski (1978):

- need \( J + 1 \) restrictions on the VCV matrix

- fix \( J \) parameters by settings last row and last column of \( \Sigma_\eta \) to 0

- fix the scale by constraining the diagonal elements of \( \Sigma_\eta \) so that \( \text{trace} \left( \Sigma_\eta \right) \) equals the variance of a standard log Weibull (they wanted to compare estimates of MNL with those independent).
2.3. HOW DO WE SOLVE THE FORECASTING PROBLEM?

2.3 How do we solve the forecasting problem?

Suppose we have two goods and add a third.

\[
\Pr(\text{choose } 1|Z^1, Z^2) = \Pr(U^1 - U^2 \geq 0) = \Pr((Z^1 - Z^2)\bar{\beta} \geq w^2 - w^1)
\]

where

\[
w^1 = Z^1(\beta - \bar{\beta}) + \eta^1
\]

\[
w^2 = Z^2(\beta - \bar{\beta}) + \eta^2
\]

Because \(w^2 - w^1\) is normally distributed,

\[
\int_{-\infty}^{(Z^1 - Z^2)\bar{\beta}} \int_{-\infty}^{(Z^1 - Z^2)\bar{\beta}} e^{-t/2} dt
\]

Now add a third good:

\[
U^3 = Z^3\bar{\beta} + Z^3(\beta - \bar{\beta}) + \eta^3
\]

Problem is that we don’t know the correlation of \(\eta^3\) with the other errors.

Suppose we knew that \(\eta^2 = \eta^3\).

Then

\[
\Pr(1 \text{ is chosen}) = \int_{-\infty}^{a} \int_{-\infty}^{b} BVN dt_1 dt_2
\]

where

\[
a = \frac{(Z^1 - Z^2)\bar{\beta}}{[\sigma_{11} + \sigma_{22} - 2\sigma_{12} + (Z^2 - Z^1)\Sigma\beta(Z^2 - Z^1)\gamma]^1/2}
\]

\[
b = \frac{(Z^1 - Z^3)\bar{\beta}}{[\sigma_{11} + (Z^2 - Z^1)\Sigma\beta(Z^3 - Z^1)\gamma]^1/2}
\]

We could also solve the forecasting problem if we make an assumption like \(\eta^2 = \eta^3\).
CHAPTER 2. MULTINOMIAL PROBIT MODEL

2.4 Application from Hausman and Wise (1978)

Developed a model of commuting choices for workers in Washington, D. C. Choice of

(a) driving car
(b) sharing ride
(c) riding bus

Utility model for person $i$ choosing mode $j$

$$U_{ij} = \beta_{i1} \log x_{1ij} + \beta_{i2} \log x_{2ij} + \beta_{i3} \frac{x_{3ij}}{x_{4i}} + \varepsilon_{ij}$$

$x_{1ij} = \text{time in vehicle}$
$x_{2ij} = \text{time out of vehicle}$
$x_{3ij} = \text{cost of mode for that person}$
$x_{4i} = \text{income of person}$
Chapter 3
Simulation Methods for MNP models

Models tend to be difficult to estimate because of high dimensional integrals that need to be evaluated at each stage of estimating the likelihood.

There are a variety of simulation methods that have been developed for estimating these types of models. Two popular methods are:

- Simulated method of moments (SMOM)
- Simulated maximum likelihood (SMLE)

References:
Lerman and Manski (1981) - Structural Analysis of Discrete Data
McFadden (1989) Econometrica
Hajivassilou and Ruud, Ch. 20, Handbook of Econometrics, Stern (1992) Econometrica
Stern (1997) survey article in Journal of Economic Literature
Gourieroux, Christin and Alain Monfort (2002): Simulation-Based Econometric Methods
Consider a model with only taste heterogeneity but no preference heterogeneity:

\[ U_j = Z_j \beta + \eta_j \]

\( \beta \) fixed

\( \eta_j \sim N(0, \Omega) \)

\( J \) choices

Let

\[ P_{ij} = \text{prob person } i \text{ chooses } j \]

\[ Y_{ij} = 1 \text{ if } i \text{ chooses } j, = 0 \text{ else} \]

The likelihood can be written as

\[ L = \prod_{i=1}^{N} \prod_{j=1}^{J} (P_{ij})^{Y_{ij}} \]

\[ \log L = \sum_{i=1}^{N} \sum_{j=1}^{J} Y_{ij} \log P_{ij} \]

\[ = \sum_{i=1}^{N} \log P_{ik} \]

where \( k \) is the alternative chosen by \( i \).

**Simulation Algorithm:**

1. For a given \( \beta, \Omega \) generate \( R \) Monte Carlo draws of the vector of residuals, \( \eta \)
2. Let

\[ \tilde{P}_k = \frac{1}{R} \sum_{r=1}^{R} 1(U_{rk} = \max\{U_{r1}, \ldots, U_{rJ}\}) \]

\[ \tilde{P}_k \] is a frequency simulator of the proportion of times that \( k \) is chosen given \( \beta, \Omega \).

3. Maximize \( \sum_{i=1}^{N} \log \tilde{P}_{ik} \) over alternative values for \( \beta, \Omega \).

Lerman and Manski found that this procedure performs poorly and requires a large number of draws to be accurate, particularly when \( P \) is close to 0 or 1.

Variance of the frequency simulator is \( \frac{1}{R^2} R\text{Var}(1()) = \frac{P_k(1-P_k)}{R} \).

McFadden (1989) provided some insights into how to improve the simulation method. He showed that simulation is a viable method of estimation even for a small number of draws provided that

(a) an unbiased simulator is used

(b) functions that are simulated appear linearly in the conditions defining the estimator

(c) same set of random draws is used to simulate the model at different parameter values

Condition (b) is violated for the crude frequency simulator, which has \( \log \tilde{P}_{ik} \).

### 3.2 Simulated method of moments

Consider model we had before, but with only preference heterogeneity and no taste heterogeneity:

\[ U_{ij} = Z_{ij} \beta = Z_{ij} \bar{\beta} + Z_{ij} \varepsilon_i \]

\[ \beta = \bar{\beta} + \varepsilon_i \]
Define
\[ P_{ij}(\gamma) = \Pr(i \text{ chooses } j | Z_i, \beta, \Sigma_{\varepsilon}) \]
\[ Y_{ij} = 1 \text{ if } i \text{ chooses } j, = 0 \text{ else} \]

The likelihood is given by:
\[
\log L = \frac{1}{N_0} \sum_{i=1}^{N} \sum_{j=1}^{J} Y_{ij} \ln P_{ij}(\gamma), \quad N_0 = NJ
\]
\[
\frac{\partial \log L}{\partial \gamma} = \frac{1}{N_0} \sum_{i=1}^{N} \sum_{j=1}^{J} Y_{ij} \left\{ \frac{\partial P_{ij}}{\partial \gamma} \right\} = 0 \quad (*)
\]

