Models with Limited Dependent Variables

Ercan Karadas

New York University
Department of Economics

Spring, 2018

Table of Contents

Resources:

- Greene (7e), pp.681-692, 760-768
- *Advanced Econometrics*, Amemiya 1985, Ch. 9
- *Discrete Choice Methods with Simulation*, Train 2009, pp.11-29, 34-52
- *Quantitative models in marketing research*, Franses and Paap, 2001, Ch. 4-6

Table of Contents

- In some cases, all we know whether a certain event took place or not. For example, whether someone is employed or unemployed
- As an another example suppose we want to explain whether a family possesses a car or not.
- As data, suppose we have data on N families ($i = 1, \dots, N$) with observations on their income x_{i2} and whether or not they own a car.
- In this case the dependent variable is binary and can be defined as follows

$$y_i = \begin{cases} 1 & \text{family } i \text{ owns a car} \\ 0 & \text{family } i \text{ does not own a car} \end{cases}$$

- Suppose we were to use a regression model to explain y_i from x_{2i} and an intercept term:

$$y_i = \beta_1 + \beta_2 x_{i2} + \varepsilon_i$$

  or in matrix form

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

  where $\mathbf{x}_i = [1 \quad x_{i2}]'$, $\boldsymbol{\beta} = [\beta_1 \quad \beta_2]'$
- We also assume that

$$E[\varepsilon_i | \mathbf{x}_i] = 0$$

- Now we are going to see what might go wrong in this model: basically two things

1) $E[y_i|\mathbf{x}_i]$ is a probability!

- First we compute

$$\begin{aligned} E[y_i|\mathbf{x}_i] &= 1 \times P[y_i = 1|\mathbf{x}_i] + 0 \times P[y_i = 0|\mathbf{x}_i] \\ &= P[y_i = 1|\mathbf{x}_i] \end{aligned}$$

  but we also have

$$E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$$

- Therefore

$$E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta} = P[y_i = 1|\mathbf{x}_i]$$

  Thus the linear model implies that $\mathbf{x}_i\boldsymbol{\beta}$ is a probability! But then $\mathbf{x}_i\boldsymbol{\beta}$ should lie between 0 and 1.

- This is only possible if the $\mathbf{x}_i$ values are bounded and if certain restrictions on $\boldsymbol{\beta}$ are satisfied. Usually this is hard to achieve in practice and therefore this is a serious problem.
- The first (and most important drawback) of the linear model: it may imply probabilities outside the [0,1] interval.

2) $\varepsilon_i|\mathbf{x}_i$ is heteroskedastic and is not Normally distributed

- The error term ε_i has a highly non-normal distribution. In fact, it can assume only two value, conditional on $\mathbf{x}_i$
- And we can compute the probabilities of these two values as

$$P[\varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta}|\mathbf{x}_i] = P[y_i = 0|\mathbf{x}_i] = 1 - \mathbf{x}_i'\boldsymbol{\beta}$$
$$P[\varepsilon_i = 1 - \mathbf{x}_i'\boldsymbol{\beta}|\mathbf{x}_i] = P[y_i = 1|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$$

- Using these results we can also show that

$$\text{var}[\varepsilon_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}(1 - \mathbf{x}_i'\boldsymbol{\beta})$$

- This implies that the variance of the error term is not constant but dependent upon the explanatory variables.
- Therefore, the second drawback is that the error term is heteroskedastic and non-normal.

Table of Contents

Linear Model with Binary Dependent Variable

Binary Choice Models

Multiple Response Models
- Ordered Response Models
- Unordered Response Models (Multinomial Models)

- Recall that when the dependent variable is binary we showed that $E[y_i|\mathbf{x}_i] = P[y_i = 1|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$, but there was nothing to assure that this quantity will be between 0 and 1. Binary choice models devise functions so as to satisfy this constraint.
- Binary choice models assume that

  $$P[y_i = 1|\mathbf{x}_i] = F(\mathbf{x}_i'\boldsymbol{\beta})$$

  where
  - $0 \leq F(\cdot) \leq 1$
  - $\mathbf{x}_i$ is a K-vector of observations
  - $\boldsymbol{\beta}$ is a K-vector of unknown parameters
- There are three common choices of F

Probit, Logit, LPM

1) When F is the standard normal distribution function

$$F(w) = \Phi(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}t^2\right\} dt$$

the binary model is called **probit model**.

