Models of Discrete Choice

Jonathan Wand

Polisci 350C
Stanford University

April 10, 2006 (revised)
Theoretical foundations of Choice models

Two main approaches to deriving models of discrete choice:

1. Discriminal process (Thurstone, 1927)
   Most often associated with models of choice based on normal distributions (probit)

2. Axiomatic derivation (Luce, 1959)
   Most often associated with models of choice based on logistic distribution (logit)
Thurstone: discriminial process

Logic of the discriminial process (abstract)

1. Assume a metric for making comparisons across items. *Thurstone deals with a single attribute, posits a psychological continuum*

2. Assume that attributes of items have a random component. *Thurstone used Normal distribution, but acknowledged arbitrariness.*

3. Assume that positive difference between two items leads to the selection of the item with the higher value. *A relative comparison, requiring only simple inequality statements.*
Thurstone: discriminable process

Logic of the discriminable process (less abstract)

1. The utility for item \( i \) is defined as \( u_i \).
2. The utility can be decomposed into two parts,
   - a systematic (observed) component \( \mu_i \)
   - a random (unobserved) component \( \epsilon_i \sim F(0, \sigma^2) \)

Assume that the overall utility of an item is the sum of the observed and unobserved components \( u_i = \mu_i + \epsilon_i \).
3. For a pair of items, choose item \( j \) over \( k \) if \( u_j > u_k \).
Thurstone: discriminial process

General binary choice

\[ P(j, k) = P(u_j > u_k) \]
\[ = P(\mu_j + \epsilon_j > \mu_k + \epsilon_k) \]
\[ = P(\mu_j - \mu_k + \epsilon_j > \epsilon_k) \]
\[ = P(\mu_j - \mu_k > \epsilon_k - \epsilon_j) \]
Notational aside

Define the standard Normal(0,1) pdf and cdf as, respectively,

\[
\phi(\epsilon) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{\epsilon^2}{2} \right\}
\]

\[
\Phi(\epsilon) = \int_{-\infty}^{\epsilon} \phi(z) \, dz
\]

Define the Type I extreme value pdf and cdf as, respectively,

\[
\lambda(\epsilon) = e^{-\epsilon} \exp\{ -e^{-\epsilon} \}
\]

\[
\Lambda(\epsilon) = \exp\{ -e^{-\epsilon} \}
\]
Thurstone: discriminable process

Case 1. Assume $F(0, \sigma_i^2)$ is a Normal distribution

Recall, for independent normal distributions

$$
\mathcal{N}(\mu_i, \sigma_i^2) - \mathcal{N}(\mu_j, \sigma_j^2) = \mathcal{N}(\mu_i - \mu_j, \sigma_i^2 + \sigma_j^2)
$$

So,

$$
P(j, k) = P(\mu_j - \mu_k > \epsilon_k - \epsilon_j)
$$

$$
= \int_{-\infty}^{\mu_j - \mu_k} \frac{1}{\sqrt{2\pi(\sigma_j^2 + \sigma_k^2)}} \exp\left\{-\frac{z^2}{2\sqrt{\sigma_j^2 + \sigma_k^2}}\right\} \, dz
$$

$$
= \Phi\left(\frac{\mu_j - \mu_k}{\sqrt{\sigma_j^2 + \sigma_k^2}}\right)
$$
Thurstone: discriminant process

Given,

\[ P(j, k) = \Phi \left( \frac{\mu_j - \mu_k}{\sqrt{\sigma_j^2 + \sigma_k^2}} \right) \]

and \( \mu_i = x_i \gamma_i \), consider what can be identified?

If \( x_j = x_k = x \), then

\[ P(j, k) = \Phi \left( x \frac{\gamma_j - \gamma_k}{\sqrt{\sigma_j^2 + \sigma_k^2}} \right) = \Phi (x \beta) \]

where \( \beta = \frac{\gamma_j - \gamma_k}{\sqrt{\sigma_j^2 + \sigma_k^2}} \).
Thurstone: discriminant process

Given,

\[ P(j, k) = \Phi \left( \frac{\mu_j - \mu_k}{\sqrt{\sigma_j^2 + \sigma_k^2}} \right) \]

and \( \mu_i = x_i \gamma_i \), consider what can be identified?