Now use the fact that
\[
\sum_{j=1}^{J} P_{ij}(\gamma) = 1
\]
which implies
\[
\sum_{j=1}^{J} \frac{\partial P_{ij}}{\partial \gamma} = 0
\]
and
\[
\sum_{j=1}^{J} \frac{\partial P_{ij}}{P_{ij}} P_{ij} = 0.
\]

We can therefore rewrite (*) as
\[
\frac{1}{N_0} \sum_{i=1}^{N} \sum_{j=1}^{J} (Y_{ij} - P_{ij}) \left\{ \frac{\partial P_{ij}}{P_{ij}(\gamma)} \right\} = 0
\]

Note that \( E(Y_{ij}) = P_{ij} \). Letting \( Z_{ij} = \frac{\partial P_{ij}}{P_{ij}(\gamma)} \) and \( \varepsilon_{ij} = Y_{ij} - P_{ij} \), we have
\[
\frac{1}{N_0} \sum_{i=1}^{N} \sum_{j=1}^{J} \varepsilon_{ij} Z_{ij} = 0,
\]
which is like a moment condition using $Z_{ij}$ as an instrument. So far, $P_{ij}$ is still a J-1 dimensional integral.

The method of simulated moments solves the first-order conditions.

\textit{Model:}

\begin{align*}
U_{ij} &= Z_{ij}\beta + Z_{ij}\varepsilon_{ij} \\
\tilde{U}_i &= \tilde{Z}_i\beta + \tilde{Z}\Gamma \tilde{e}_i \\
\tilde{U}_i &\text{ is } J \times 1 \\
\tilde{Z}_i &\text{ is } J \times K \\
\tilde{e}_i &\text{ is } K \times 1 \\
\Gamma &\text{ is } K \times K \\
\end{align*}

\begin{align*}
\Gamma\Gamma' &= \Sigma_\varepsilon \text{ (cholesky decomposition)} \\
\tilde{e}_i^{-1} &\sim N(0, I_k)
\end{align*}

\textit{Simulation algorithm:}

(i) Generate $\tilde{e}_i$ for each $i$ such that $\tilde{e}_i$ are iid across persons and distributed $N(0, I_k)$. In total, generate $NK^R$ draws, where $N$ is the sample size, $K$ is the number of characteristics, and $R$ is the number of Monte Carlo draws.

(ii) Fix matrix $\Gamma$ and obtain

\[ \eta_{ij} = Z_{ij}\Gamma \tilde{e}_i \]

and form vector

\[
\begin{pmatrix}
z_{i1} \Gamma \tilde{e}_i \\
z_{i2} \Gamma \tilde{e}_i \\
\vdots \\
z_{ij} \Gamma \tilde{e}_i
\end{pmatrix}
\]

for each person.

(iii) Fix $\tilde{\beta}$ and generate:

\[ \tilde{U}_{ij} = Z_{ij} \tilde{\beta} + \eta_{ij} \]

for all $i$.  

(iv) Find relative frequency that the ith person chooses alternative j across Monte Carlo draws:

\[ \tilde{P}_{ij}(\gamma) = \frac{1}{R} \sum_{r=1}^{R} 1(\bar{U}_{i1} > \bar{U}_{im} \text{ for all } m \neq j) \]

\( \tilde{P}_{ij}(\gamma) \) is a simulator for \( P_{ij}(\gamma) \). Stack to get \( \tilde{P}_i(\gamma) \).

(v) To get \( \tilde{P}_{ij}(\gamma) \) for different values of \( \gamma \), repeat steps (i) through (iv) using the same random variable draws \( \bar{e}_i \) from step (i).

(vi) Obtain a numerical simulator for \( \frac{\partial P_{ij}}{\partial \gamma} \) : For small \( h \),

\[ \frac{\partial P_{ij}(\gamma)}{\partial \gamma} = \frac{P_{ij}(\gamma + h\epsilon_m) - P_{ij}(\gamma - h\epsilon_m)}{2h} \]

where \( \epsilon_m \) is a vector with a 1 in the mth place.

Now, we have the ingredients for evaluating the moment conditions. Can use a method such as Gauss-Newton method to solve the moment conditions.

**Limitations of Simulation Methods**

1. \( \tilde{P}_{ij}(\gamma) \) simulator not smooth due to presence of indicator function (causes difficulties in deriving the asymptotic distribution - need to use methods developed by Pakes and Pollard (1989) for nondifferentiable functions)

2. \( \tilde{P}_{ij} \) cannot equal 0 (causes problems in denominator when close to 0)

3. Simulating a small \( \tilde{P}_{ij} \) may require a large number of draws


Replaces the indicator function with a smooth function, which is differentiable

\[ P_{ij}(\gamma) = \frac{1}{R} \sum_{r=1}^{R} k(\bar{U}_{ij} - \bar{U}_{im}) \]

The smoothed function goes smoothly from 0 to one around 0 instead of changing abruptly at 0.
3.2. SIMULATED METHOD OF MOMENTS

How does simulation affect the asymptotic distribution?

Define

\[ W_i = \frac{\partial P_i}{\partial \gamma} P_i(\gamma). \]

The variance of the method of moments estimator is given by

\[
D(\gamma) = \text{plim} \left[ \frac{1}{N} \sum_{i=1}^{N} W_i' \frac{\partial P_i}{\partial \gamma} \right]^{-1} \\
\text{plim} \left[ \frac{1}{N} \sum_{i=1}^{N} (1 + \frac{1}{R}) P_i(1 - P_i) W_i' W_i \right] \\
\text{plim} \left[ \frac{1}{N} \sum_{i=1}^{N} W_i' \frac{\partial P_i}{\partial \gamma} \right]^{-1}
\]

where the term \( \frac{1}{R} \) adjusts for the simulation error.

Remarks:

- An advantage of simulated methods of moments over simulated maximum likelihood methods is that SMM does not require that the number of draws \( R \) goes to infinity and in fact has been shown to perform well even for a modest number of draws.

- Simulated maximum likelihood (where the simulator is plugged directly into the likelihood instead of working with the FOC, does require that the number of draws go to infinity.
Chapter 4
Nonrandom sampling

References: Amemiya (Chapter 9 and 10), Manski and McFadden volume (Chapters 1 and 2), Manski and Lerman (1978 *Econometrica*), Amemiya

Different Types of Sampling

- random sampling
- censored sampling
- truncated sampling
- other nonrandom
  - truncated sampling
  - choice-based sampling

4.1 Some examples of choice-based sampling

- Study transportation mode choices and we gather data at the train, station, subway station, and at car checkpoints (e.g. toll booths).

- Study the effectiveness of a new treatment for heart attacks and we follow people who have had the treatment along with a control group of people who have not had the treatment.
• Evaluate the effects of a job training program and have data on a group of program participants and a group of nonparticipants. Usually participants are oversampled relative to their frequency in a random population.