2) When F is the standard logistic distribution function

$$F(w) = \Lambda(w) = \frac{e^w}{1 + e^w}$$

the binary model is called **logit model**.
The logistic distribution Λ is similar to a normal distribution but has a much simpler form so it's commonly used. It has zero mean and variance $\pi^2/3$.

3) A third choice corresponds to a uniform distribution over the interval [0, 1] with distribution function

$$F(w) = \begin{cases} 0 & w < 0 \\ 1 & 0 \leq w \leq 1 \\ 1 & w > 1 \end{cases}$$

This results in the so-called **linear probability model (LPM)**, which is similar to the linear regression model, but the probabilities are set to 0 or 1 if $\mathbf{x}_i'\boldsymbol{\beta}$ exceeds the lower or upper limit, respectively.

A Latent Variable Representation of Binary Models

- It is possible to derive a binary choice model from underlying behavioural assumptions.
- This method is called a **latent variable representation** of the binary model.
- This method is in common use even when such behavioral assumptions are not made because its methodology is applicable to some other cases.
- As an example let us look at an agent's decision to accept a job or not.
- It is reasonable to think that the agent will accept the job if the utility difference between having a paid job and not having one is above a certain threshold, which we can normalize to zero
- This utility difference might depend upon the wage as well as other personal characteristics, like age and education, etc.

- Thus, for each person i we can write the utility difference as a function of observed characteristics, $\mathbf{x}_i$ say, and unobserved characteristics, ε_i, say.
- Assuming a linear additive relationship, we obtain for the utility difference, denoted y_i^*, the following regression

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

- Since y_i^* is unobserved, it is referred to as a **latent variable**.
- What we observe is agent i's job status y_i. And let's say $y_i = 1$ if employed and $y_i = 0$ if unemployed.
- Combining this with our previous discussion that the agent would take a job if and only if the utility difference is positive we obtain

$$y_i = \begin{cases} 1 & y_i^* > 0 \\ 0 & y_i^* \leq 0 \end{cases}$$

- Now we can compute the probability of employment as

$$\begin{aligned} P[y_i = 1|\mathbf{x}_i] &= P[y_i^* > 0|\mathbf{x}_i] \\ &= P[\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i > 0|\mathbf{x}_i] \\ &= P[-\varepsilon_i \leq \mathbf{x}_i'\boldsymbol{\beta}|\mathbf{x}_i] \\ &= F[\mathbf{x}_i'\boldsymbol{\beta}] \end{aligned}$$

  where F is the distribution of $-\varepsilon_i$. But this is exactly the same expression that we used to model binary dependent variables.

- The form of the binary model depends upon the distribution of ε_i, F.
- If a standard normal distribution is chosen, one obtains the probit model; for the logistic distribution the logit model is obtained.
- We can express the probit model concisely as

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim NID(0,1)$$

$$y_i = \begin{cases} 1 & y_i^* > 0 \\ 0 & y_i^* \leq 0 \end{cases}$$

  The logit model can be represented similarly.

Estimation of the Binary Model

- Since y_i is a binary variable with $y_i \in \{0, 1\}$, the likelihood contribution of observation i can be written as

$$P[y_i = 1|\mathbf{x}_i; \boldsymbol{\beta}]^{y_i} P[y_i = 0|\mathbf{x}_i; \boldsymbol{\beta}]^{1-y_i}$$

- Then the likelihood function for the entire sample is thus given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} P[y_i = 1|\mathbf{x}_i; \boldsymbol{\beta}]^{y_i} P[y_i = 0|\mathbf{x}_i; \boldsymbol{\beta}]^{1-y_i}$$

- Substituting $P[y_i = 1|\mathbf{x}_i; \boldsymbol{\beta}] = F[\mathbf{x}_i'\boldsymbol{\beta}]$ and taking logs we obtain the loglikelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i \log F[\mathbf{x}_i'\boldsymbol{\beta}] + \sum_{i=1}^{N} (1 - y_i) \log(1 - F[\mathbf{x}_i'\boldsymbol{\beta}])$$

- To find the mle estimators we proceed as usual: take first order conditions, etc.