If \( \gamma_j = \gamma_k = \gamma \), then

\[ P(j, k) = \Phi \left( \frac{(x_j - x_k) \gamma}{\sqrt{\sigma_j^2 + \sigma_k^2}} \right) = \Phi \left((x_j - x_k) \beta\right) \]

where \( \beta = \frac{\gamma}{\sqrt{\sigma_j^2 + \sigma_k^2}} \).
Thurstone: discriminative process

Case 2. Assume $F$ is a Type I extreme value distribution

\[
P(j, k) = P(u_j > u_k) \\
= P(\mu_j - \mu_k + \epsilon_j > \epsilon_k) \\
= \int_{-\infty}^{\infty} \lambda(\epsilon_j) \int_{-\infty}^{\mu_j - \mu_k + \epsilon_j} \lambda(\epsilon_k) \\
= \int_{-\infty}^{\infty} \lambda(\epsilon_j) \Lambda(\mu_j - \mu_k + \epsilon_j) \\
= \int_{-\infty}^{\infty} \lambda(\epsilon_j) \Lambda(\mu_j - \mu_k + \epsilon_j) \\
= \frac{1}{w} \int_{-\infty}^{\infty} -e^{-\epsilon_j w} \exp\{-e^{-\epsilon_j w}\} \\
= \frac{1}{w} \frac{1}{1 + \exp\{-(\mu_1 - \mu_2)\}}
\]
Thurstone: discriminial process

Given,

\[ P(j, k) = \frac{1}{1 + \exp\{-(\mu_1 - \mu_2)\}} \]

and \( \mu_i = x_i \gamma_i \), consider what can be identified?

If \( x_j = x_k = x \), then

\[ P(j, k) = \frac{1}{1 + \exp\{-x(\gamma_1 - \gamma_2)\}} = \frac{1}{1 + \exp\{-x \beta\}} \]

where \( \beta = \gamma_1 - \gamma_2 \).

If \( \gamma_j = \gamma_k = \gamma \), then

\[ P(j, k) = \frac{1}{1 + \exp\{-\beta(x_1 - x_2)\}} \]

where \( \beta = \gamma \).
Axiomatic Foundations of Choice Models

Assumptions

D1 Let $R \subset S \subset T \subset U$.

D2 Let $x, y, z \in T$, arbitrary elements of choice set.

D3 Let $P(x, y)$ be the probability of choosing $x$ instead of $y$, $0 < P(x, y) < 1$.

D4 $P_S(R)$ is the probability of choosing $R$ given choice from among alternatives in $S$.

Choice Axiom

(i) $P_T(R) = P_S(R)P_T(S)$

(ii) If $P(x, y) = 0$ for some $x, y \in T$, $P_T(S) = P_{T-\{x\}}(S - \{x\})$
Axiomatic Foundations of Choice Models

Axiom of Choice

• Defines relationship which defines how choices within subsets are related in the context of an individual making probabilistic choices.

• Can be rewritten as $P_T(R | S)P_T(S) = P_T(R)$

• Two core implications,
  Lemma 3: Independence of Irrelevant Alternatives (IIA)
  Theorem 3: Probability must satisfy a ratio scale
Axiomatic Foundations of Choice Models

Lemma 3 (Independence from irrelevant alternatives):
For \( x, y \in S \),

\[
\frac{P(x, y)}{P(y, x)} = \frac{P_S(x)}{P_S(y)}
\]

Proof:
By Axiom we have

\[
P_S(x) = P(x, y)[P_S(x) + P_S(y)]
\]

So

\[
P_S(x) = P(x, y)[P_S(x) + P_S(y)]
\]
\[
P_S(x) = P(x, y)P_S(x) + P(x, y)P_S(y)
\]
\[
(1 - P(x, y))P_S(x) = P(x, y)P_S(y)
\]
\[
P(y, x)P_S(x) = P(x, y)P_S(y)
\]
\[
P(x, y) = P_S(x)
\]
\[
P(y, x) = P_S(y)
\]
Axiomatic Foundations of Choice Models

Lemma 3 (Independence from irrelevant alternatives):

- Most famous implication: relative probability of choosing two alternatives is invariant to the composition of the larger set of alternatives.
- Again, only ratio that is invariant not probabilities themselves.
- Might also hear that log-odds of two choices are constant: \( \log(P_S(x)) - \log(P_S(y)) = c \).
- Can estimate parameters defining utility of choices even with only a subset. Not possible if IIA does not hold (e.g., correlation between utilities of choices)—then to estimate any choice must model all choices.
Axiomatic Foundations of Choice Models

Theorem 3: choice probability is ratio scale

$\exists v : T \rightarrow \mathbb{R}_+$, unique up to multiplication by $k > 0$, such that

$$P_S(x) = \frac{v(x)}{\sum_{y \in S} v(y)} = \frac{1}{1 + \sum_{y \in S \setminus \{x\}} v(y)/v(x)}$$

McFadden and Yellot have each pointed out the connection to logit models by setting $v(x) = e^x$.

E.g.,

$$P(x, y) = \frac{v(x)}{v(x) + v(y)} = \frac{e^{\mu x}}{e^{\mu x} + e^{\mu y}} = \frac{1}{1 + e^{\mu y}/e^{\mu x}} = \frac{1}{1 + e^{-(\mu x - \mu y)}}$$

Yellot (1977) shows that discriminial process based on Type I discrete value distribution is uniquely equivalent to Choice Axiom.
Likelihood for a dichotomous choice

Consider a Bernoulli process for observing \( y \in \{0, 1\} \), where \( P(y = 1) = p \), and therefore \( P(y = 0) = 1 - p \).