In all of these cases, we observe characteristics of populations conditional on the choices they made.

There are different types of stratification:

(i) exogenous
(ii) endogenous

- For example, in administering surveys, there is often oversampling of high density population areas (often done to reduce sampling costs or to increase representation of certain groups). The sampling could be exogenous or endogenous, depending on the phenomenon being studied.

- Datasets usually provide sampling weights that are required to adjust for the nonrandom sampling

4.2 Choice-based sampling

Let’s look at how to use weights to adjust for choice-based sampling: Let $P_i = P(i \mid Z_i)$ in a random sample.

Let $P_i^*$ the corresponding probability in a choice-based sample.

Under choice-based sampling, sampling is random within $i$ partitions of the data

$$P(Z \mid i) = P^*(Z \mid i)$$

but

$$P(Z) \neq P^*(Z).$$
Suppose we want to recover $P(i \mid Z)$ from choice-based sampled data. We observe:

\[
P^*(i \mid Z) \\
P^*(Z) \\
P^*(i) \\
P^*(Z \mid i) = P(Z \mid i)
\]

Will apply Bayes’ rule:

\[
P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}
\]

By Bayes’ rule, write

\[
P^*(i \mid Z) = \frac{P^*(Z \mid i)P^*(i)}{P^*(Z)} \\
P(i \mid Z) = \frac{P(Z \mid i)P(i)}{P(Z)} \\
\Rightarrow \\
P(i \mid Z) = \frac{\left[ \frac{P^*(i \mid Z)P^*(Z)}{P^*(i)} \right] P(i)}{P(Z) = \sum_j P(Z \mid j)P(j)} \\
= \frac{\left[ \frac{P^*(i \mid Z)P^*(Z)}{P^*(i)} \right]}{\sum_j P^*(Z \mid j)P(j)} P(i) \\
= \frac{\left[ \frac{P^*(i \mid Z)P^*(Z)}{P^*(i)} \right]}{\sum_j P^*(j \mid Z)P^*(Z)P(j)} P(i) \\
= \frac{\left[ \frac{P^*(i \mid Z)P^*(Z)}{P^*(i)} \right]}{\sum_j P^*(j \mid Z)P^*(Z)P(j)}
\]

This shows that to recover $P(i \mid Z)$ from choice-based sampled data, we need to know $P(j)$ and $P^*(j)$ for all choices $j$. $P^*(j)$ can be estimated from the sampling, but $P(j)$ requires outside information (usually from some other dataset). Need to form the weights $\frac{P(i)}{P^*(i)}$. 

---

4.2. CHOICE-BASED SAMPLING

Suppose we want to recover $P(i \mid Z)$ from choice-based sampled data. We observe:

\[
P^*(i \mid Z) \\
P^*(Z) \\
P^*(i) \\
P^*(Z \mid i) = P(Z \mid i)
\]

Will apply Bayes’ rule:

\[
P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}
\]

By Bayes’ rule, write

\[
P^*(i \mid Z) = \frac{P^*(Z \mid i)P^*(i)}{P^*(Z)} \\
P(i \mid Z) = \frac{P(Z \mid i)P(i)}{P(Z)} \\
\Rightarrow \\
P(i \mid Z) = \frac{\left[ \frac{P^*(i \mid Z)P^*(Z)}{P^*(i)} \right] P(i)}{P(Z) = \sum_j P(Z \mid j)P(j)} \\
= \frac{\left[ \frac{P^*(i \mid Z)P^*(Z)}{P^*(i)} \right]}{\sum_j P^*(Z \mid j)P(j)} P(i) \\
= \frac{\left[ \frac{P^*(i \mid Z)P^*(Z)}{P^*(i)} \right]}{\sum_j P^*(j \mid Z)P^*(Z)P(j)} P(i) \\
= \frac{\left[ \frac{P^*(i \mid Z)P^*(Z)}{P^*(i)} \right]}{\sum_j P^*(j \mid Z)P^*(Z)P(j)}
\]

This shows that to recover $P(i \mid Z)$ from choice-based sampled data, we need to know $P(j)$ and $P^*(j)$ for all choices $j$. $P^*(j)$ can be estimated from the sampling, but $P(j)$ requires outside information (usually from some other dataset). Need to form the weights $\frac{P(i)}{P^*(i)}$. 

---
4.2.1 Manski and Lerman (1978)

Consider how to introduce weights to adjust for choice-based sampling in a standard probit model.

\textit{Binary choice model}

\[
D_i = \begin{cases} 
1 & \text{if } x_i \beta + \varepsilon_i > 0 \\
0 & \text{else} 
\end{cases}
\]

Denote choice-based sampling weights

\[
w_{1i} = \frac{P(D_i = 1)}{P^*(D_i = 1)} \\
w_{0i} = \frac{P(D_i = 0)}{P^*(D_i = 0)}
\]

Likelihood

\[
L = \prod_{i=1}^{n} [1 - \Phi(-x_i \beta)]^{w_{1i}D_i} \Phi(-x_i \beta)]^{w_{0i}(1-D_i)}
\]
Chapter 5

Limited Dependent Variable Models

We next consider a class of models where some of the outcome variables are partially observed, called limited dependent variable models.

5.1 Type I Tobit Model

Tobin’s (1958) example: expenditure on a durable good is only observed if good is purchased. Some people chose not to buy the good, but might have bought it if the price was lower.

Define a latent variable that represents desired expenditure amount:

\[ y_i^* = x_i \beta + u_i \]

Observe

\[ y_i = \begin{cases} y_i^* & \text{if } y_i^* \geq y_0 \\ 0 & \text{if } y_i^* < y_0 \end{cases} \]

where \( y_0 \) is the minimum price. Or, we can write

\[ y_i = 1(y_i \geq y_0) y_i^*. \]
CHAPTER 5. LIMITED DEPENDENT VARIABLE MODELS

Estimation: Assume

\[ u_i \sim N(0, \sigma^2) \]

\[ y_i^* | x_i \sim N(x_i \beta, \sigma_u^2) \]

\[ D_i = \begin{cases} 1 & \text{if } y_i^* \geq y_0, \\ 0 & \text{else} \end{cases} \]

The likelihood can be written as

\[ L = \prod_{i=1}^{n} \Pr(y_i^* < y_0)^{(1-D_i)} \{ \Pr(y_i^* \geq y_0) f(y_i^* | y_i^* \geq y_0) \}^{D_i}. \]

\[ \Pr(y_i^* < y_0) = \Pr \left( \frac{x_i \beta + u_i}{\sigma_u} < \frac{y_0}{\sigma_u} \right) = \Phi \left( \frac{y_0 - x_i \beta}{\sigma_u} \right) \]

\[ f(y_i^* | y_i^* \geq y_0) = \frac{1}{\sigma_u} \phi \left( \frac{y_i^* - x_i \beta}{\sigma_u} \right) \]

Now, note that the likelihood can be written as:

\[ L = \prod_{i=1}^{n} \Phi \left( \frac{y_0 - x_i \beta}{\sigma_u} \right) \prod_{i=1}^{n} \left\{ 1 - \Phi \left( \frac{y_0 - x_i \beta}{\sigma_u} \right) \right\} \prod_{i=1}^{n} \frac{1}{\sigma_u} \phi \left( \frac{y_i^* - x_i \beta}{\sigma_u} \right) \]

Remarks

(i) The first two products are the information that you would get from whether \( y_i^* \geq y_0 \) (i.e. whether bought the good or not). This information alone could be used to estimate \( \beta \) up to scale by a binary probit.