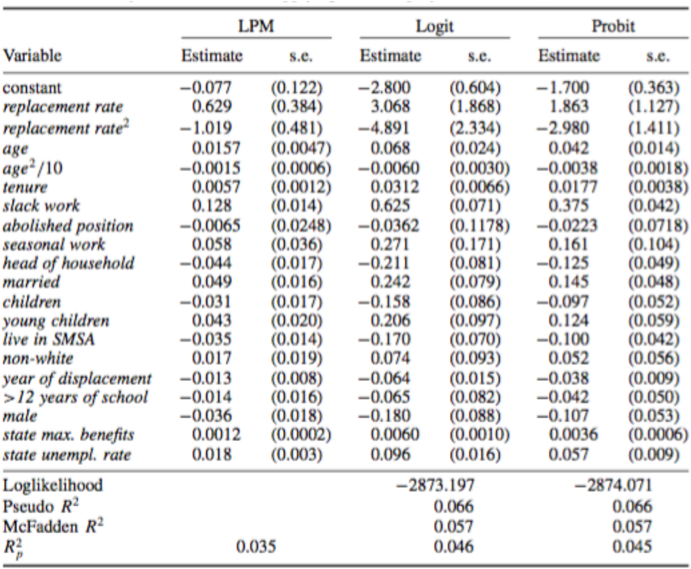
Application: Unemployment Benefits

- As an example we consider a sample of 4877 blue-collar workers who lost their jobs in the United States between 1982 and 1991, taken from a study by McCall (1995).
- Not all unemployed workers eligible for unemployment insurance (UI) benefits apply for it, probably owing to the associated pecuniary and psychological costs. The percentage of eligible unemployed blue-collar workers that actually apply for UI benefits is called the take-up rate, and it was only 68% in the available sample.
- It is therefore interesting to investigate what makes people decide not to apply.
- Here the dependent variable y_i can be expressed as

$$y_i = \begin{cases} 1 & \text{applied for UI} \\ 0 & \text{otherwise} \end{cases}$$

- The set of independent variables in $\mathbf{x}_i$ is shown in the first column of the following table. These variables include some personal characteristics(schooling, age, racial, gender) as well as some other variables beyond the agent's control (for example state unemployment).
- Motivating each variable is not the point of this example, but we are going to focus on interpretation of the results and how three binary model compare to each other

Estimation Results



	LPM		Logit		Probit	
Variable	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
constant	−0.077	(0.122)	−2.800	(0.604)	−1.700	(0.363)
replacement rate	0.629	(0.384)	3.068	(1.868)	1.863	(1.127)
replacement rate^2	−1.019	(0.481)	−4.891	(2.334)	−2.980	(1.411)
age	0.0157	(0.0047)	0.068	(0.024)	0.042	(0.014)
age^2*/10*	−0.0015	(0.0006)	−0.0060	(0.0030)	−0.0038	(0.0018)
tenure	0.0057	(0.0012)	0.0312	(0.0066)	0.0177	(0.0038)
slack work	0.128	(0.014)	0.625	(0.071)	0.375	(0.042)
abolished position	−0.0065	(0.0248)	−0.0362	(0.1178)	−0.0223	(0.0718)
seasonal work	0.058	(0.036)	0.271	(0.171)	0.161	(0.104)
head of household	−0.044	(0.017)	−0.211	(0.081)	−0.125	(0.049)
married	0.049	(0.016)	0.242	(0.079)	0.145	(0.048)
children	−0.031	(0.017)	−0.158	(0.086)	−0.097	(0.052)
young children	0.043	(0.020)	0.206	(0.097)	0.124	(0.059)
live in SMSA	−0.035	(0.014)	−0.170	(0.070)	−0.100	(0.042)
non-white	0.017	(0.019)	0.074	(0.093)	0.052	(0.056)
year of displacement	−0.013	(0.008)	−0.064	(0.015)	−0.038	(0.009)
>12 years of school	−0.014	(0.016)	−0.065	(0.082)	−0.042	(0.050)
male	−0.036	(0.018)	−0.180	(0.088)	−0.107	(0.053)
state max. benefits	0.0012	(0.0002)	0.0060	(0.0010)	0.0036	(0.0006)
state unempl. rate	0.018	(0.003)	0.096	(0.016)	0.057	(0.009)
Loglikelihood			−2873.197		−2874.071	
Pseudo R^2			0.066		0.066	
McFadden R^2			0.057		0.057	
R_p^2	0.035		0.046		0.045	