For an individual \( i \), the likelihood of observing \( y_i \) is

\[
\mathcal{L}_i = p^{y_i}(1 - p)^{1-y_i}
\]  

(1)

The log-likelihood,

\[
L_i = y_i \log(p) + (1 - y_i) \log(1 - p)
\]  

(2)

The score of the likelihood

\[
\frac{L_i}{p} = \frac{y_i}{p} + \frac{1 - y_i}{1 - p}(-1) = \frac{y_i - p}{p(1 - p)} = \frac{y_i - E(y_i)}{V(y_i)}
\]  

(3)

Here, \( p \) is a single parameter for all people, but same setup for more general parameterization.
Likelihood for a dichotomous choice

Normally, would like \( p \) to be a function of covariates \( x_i \).

Therefore we seek a mapping: \( x_i \rightarrow P(y_i \mid x_i) \).

In general, will need to specify two things:

1. **Index function:** \( a(x, \beta) \) that maps k-vector to scalar
   - 1.1 e.g., additive-linear: \( a(x_i, \beta) = x_i \beta \)
   - 1.2 e.g., non-linear \( a(x_i, \beta) = x_1^{\beta_1} + x_2 \beta_2 \)

2. **Transformation function** \( F() \) with properties,
   - 2.1 \( F(-\infty) = 0 \)
   - 2.2 \( F(\infty) = 1 \)
   - 2.3 \( \frac{\partial F(z)}{\partial z} = f(z) > 0 \)

i.e., monotonic function that maps real line to unit interval
Likelihood for a dichotomous choice

Standard parameterizations for k-vector of covariates \( x_i \),

1. Probit:
   1.1 \( a(x, \beta) = x^{\beta} \)
   1.2 \( F(x^{\beta}) = \int_{-\infty}^{x^{\beta}} \frac{1}{\sqrt{2\pi}} \exp\{- \frac{z^2}{2}\} = \int_{-\infty}^{x^{\beta}} \phi(z) = \Phi(x^{\beta}) \)
   1.3 \( L_i = y_i \log(\Phi(x_i^{\beta})) + (1 - y_i) \log(1 - \Phi(x_i^{\beta})) \)

2. Logit
   2.1 \( a(x, \gamma) = x^{\gamma} \)
   2.2 \( F(x^{\gamma}) = \frac{1}{1 + \exp\{-x^{\gamma}\}} = \Lambda(x^{\gamma}) \)
   2.3 \( L_i = y_i \log(\Lambda(x_i^{\gamma})) + (1 - y_i) \log(1 - \Lambda(x_i^{\gamma})) \)
Logit and Probit, CDF

\[ P(y=1 | x) \]

\[ \beta = (0,1) \]
Logit and Cauchit (t), CDF

\[ P(y=1 \mid x) \]

\[ \text{beta} = (0,1) \]
Likelihood for a dichotomous choice

Let \( p_i = P(y_i \mid x_i) = F(a(x_i, \beta)) \).

Recall the log-likelihood,

\[
L_i = y_i \log(p) + (1 - y_i) \log(1 - p)
\]

and re-write the score of the log-likelihood,

\[
S_i(\beta) = \frac{\partial L_i}{\partial \beta} = \frac{\partial L_i}{\partial p_i} \frac{\partial p_i}{\partial \beta} = \frac{y_i - p_i}{p_i(1 - p_i)} \frac{\partial F(a(x_i, \beta))}{\partial \beta} \frac{\partial a(x_i, \beta)}{\partial \beta}
\]

\[
S_i(\beta) = \frac{\partial L_i}{\partial \beta} = (y_i - F_i) \frac{f_i}{F_i(1 - F_i)} x_i
\]

FOC to maximize \( L \) defines ML \( \hat{\beta} \), such that \( \sum_{i=1}^{n} S_i(\hat{\beta}) = 0 \)
A quick aside
Let’s remind ourselves of logic of ML using linear model.

\[ L_i = -\frac{1}{2} \log(2\pi) - \frac{1}{2} (y_i - x_i\theta)^2 \]

Score function for an individual

\[ S_i(\theta) = \frac{\partial L_i}{\partial \theta} = (y_i - x_i\theta)x_i \]

FOC, in matrix notation,

\[ \sum_{i=1}^{n} S_i(\hat{\theta}) = X^\top (Y - X\hat{\theta}) = X^\top Y - X^\top X\hat{\theta} = 0 \]

so we have

\[ \hat{\theta} = (X^\top X)^{-1} X^\top Y \]
Particularly important concepts from this lecture

1. How to derive binary choice from models using normal and extreme value distributions
2. How to interpret the meaning of the coefficients (what is identified by reduced form parameters?)
3. The (log)likelihood and score of a Bernoulli process
4. How to parametrize a logit and probit (log) likelihood, and derivation of score
5. How the logit and probit differ (i) in motivation (2) in tail behavior