(ii) With the additional information on expenditure \( (y_i^*) \), can get a more efficient estimate of \( \beta/\sigma_u \) and an estimate of \( \sigma_u \).

5.1.1 Truncated Type I Tobit Model

Suppose you only observe \( y_i^* \) for \( y_i^* \geq y_0 \). In that case, the likelihood is:

\[ L = \prod_i \frac{1}{\sigma_u} \phi \left( \frac{y_i^* - x_i \beta}{\sigma_u} \right) \]

\[ 1 - \Phi \left( \frac{y_0 - x_i \beta}{\sigma_u} \right) \]
5.2 Different Ways of Estimating Tobit Models

(a) MLE, or

(b) If Censored, just estimate probit model, get $\frac{\beta}{\sigma_u}$

Then estimate OLS model on observations for which $y_i^*$ is observed. For example, with $y_0 = 0$:

$$E(y_i|x_i \beta + u_i \geq 0) = x_i \beta + E(u_i|x_i \beta + u_i \geq 0)$$
$$= x_i \beta + \sigma_u E\left(\frac{u_i}{\sigma_u} \cdot \frac{u_i}{\sigma_u} \geq -x_i \beta\right)$$

for $\varepsilon \sim N(0,1)$

$$\lambda(c) = E(\varepsilon|\varepsilon \geq c)$$
$$= \frac{\phi(c)}{1 - \Phi(c)} = \frac{\phi(c)}{\Phi(-c)}$$

$\lambda(c)$ known as the “mill’s ratio”

$$E(\varepsilon|\varepsilon < c) = \frac{-\phi(c)}{\Phi(c)}$$

$$E\left(\frac{u_i}{\sigma_u} \cdot \frac{u_i}{\sigma_u} \geq -x_i \beta\right) = \frac{\phi\left(\frac{-x_i \beta}{\sigma_u}\right)}{1 - \Phi\left(\frac{-x_i \beta}{\sigma_u}\right)} = \frac{\phi\left(\frac{\tilde{x_i \beta}{\sigma_u}}{\sigma_u}\right)}{\Phi\left(\frac{\tilde{x_i \beta}}{\sigma_u}\right)}$$

Two Step Estimation Procedure (Heckman, 1974, 1976)

(i) Estimate $\frac{\tilde{\beta}}{\sigma_u}$ by probit

(ii) Form $\tilde{\lambda}_i = \lambda\left(\frac{\tilde{x_i \beta}}{\sigma_u}\right)$

Regress

$$y_i = x_i \beta + \sigma \tilde{\lambda}_i + (v_i + \varepsilon_i)$$
$$v_i = \sigma \{\lambda_i - \tilde{\lambda}_i\}$$
$$\varepsilon_i = u_i - E(u_i|u_i > -x_i \beta)$$
CHAPTER 5. LIMITED DEPENDENT VARIABLE MODELS

Note that $\varepsilon_i$ has conditional mean equal to zero by construction. Here, $v_i$ is the additional error due to the fact that $\hat{\lambda}_i$ is estimated. The composite error term $(v_i + \varepsilon_i)$ will be heteroskedastic, which implies that weights can be used to improve efficiency. Need to account for the fact that $\hat{\lambda}_i$ is estimated in computing standard errors for $\hat{\beta}$. This is called an “estimated regressor” problem (also sometimes called the “Durbin problem.”

Two ways of doing this

- Delta method
- GMM (Newey, Economic Letters, 1984) - set up moment conditions for the OLS estimation and for the probit estimation program, with optimal weighting matrix. Estimating jointly will give corrected standard errors.

Correction for estimated regressor is incorporated into stata.

Suppose you used all the data, including the zeros,

$$E(y_i|x_i) = \Pr(y_i^* < 0) \cdot 0 + \Pr(y_i^* > 0)E(y_i^*|y_i^* \geq 0)$$

$$= \Phi \left( \frac{x_i\beta}{\sigma_u} \right) \left[ x_i\beta + \sigma_u \lambda \left( \frac{x_i\beta}{\sigma_u} \right) \right]$$

Could estimate by ols, replacing $\Phi$ with $\hat{\Phi}$ and $\lambda$ with $\hat{\lambda}$, or could do nonlinear least squares.

5.3 Type II Tobit Model

One outcome equation and one switching equation:

$$y_{1i}^* = x_{1i}\beta + u_{1i}$$
$$y_{2i}^* = x_{2i}\beta + u_{2i}$$
$$y_{2i} = y_{2i}^* \text{ if } y_{1i}^* \geq 0$$
$$= 0 \text{ else}$$
Here, we observe $x_{1i}$ and $x_{2i}$. $y^*_{1i}$ is not directly observed, although we observe whether $y^*_{1i} \geq 0$. $y^*_2$ is only observed if $y^*_{1i} \geq 0$.

Example: Suppose we wish to estimate the effect of school quality on test scores in a developing country. Quality can be measured in terms of school infrastructure, textbooks or teacher characteristics. We only observe test scores for children who choose to go to school and quality might also affect school-going behavior.

\[
\begin{align*}
y_{2i} &= \text{student’s test scores} \\
y_{1i} &= \text{index representing parent’s propensity to enroll child in school}
\end{align*}
\]

### 5.3.1 The likelihood:

\[
L = \prod_i \Pr(y^*_{1i} \geq 0) f(y_{2i} | y^*_{1i} \geq 0) \prod_i \Pr(y^*_{1i} < 0)
\]

where

\[
f(y_{2i} | y^*_{1i} \geq 0) = \frac{\int_0^\infty f(y^*_{1i}, y^*_{2i}) dy^*_{1i}}{\int_0^\infty f(y^*_{1i}) dy^*_{1i}}
\]

\[
= f(y^*_{2i}) \frac{\int_0^\infty f(y^*_{1i}, y^*_{2i}) dy^*_{1i}}{\int_0^\infty f(y^*_{1i}) dy^*_{1i}},
\]

where $f(y^*_{2i})$ and $f(y^*_{1i})$ are the marginal densities.