Figure: Binary choice models for applying for Unemp Benefits (blue-collar workers)

Comments on the Results

- The LPM is estimated by OLS, so no corrections for heteroskedasticity are made, and no attempt is made to keep the implied probabilities between 0 and 1.
- The logit and probit models are both estimated by maximum likelihood.
- The estimates of β obtained from the logit model are roughly a factor $\pi/\sqrt{3}$ larger than those obtained from the probit model, acknowledging the small differences in the shape of the distributions. This is because the logistic distribution has a variance of $\pi^2/3$ whereas the variance in probit model is 1.
- The signs of the coefficients are identical across the different specifications, while the statistical significance of the explanatory variables is also comparable.
- The dummy variable that indicates whether the job was lost because of slack work is highly significant in all specifications, which is not surprising given that these workers typically will find it hard to get a new job.
- The higher the state unemployment rate and the higher the maximum benefit level, the more likely it is that individuals will apply for benefits, which is intuitively reasonable.
- Usually, goodness-of-fit is fairly low for discrete choice models.

Table of Contents

Linear Model with Binary Dependent Variable

Binary Choice Models

Multiple Response Models

- Ordered Response Models
- Unordered Response Models (Multinomial Models)

- In many applications, the number of alternatives that can be chosen is larger than 2. For example, instead of being interested in only whether an individual works or not, we might be interested in learning about how people choose between full-time work, part-time work or not working.
- As an another example, we might be interested in the choice of a company to invest in Europe, Asia or the United States.
- An important distinction exists between ordered response models and unordered models:
  - An ordered response model is generally more parsimonious but is only appropriate if there exists a logical ordering of the alternatives. The reason is that it assumes there is one underlying latent variable that drives the choice between the alternatives. In other words, the results will be sensitive to the ordering of the alternatives, so this ordering should make sense.
  - Unordered models are not sensitive to the way in which the alternatives are numbered. In many cases, they can be based upon the assumption that each alternative has a utility level and that individuals choose the alternative that yields highest utility.

Ordered Response Models

- Let us consider the choice between M alternatives, numbered from 1 to M
- If there is a logical ordering in these alternatives (e.g. no car, one car, more than one car), a so-called **ordered response model** can be used. This model is also based on one underlying latent variable but with a different match from the latent variable, y_i^*, to the observed one $y_i \in \{1, 2, \ldots, M\}$.
- We can express an ordered response model as follows

$$
\begin{aligned}
y_i^* &= \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i \\
y_i = j &\iff \gamma_{j-1} < y_i^* \leq \gamma_j
\end{aligned}
$$

  for unknown parameters with $\gamma_0 = -\infty$, $\gamma_1 = 0$ and $\gamma_M = \infty$ (Setting $\gamma_1 = 0$ is just a normalization).
- Consequently, the probability that alternative j is chosen is the probability that the latent variable y_i^* is between two boundaries γ_{j-1} and γ_j.
- Assuming that ε_i is i.i.d. standard normal results in the **ordered probit model**. The logistic distribution gives the **ordered logit model**.
- For $M = 2$ we are back at the binary choice model.