Assume

\[
\begin{align*}
y^*_{1i} &\sim N(x_{1i}\beta_1, \sigma^2_1) \\
y^*_{2i} &\sim N(x_{2i}\beta_2, \sigma^2_2)
\end{align*}
\]

\[
\begin{align*}
y^*_{1i} | y^*_{2i} &\sim N(x_{1i}\beta_1 + \frac{\sigma^2_{12}}{\sigma^2_2} (y^*_{2i} - x_{2i}\beta_2), \sigma^2_1 - \frac{\sigma^2_{12} \sigma^2_2}{\sigma^2_2}) \\
y^*_{2i} | y^*_{1i} &\sim N(x_{2i}\beta_2 + \frac{\sigma^2_{12}}{\sigma^2_1} (y^*_{1i} - x_{1i}\beta_1), \sigma^2_2 - \frac{\sigma^2_{12} \sigma^2_1}{\sigma^2_1})
\end{align*}
\]

*Estimation by MLE:*
CHAPTER 5. LIMITED DEPENDENT VARIABLE MODELS

\[ L = \prod_0 \left[ 1 - \Phi \left( \frac{x_{1i} \beta_1}{\sigma_1} \right) \right] \prod_1 \Phi \left( \frac{x_{1i} \beta_1}{\sigma_1} \right) \frac{f(y_{1i}^*)}{\int_0^\infty f(y_{1i}^*)dy_{1i}} \]

\[ \prod_0 \left[ 1 - \Phi \left( \frac{x_{1i} \beta_1}{\sigma_1} \right) \right] \times \]

\[ \prod_1 \left\{ \frac{\phi \left( \frac{y_{2i} - x_{2i} \beta_2}{\sigma_2} \right)}{\sigma_2} \right\} \left\{ 1 - \Phi \left( \frac{-\{x_{1i} \beta_1 + \frac{\sigma_{12}}{\sigma_2^2} (y_{2i} - x_{2i} \beta_2)\}}{\sigma^*} \right) \right\} \]

\[ \sigma^* = \sqrt{\frac{\sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}}{}} \]

The likelihood depends on \( \sigma_1 \) only through \( \beta_1 \sigma_1^{-1} \) and \( \sigma_{12}/\sigma_1^{-1} \), so we cannot separately identify \( \sigma_1 \) and WLOG can impose a normalization such as \( \sigma_1 = 1 \).

**Estimation by Two Step Approach**

Using data on \( y_{2i} \) for which \( y_{1i} > 0 \).

\[ E(y_{2i}|y_{1i}^* > 0) = x_{2i} \beta_2 + E(u_{2i}|x_{1i} \beta_1 + u_{1i} > 0) \]

\[ = x_{2i} \beta_2 + \sigma_2 E \left( \frac{u_{2i}}{\sigma_2} \left| \frac{u_{1i}}{\sigma_1} > \frac{-x_{1i} \beta_1}{\sigma_1} \right. \right) \]

\[ = x_{2i} \beta_2 + \frac{\sigma_{12}}{\sigma_1} E \left( \frac{u_{1i}}{\sigma_1} \left| \frac{u_{1i}}{\sigma_1} > \frac{-x_{1i} \beta_1}{\sigma_1} \right. \right) \]

\[ = x_{2i} \beta_2 + \frac{\sigma_{12}}{\sigma_1} \phi \left( \frac{-x_{1i} \beta_1}{\sigma_1} \right) \]

\[ = x_{2i} \beta_2 + \frac{\sigma_{12}}{\sigma_1} \phi \left( \frac{-x_{1i} \beta_1}{\sigma_1} \right) \]

\[ = x_{2i} \beta_2 + \frac{\sigma_{12}}{\sigma_1} \phi \left( \frac{-x_{1i} \beta_1}{\sigma_1} \right) \]

\[ = x_{2i} \beta_2 + \alpha \left( \frac{-x_{1i} \beta_1}{\sigma_1} \right) \]

\[ = x_{2i} \beta_2 + \alpha \left( \frac{-x_{1i} \beta_1}{\sigma_1} \right) \]
5.3. TYPE II TOBIT MODEL

5.3.2 Examples

Gronau’s Female Labor Supply Model

\[
\max_{C,L} U(C,L)
\]
\[
C = wH + A
\]
\[
H = 1 - L
\]
\[
w = \text{wage rate (given)}
\]
\[
P_c = 1
\]
\[
L = \text{time spent at home for child care}
\]

First-order condition:

\[
MRS = \frac{\partial U}{\partial L} = \frac{\partial U}{\partial C} = w \text{ when } L < 1
\]
\[
\frac{\partial U}{\partial L} > \frac{\partial U}{\partial C} > w \text{ if } L = 1
\]

Define reservation wage

\[
w^R = MRS |_{H=0}.
\]

We don’t observe reservation wage directly, so model

\[
w^o = x \beta + u \quad \text{wage offer}
\]
\[
w^R = z \gamma + v \quad \text{reservation wage}.
\]

Only observe the wage offer if the offer exceeds the reservation wage:

\[
w_i = w_i^o \text{ if } w_i^o > w_i^R
\]
\[
= 0 \text{ else}.
\]

This model fits within the Tobit Type II framework if we set

\[
y_{1i}^* = x_i \beta - z_i \gamma + u_i - v_i = w_i^o - w_i^R
\]
\[
y_{2i}^* = w_i^o
\]

Gronau did not develop a model to explain the choice of \(H\) (hours worked).
5.4 Type III Tobit: Heckman’s Model (Incorporates Choice of H)

Adopts a functional form for the utility function that implies a functional form relationship between the marginal rate of substitution and hours worked:

\[ w_i^o = x_{2i}^\beta_2 + u_{2i} \]
\[ MRS = \gamma H_i + z_i^\alpha + v_i \]
\[ MRS |_{H=0} = w^R = z_i^\alpha + v_i \]

An individual works if

\[ w_i^o > w_i^R, \text{ or } \]
\[ x_{2i}^\beta_2 + u_{2i} > z_i^\gamma + v_i \]

If works, then

\[ w_i^o = MRS \]
\[ \implies x_{2i}^\beta_2 + u_{2i} = \gamma H_i + z_i^\alpha + v_i \]
\[ \implies H_i = \frac{x_{2i}^\beta_2 - z_i^\alpha + u_{2i} - v_i}{\gamma} \]
\[ = x_{1i}^\beta_1 + u_{1i} \]

where

\[ x_{1i}^\beta_1 = (x_{2i}^\beta_2 - z_i^\alpha)^{-1} \]
\[ u_{1i} = (u_{2i} - v_i)^{-1} \]

Observe both hours and wages if hours worked positive:

\[ y_{1i}^* = x_{1i}^\beta_1 + u_{1i} \text{ hours } (= H_i) \]
\[ y_{2i}^* = x_{2i}^\beta_2 + u_{2i} \text{ wages } (= w_i^*) \]
\[ y_{1i} = y_{1i}^* \text{ if } y_{1i}^* > 0 \]
\[ = 0 \text{ if } y_{1i}^* \leq 0 \]
\[ y_{2i} = y_{2i}^* \text{ if } y_{1i}^* > 0 \]
\[ = 0 \text{ if } y_{1i}^* \leq 0. \]
5.5 Type IV Tobit Model

Adds another equation

\[ y_{3i} = \begin{cases} y_{3i}^* & \text{if } y_{1i}^* > 0 \\ 0 & \text{if } y_{1i}^* \leq 0. \end{cases} \]

Can estimate these models by

(i) maximum likelihood

(ii) Two-step method

\[ E(u_i^c | H_i > 0) = \gamma H_i + z_i \alpha + E(v_i | H_i > 0) \]

5.6 Type V Tobit: Switching Regression Model

\[ \begin{align*}
  y_{1i}^* &= x_{1i}' \beta_1 + u_{1i} \\
  y_{2i}^* &= x_{2i}' \beta_1 + u_{2i} \\
  y_{3i}^* &= x_{3i}' \beta_1 + u_{3i} \\
  y_{2i} &= y_{2i}^* & \text{if } y_{1i}^* > 0 \\
  &= 0 & \text{if } y_{1i}^* \leq 0 \\
  y_{3i} &= y_{3i}^* & \text{if } y_{1i}^* \leq 0 \\
  &= 0 & \text{if } y_{1i}^* > 0 \quad i = 1, \ldots, n
\end{align*} \]

Only the sign of \( y_{1i}^* \) is observed.

Assume \((u_{1i}, u_{2i}, u_{3i})\) are drawings from a trivariate normal distribution. The likelihood can be written as,

\[ L = \Pi_0 \int_{-\infty}^{0} f_3(y_{1i}^*, y_{3i})dy_{1i}^* \Pi_1 \int_{-\infty}^{0} f_2(y_{1i}^*, y_{3i})dy_{1i}^* \]

where \(f_3(y_{1i}^*, y_{3i})\) is the joint density of \(y_{1i}^*\) and \(y_{3i}^*\) and \(f_2(y_{1i}^*, y_{2i})\) is the joint density of \(y_{1i}^*\) and \(y_{2i}^*\).
5.6.1 Examples

Lee (1978) unionism model

In goal of Lee’s application is to estimate the effect of unions on worker wages. $y^*_2$ represents the log of the wage rate if the ith worker joins a union, and $y^*_3$ represents the wage rate if don’t join a union.

$$y^*_1 = y^*_2 - y^*_3 + z_i \alpha + v_i,$$

where $z_i \alpha$ captures costs (monetary and psychic) of joining a union. This set-up is close to the Roy model, which we will consider later.

Heckman (1978) model of effects of antidiscrimination laws

Used to estimate the effect of antidiscrimination laws on income of African Americans in a way that takes into account that passage of the law may be endogenous. Let

$$y_{2i} = \text{average income of African Americans within a state}$$

$$y^*_1 = \text{unobservable sentiment toward blacks in the ith state}$$

$$w_i = 1 \text{ if } y^*_1 > 0$$

$$= 0 \text{ if } y^*_1 \leq 0$$

$$y^*_1 = \gamma_1 y_{2i} + x'_1 \beta_1 + \delta_1 w_i + u_{1i}$$

$$y_{2i} = \gamma_2 y^*_1 + x'_2 \beta_2 + \delta_2 w_i + u_{2i}$$

If we solve for $y^*_1$ the solution should not depend on $w_i$, which requires the assumption that

$$\gamma_1 \delta_2 + \delta_1 = 0$$

for the model to be logically consistent.
Chapter 6

The Roy Model

• The Roy model is a two-sector econometric model of self-selection that has been very influential in economics, especially in labor economics.

• It was first developed by A.D. Roy (1951) in his analysis of earnings in two occupational sectors, in which individuals self-select into the sector with the highest earnings. In Roy’s original example, the two occupations were hunting and fishing.

• Roy assumed that individuals have some endowed skill in both sectors and that they choose to work in the sector that would give them the highest earnings.

• Although originally developed in studying occupational choice, the model has also been used to study the effects of unions on earnings, the effects of education (high school or college) on earnings and the effects of federal job training programs.

• Heckman and Honore (1990) reconsider the Roy model and derive a number of new results, generalizing it to non-normal distributions.

6.1 Model

To describe Roy’s set-up, let \((F, H)\) denote an individual’s endowment of skills (talent) in the two occupations, fishing and hunting. It is assumed that skills are log normally distributed.
CHAPTER 6. THE ROY MODEL

Distribution of skills

- Talent matters more in occupation with higher variance in skills

Denote the earnings for fisherman and hunters by:

\[ y_i^F = F_i p_F \]
\[ y_i^H = H_i p_H \]

where \( p_F \) is the unit price of fish and \( p_H \) the unit price of hunting output.

It is assumed that the individual chooses the occupation that maximizes their earnings, so the choice rule is:

\[ y_i = \max\{y_i^F, y_i^H\} \]

An individual chooses fishing if

\[ F_i p_F > H_i p_H \]

and chooses hunting if

\[ H_i p_H > F_i p_F. \]

The line of indifference is:

\[ F_i = \frac{p_H}{p_F} H_i \]

Let \( g(F_i, H_i) \) denote the distribution of skills in the population (a two dimensional density).
6.2. MEAN EARNINGS GIVEN SELF-SELECTION

We can ask, what is the average amount of fish caught by people who choose the fishing occupation?

\[
E(F_i|y_i^F > y_i^H) = E(F_i|F_i > \frac{p_H}{p_F}H_i) = \frac{\int_0^\infty \int_0^\infty F_i g(F_i, H_i) dF_i dH_i}{\int_0^\infty \int_0^\infty g(F_i, H_i) dF_i dH_i}
\]

The total supply of fish in the population will be

\[
E(F_i|F_i > \frac{p_H}{p_F}H_i) Pr(y_i^F > y_i^H) + 0 Pr(y_i^F < y_i^H) = E(F_i|F_i > \frac{p_H}{p_F}H_i) \int_0^\infty \int_0^\infty g(F_i, H_i) dF_i dH_i
\]

as the price of fish increases, more people will choose the fishing occupation and the supply of fish will increase.

A similar expression can be derived for the average amount of hunting output for people who choose hunting as their profession and for the total amount of hunting output in the economy.