Example

- Suppose an agent answers the question *How much would you like to work?* in three categories *not*, *part-time* and *full-time*.
- To model the outcomes, $y_i = 1$ (not working), $y_i = 2$ (part-time working) and $y_i = 3$ (full-time working), we note that there appears to be a logical ordering in these answers.
- To be precise, the question is whether it is reasonable to assume that there exists a single index $y_i^* = \mathbf{x}_i'\boldsymbol{\beta}$ such that higher values for this index correspond to, on average, larger values for y_i. Here, we can interpret y_i^* as 'willingness to work'
- We can express the ordered response model for this example as

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

$$y_i = \begin{cases} 1 & y_i^* \leq 0 \\ 2 & 0 < y_i^* \leq \gamma \\ 3 & y_i^* > \gamma \end{cases}$$

- For the rest of this example assume that $\varepsilon_i \sim NID(0, \sigma^2)$
- Then we can compute the probability of each outcome as

$$
\begin{aligned}
P[y_i = 1|\mathbf{x}_i] &= P[y_i^* \leq 0|\mathbf{x}_i] = \Phi\left(-\mathbf{x}_i'\boldsymbol{\beta}\right) \\
P[y_i = 3|\mathbf{x}_i] &= P[y_i^* > \gamma|\mathbf{x}_i] = 1 - \Phi\left(\gamma - \mathbf{x}_i'\boldsymbol{\beta}\right) \\
P[y_i = 2|\mathbf{x}_i] &= P[y_i^* > \gamma|\mathbf{x}_i] = \Phi\left(\gamma - \mathbf{x}_i'\boldsymbol{\beta}\right) - \Phi\left(-\mathbf{x}_i'\boldsymbol{\beta}\right)
\end{aligned}
$$

- γ is an unknown parameter that is estimated jointly with $\boldsymbol{\beta}$. This point is interesting: so, we choose the threshold value as well. In an MLE framework this means that we choose this value to maximize the likelihood of the sample.
- Using these probabilities we can write the likelihood function for a sample of size N as

$$
L(\gamma, \boldsymbol{\beta}) = \prod_{i=1}^{N} P[y_i = 1|\mathbf{x}_i; \gamma, \boldsymbol{\beta}]^{a_1} P[y_i = 2|\mathbf{x}_i; \gamma, \boldsymbol{\beta}]^{a_2} P[y_i = 1|\mathbf{x}_i; \gamma, \boldsymbol{\beta}]^{a_3}
$$

  where

$$
\begin{aligned}
a_1 &= (2 - y_i)(3 - y_i) - 1 \\
a_2 &= -(1 - y_i)(3 - y_i) \\
a_3 &= (1 - y_i)(2 - y_i) - 1
\end{aligned}
$$

- Instead of a_1, a_2, a_3 we could use three dummies defined as: $d_{ij} = 1$ if $y_i = j$, $j = 1, 2, 3$.

An Unordered Response Model: Multinomial Logit Model

- In several cases, there is no natural ordering in the alternatives, and it is not realistic to assume that there is a monotonic relationship between one underlying latent variable and the observed outcomes.
- Consider, for example, modeling the mode of transportation (bus, train, car, bicycle, walking).
- In such cases, an alternative framework has to be used to put some structure on the different probabilities.
- A popular framework is a **random utility model**, in which the utility of each alternative is a linear function of observed characteristics (individual and/or alternative specific) plus an additive unobservable disturbance term.
- Individuals are assumed to choose the alternative that has the highest utility.
- With appropriate distributional assumptions on the disturbance terms, this approach leads to manageable expressions for the probabilities implied by the model.

- There is a choice between M alternatives, indexed $j = 1, 2, \ldots, M$, noting that the order is arbitrary.
- Next, assume that the utility level that individual i attaches to each of the alternatives is given by U_{ij}
- Then alternative j is chosen by individual i if it gives highest utility, that is, if

$$U_{ij} = \max\{U_{i1}, \ldots, U_{iM}\}$$

- Since these utility levels are not observed, and we need to make some additional assumptions to make this set-up operational.
- Let us assume that

$$U_{ij} = \mu_{ij} + \varepsilon_{ij}$$

  where μ_{ij} is a nonstochastic function of observables and a small number of unknown parameters, and ε_{ij} is an unobservable error term.
- We will talk about how μ_{ij} is determined later.

- From this we compute the probability of choosing j as

$$\begin{aligned} P[y_i = j] &= P\left[U_{ij} = \max\{U_{i1}, \ldots, U_{iM}\}\right] \\ &= P\left[\mu_{ij} + \varepsilon_{ij} > \max_{k=1,\ldots,M,\, k\neq j}\{\mu_{ik} + \varepsilon_{ik}\}\right] \\ &= P\left[\varepsilon_{ij} - \varepsilon_{ik} > \mu_{ik} - \mu_{ij}, k = 1, \ldots, M,\ k \neq j\right] \end{aligned}$$

- To compute this probability we need to specify a distribution for ε_{ij}.
- Gumbell distribution (also known as a type I extreme value distribution) leads to a very simple expression for these probabilities

$$F(\varepsilon_{ij}) = \exp\{-e^{-\varepsilon_{ij}}\}$$

- Under these assumptions, it can be shown that the probability of choosing item j is

$$P[y_i = j] = \frac{e^{\mu_{ij}}}{e^{\mu_{i1}} + e^{\mu_{i2}} + \cdots + e^{\mu_{iM}}}$$

  (The derivation of this result is given at the end if you are interested)
- In this model the probability of an individual choosing alternative j is a simple function of the explanatory variables, by virtue of the convenient assumptions made about the distribution of the unobservables ε_{ij}

Normalization of Utility: Examples

- Consider a person who can take either a Car or a Bus to work. There are two *attributes (characteristics)* that the consumer cares about: Time and Money (cost). So the utility of consumer i from alternative j is

  $$U_{ij} = \beta_1 T_j + \beta_2 M_j + \varepsilon_{ij},$$

  where T_j (M_j) is the time (money(cost)) that consumer i incurs traveling to work using j, $j \in \{\text{Car}, \text{Bus}\}$.
- Now suppose that the researcher thinks that there is a minimum utility k_j that each alternative brings to the consumer. Basically we want to include alternative-specific constants to model. But then

  $$U_{ic} = \beta_1 T_c + \beta_2 M_c + k_c^0 + \varepsilon_{ic}$$
  $$U_{ib} = \beta_1 T_b + \beta_2 M_b + k_b^0 + \varepsilon_{ib}$$

  is equivalent to a model

  $$U_{ic} = \beta_1 T_c + \beta_2 M_c + k_c^1 + \varepsilon_{ic}$$
  $$U_{ib} = \beta_1 T_b + \beta_2 M_b + k_b^1 + \varepsilon_{ib}$$

  as long as $k_b^0 - k_c^0 = k_b^1 - k_c^1$ since only differences in utilities matter.
- Therefore, it is impossible to estimate the two constants themselves, since an infinite number of values of the two constants (any values that have the same difference) result in the same choice probabilities.
- To account for this fact, the researcher must normalize the absolute levels of the constants. The standard procedure is to normalize one of the constants to zero. For example, the following specification would work

  $$U_{ic} = \beta_1 T_c + \beta_2 M_c + \varepsilon_{ic}$$
  $$U_{ib} = \beta_1 T_b + \beta_2 M_b + k_b + \varepsilon_{ib}$$
- Therefore, with J alternatives, at most $J - 1$ alternative-specific constants can enter the model, with one of the constants normalized to zero.