6.2 Mean earnings given self-selection

Next, we will analyze the implications of self-selection into occupations for mean earnings in that occupation. We will also consider how to recover the parameters of the underlying skill distribution. Let \( Y_0 \) denote log earnings in the fishing occupation, which is assumed to depend on some observed characteristics \( X \) and some unobservables \( \varepsilon_0 \). Also, let \( Y_1 \) denote log earnings in the fishing occupation, which is assumed to depend on some observed characteristics \( X \) and some unobservables \( \varepsilon_1 \).

\[
Y_{0i} = \ln y_i^F = \ln p_i^F + \ln F_i = X_i \beta_0 + \varepsilon_{0i}
\]

\[
Y_{1i} = \ln y_i^H = \ln p_i^H + \ln H_i = X_i \beta_1 + \varepsilon_{1i}
\]

The prices will enter into the intercepts of the log earnings equation (the constant terms). We will assume that the residuals are joint normally distributed.

\[
\begin{pmatrix}
\varepsilon_{0i} \\
\varepsilon_{1i}
\end{pmatrix}
\sim N
\left(
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
\sigma_0^2 & \sigma_{01} \\
\sigma_{10} & \sigma_1^2
\end{pmatrix}
\right)
\]
Define the correlation coefficient:

\[ \rho = \frac{\sigma_{10}}{\sqrt{\sigma_1^2 \sqrt{\sigma_0^2}}} = \frac{\text{sigma}_{10}}{\sigma_1 \sigma_0} \leq \sigma_{10} = \rho \sigma_0 \sigma_1 \]

To recover the underlying skill distribution in the population, we need to recover the parameters \( \beta_0, \beta_1, \sigma_0^2, \sigma_1^2, \) and \( \sigma_{10}. \)

Define \( D = 1 \) if an individual chooses occupation \( H, \) else \( D = 0. \)

\[
Prob(D = 1|X) = Prob(Y_1 > Y_0|x)
= Prob(X\beta_1 + \varepsilon_1 \geq X\beta_0 + \varepsilon_0)
= Prob\left(\frac{\varepsilon_0 - \varepsilon_1}{\sigma} \leq \frac{X(\beta_1 - \beta_0)}{\sigma}\right)
\]

Define \( \xi = \frac{\varepsilon_0 - \varepsilon_1}{\sigma} \) and \( \sigma = \sqrt{\text{var}(\varepsilon_0 - \varepsilon_1)} = \sqrt{\sigma_0^2 + \sigma_1^2 - 2\sigma_{10}} \)

If occupations were chosen or assigned at random (without regard to earnings potential), then average earnings would be

\[
E(Y_0|X) = X\beta_0,
E(Y_1|X) = X\beta_1
\]

Now, ask, what are average earnings in an occupation, given the fact that individuals choose the occupation that maximizes their earnings. First, let’s consider earnings in section 1 for those who chose sector 1:

\[
E(Y_1|X, D = 1) = X\beta_1 + \sigma_1 E\left(\frac{\varepsilon_1}{\sigma_1}|\xi \leq \frac{X(\beta_1 - \beta_0)}{\sigma}\right)
\]

To derive an expression for the last term, we will use the fact that for normally distributed mean zero random variables, the projection of \( U \) on \( V \) is

\[
U = \frac{\text{Cov}(U,V)}{\text{Var}(V)} + \eta
\]

Define \( \xi = \frac{\varepsilon_0 - \varepsilon_1}{\sigma} \)
6.2. MEAN EARNINGS GIVEN SELF-SELECTION

$$E(Y_1|X) = X\beta_1 + \sigma_1 E(\frac{\varepsilon_1}{\sigma_1} | \xi \leq \frac{X(\beta_1 - \beta_0)}{\sigma})$$

$$= X\beta_1 + \frac{\sigma_1 \sigma_0 - \sigma_1^2}{\sigma} E(\xi|\xi \leq \frac{X(\beta_1 - \beta_0)}{\sigma})$$

$$= X\beta_1 + \frac{\rho \sigma_1 \sigma_0 - \sigma_1^2}{\sigma} E(\xi|\xi \leq \frac{X(\beta_1 - \beta_0)}{\sigma})$$

The term $E(\xi|\xi \leq \frac{X(\beta_1 - \beta_0)}{\sigma}) < 0$, because it is a standard normal random variable that is truncated from above.\(^1\) We can rewrite the expression as

$$E(Y_1|X, D = 1) = X\beta_1 + \frac{\sigma_1 \sigma_0}{\sigma} (\rho - \frac{\sigma_1}{\sigma_0}) E(\xi|\xi \leq \frac{X(\beta_1 - \beta_0)}{\sigma})$$

We have that

$$E(Y_1|X, D = 1) > E(Y_1) \text{ if } \rho < \frac{\sigma_1}{\sigma_0}$$

$$E(Y_1|X, D = 1) < E(Y_1) \text{ if } \rho > \frac{\sigma_1}{\sigma_0}$$

Similarly, we can obtain

$$E(Y_0|X, D = 0) = X\beta_0 + \frac{\sigma_1 \sigma_0}{\sigma} (\frac{\sigma_0}{\sigma_1} - \rho) E(\xi|\xi \geq \frac{X(\beta_1 - \beta_0)}{\sigma})$$

where $E(\xi|\xi \geq \frac{X(\beta_1 - \beta_0)}{\sigma}) > 0$, because it is the mean of a standard normal random variable truncated from below.

From these expressions, we see that

$$E(Y_0|D = 0, X) > E(Y_0|X) \text{ if } \rho < \frac{\sigma_0}{\sigma_1}$$

$$E(Y_0|D = 0, X) < E(Y_0|X) \text{ if } \rho > \frac{\sigma_0}{\sigma_1}$$

\(^1\)The mean without truncation is zero, $E(\varepsilon) = 0$, so $E(\varepsilon < c) \leq 0.$
Can we have both $E(Y_{1}\mid X, D = 1) < E(Y_{1}\mid X)$ and $(E(Y_{0}\mid X, D = 0) < E(Y_{0}\mid X)$? No,

$$\text{var}(\xi) = \sigma_{1}^{2} + \sigma_{0}^{2} - 2\sigma_{10}$$

which implies that either

$$\frac{\sigma_{1}^{2}}{\sigma_{1}, \sigma_{0}} > \frac{\sigma_{10}}{\sigma_{1}, \sigma_{0}} \text{ or } \frac{\sigma_{0}^{2}}{\sigma_{1}, \sigma_{0}} > \frac{\sigma_{10}}{\sigma_{1}, \sigma_{0}}$$

or both conditions hold

$$\Leftrightarrow$$

$$\rho < \frac{\sigma_{0}}{\sigma_{0}} \text{ or } \rho < \frac{\sigma_{1}}{\sigma_{1}}$$

or both.

### 6.3 Estimation

How do we estimate this model? Write

$$Y_{1} = X\beta_{1} + \sigma_{1}E(\frac{\varepsilon_{1}}{\sigma_{1}}|\varepsilon_{0} - \varepsilon_{1}) \leq \frac{X(\beta_{1} - \beta_{0})}{\sigma}) + v_{1}$$

$$v_{1} = \varepsilon_{1} - E(\varepsilon_{1}\mid X, D = 1)$$

$$Y_{1} = X\beta_{1} + \sigma_{10} - \sigma_{1}^{2} - \phi(\frac{X(\beta_{1} - \beta_{0})}{\sigma}) + v_{1}$$

Let
6.3. ESTIMATION

\[ \lambda_1 \left( \frac{X(\beta_1 - \beta_0)}{\sigma} \right) = -\phi \left( \frac{X(\beta_1 - \beta_0)}{\sigma} \right) \Phi \left( \frac{X(\beta_1 - \beta_0)}{\sigma} \right) \]

Then

\[ Y_1 = X \beta_1 + \alpha_1 \lambda_1 + v_1 \]

Similarly, for \( D = 0 \), write

\[ Y_0 = X \beta_0 + \sigma_0 E \left( \frac{\varepsilon_0}{\sigma_0} \frac{\varepsilon_0 - \varepsilon_1}{\sigma} > \frac{X(\beta_1 - \beta_0)}{\sigma} \right) + v_0 \]

\[ v_0 = \varepsilon_0 - E(\varepsilon_0 | X, D = 0) \]

\[ Y_0 = X \beta_0 + \frac{\sigma_0^2 - \sigma_{10}}{\sigma} \phi \left( \frac{X(\beta_1 - \beta_0)}{\sigma} \right) 1 - \Phi \left( \frac{X(\beta_1 - \beta_0)}{\sigma} \right) + v_0 \]

Let

\[ \lambda_0 = \left( \frac{X(\beta_1 - \beta_0)}{\sigma} \right) = \frac{\phi \left( \frac{X(\beta_1 - \beta_0)}{\sigma} \right)}{1 - \Phi \left( \frac{X(\beta_1 - \beta_0)}{\sigma} \right)} \]

Then

\[ Y_0 = X \beta_0 + \alpha_0 \lambda_0 + v_0 \]

Heckman (1976) proposed a two-step estimation approach for this model.

Step 1: First estimate a probit model for \( Pr(D = 1|X) \) and \( Pr(D = 0|X) \). From that, get estimates of \( X(\hat{\beta}_1 - \hat{\beta}_0) \) and form \( \hat{\lambda}_1 \) and \( \hat{\lambda}_0 \).

Step 2: Next estimate the \( Y_1 \) and \( Y_0 \) regressions, including \( \hat{\lambda}_1 \) and \( \hat{\lambda}_0 \) as “control functions.” From this, we get estimates of \( \beta_1, \beta_0, \alpha_1 \) and \( \alpha_0 \).

To separately identify \( \sigma_1^2, \sigma_0^2 \), and \( \sigma_{10} \), we would also have to estimate models for
Var\( (Y_0|D = 1, X) \)
Var\( (Y_1|D = 0, X) \)

and use the fact that for a normal random variable

\[
\text{Var}(u|u > c) = 1 - \frac{\phi(c)}{\Phi(-c)}c - \frac{\phi(c)}{\Phi(-c)}^2.
\]

### 6.4 Constructing counterfactuals

Once the parameters described above are estimated, we can ask what average earnings would be for sector 0 persons had they chose sector 1 instead.

\[
E(Y_1|X, D = 0) = X\hat{\beta}_1 + E(\varepsilon_1|X, X\beta_0 + \varepsilon_0 > X\beta_1 + \varepsilon_1)
\]
\[
E(Y_1|X, D = 0) = X\beta_1 + \sigma_1 E\left(\frac{\varepsilon_0 - \varepsilon_1}{\sigma} \sigma > \frac{X(\beta_1 - \beta_0)}{\sigma}\right)
\]
\[
E(Y_1|X, D = 0) = X\beta_1 + \frac{\sigma_{10} - \sigma_1^2}{\sigma} E(\xi|\xi > X(\beta_1 - \beta_0))
\]
\[
E(Y_1|X, D = 0) = X\beta_1 + \frac{\sigma_{10} - \sigma_1^2}{\sigma} \frac{\phi\left(\frac{X(\beta_1 - \beta_0)}{\sigma}\right)}{1 - \Phi\left(\frac{X(\beta_1 - \beta_0)}{\sigma}\right)}
\]

which we can estimate by

\[
X\hat{\beta}_1 + \hat{\alpha}_1 - \frac{\phi\left(\frac{X(\beta_1 - \beta_0)}{\sigma}\right)}{1 - \Phi\left(\frac{X(\beta_1 - \beta_0)}{\sigma}\right)}
\]

Similarly, average earnings for sector 1 persons had they chose sector 0
6.4. CONSTRUCTING COUNTERFACTUALS

\[
E(Y_0 | X, D = 1) = X\beta_0 + E(\varepsilon_0 | X, X\beta_0 + \varepsilon_0 \leq X\beta_1 + \varepsilon_1)
\]
\[
E(Y_0 | X, D = 1) = X\beta_0 + \sigma_0 E(\frac{\varepsilon_0}{\sigma_0} | \varepsilon_0 - \varepsilon_1 \leq \frac{X(\beta_1 - \beta_0)}{\sigma})
\]
\[
E(Y_0 | X, D = 1) = X\beta_0 + \frac{\sigma_0^2 - \sigma_{10}}{\sigma} E(\xi | \xi \leq \frac{X(\beta_1 - \beta_0)}{\sigma})
\]
\[
E(Y_0 | X, D = 1) = X\beta_0 + \frac{\sigma_0^2 - \sigma_{10}}{\sigma} - \phi\left(\frac{X(\beta_1 - \beta_0)}{\sigma}\right) - \Phi\left(\frac{X(\beta_1 - \beta_0)}{\sigma}\right)
\]
which we can estimate by

\[
X\hat{\beta}_0 + \hat{\alpha}_0 - \phi\left(\frac{X(\hat{\beta}_1 - \hat{\beta}_0)}{\sigma}\right) - \Phi\left(\frac{X(\hat{\beta}_1 - \hat{\beta}_0)}{\sigma}\right)
\]

The Roy model is very powerful in that it provides a way of inferring the distribution of underlying skills even though earnings are only observed for the self-selected group of individuals who chose a particular sector. The model makes some strong assumptions, though, and recent work has focused on relaxing some of those assumptions. Namely, the model assumes

- That agents choose sectors without incurring pecuniary costs
- That there is no uncertainty in the future rewards from choosing a sector.
- That the \(Y_0\) and \(Y_1\) are normally distributed.

Rosen and Willis (1979) present an application of an extended version of the Roy model to studying the decision to go to college, where the two sectors represent stopping at a high school degree or going on to college. One of the focuses of their study is on the parameter \(\sigma_{10}\), which tells us whether people who have unobserved skills that make them higher earners without college also tend to be higher earners with college (conditional on \(X\)). They find evidence of comparative advantage, that individuals who go to college would not tend to be the highest earners had they stopped at high school and vice versa. The Roy model has also been applied to study the decision to go to public or private school, the decisions to get a union wage job or not, and the decision to go to a job training program.