Normalization of Utility (Cont'd)

- Now extend the previous example to include the effect of a person's income (Y) on the decision whether to take bus or car to work:

$$U_{ic} = \beta_1 T_c + \beta_2 M_c + \beta_{3c} Y_i + \varepsilon_{ic}$$
$$U_{ib} = \beta_1 T_b + \beta_2 M_b + \beta_3 Y_i + k_b + \varepsilon_{ib}$$

- Note that coefficients of Y should be alternative specific (one for car and one for bus) because otherwise they would vanish in the difference of utilities
- Again since only differences in utility matter, the absolute levels of β_{3c} and β_{3b} cannot be estimated, only their difference $\beta_{3c} - \beta_{3b}$. To set the level, one of these parameters is normalized to zero. The model becomes

$$U_{ic} = \beta_1 T_c + \beta_2 M_c + \varepsilon_{ic}$$
$$U_{ib} = \beta_1 T_b + \beta_2 M_b + \beta_3 Y_i + k_b + \varepsilon_{ib}$$

  where $\beta_3 = \beta_{3c} - \beta_{3b}$ and is interpreted as the differential effect of income on the utility of bus compared to car.

- Just as adding a constant to the utility of all alternatives does not change the decision maker's choice, neither does multiplying each alternative's utility by a constant. The alternative with the highest utility is the same no matter how utility is scaled. Therefore, a model with $U_{ij} = \mu_{ij} + \varepsilon_{ij}$ is equivalent to a model with $\tilde{U}_{ij} = \lambda\mu_{ij} + \lambda\varepsilon_{ij}$, $\lambda > 0$.
- In the next slide we are going to summarize these issues and see how they are commonly handled in practice.

Normalization of Utility

Scale

- *The Overall Scale of Utility Is Irrelevant:* Multiplying each alternative's utility by a constant does not change the decision maker's choice (scale is indeterminate)
- The scaling of the model is determined by the distribution of ε_{ij} (basically through its mean and variance).

Location

- *Only Differences in Utility Matter:* If a constant is added to the utility of all alternatives, the alternative with the highest utility doesnt change (location is indeterminate)
- To determine the location of utility, it is common to normalize by focusing on the differences between utilities of all the other alternatives from a reference one, say alternative one μ_{i1}:

$$P[y_i = j] = \frac{e^{\mu_{ij} - \mu_{i1}}}{1 + e^{\mu_{i2} - \mu_{i1}} + \cdots + e^{\mu_{iM} - \mu_{i1}}}$$

- From now on we will use this normalized formula.

Specification of μ_{ij}

- If μ_{ij} is assumed to be a linear function of observable variables, which may depend upon the individual (i), the alternative (j) or both, we can write

  $$\mu_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta}$$

  For example, when explaining the mode of transportation, this may include variables like traveling time and costs, which may vary from one person to another.

- In some applications, we may observe the characteristics of the decision-makers, for example their age, gender and income. In this case, it is appropriate to impose

  $$\mu_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j$$

  where $\mathbf{x}_i$ is a K-dimensional vector containing the characteristics of individual i (including an intercept term) and β_j denotes a vector of alternative specific coefficients.

Example 1

- Suppose that a number of respondents are asked to pick their preferred coffee-maker.
- Assume that the utility derived from a certain coffee-maker depends on two characteristics: capacity, price.
- Suppose the current market for coffee-makers consists of two products:
  - machine 1: 10 cups for 25 dollars
  - machine 2: 15 cups for 35 dollars
- In this case, characteristics are independent of respondents so we can write

$$x_{i1} = \begin{bmatrix} 10 \\ 25 \end{bmatrix}, \quad x_{i2} = \begin{bmatrix} 15 \\ 35 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad i = 1, 2, \ldots, N$$

  Here we could drop the subscript denoting the individuals and simply use $\mathbf{x}_{ij} = \mathbf{x}_j$

Example 2

- The same two machines as in Example 1
- But now suppose that in order to but the second machine agent i should pay a transportation cost $c_i = 0.1i$ (think of that agents are sorted according to their distance to the location of machine 2 so that the c_i is highest for the agent with highest index).
- In this case

$$x_{i1} = \begin{bmatrix} 10 \\ 25 \end{bmatrix}, \quad x_{i2} = \begin{bmatrix} 15 \\ 35 + 0.1i \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad i = 1, 2, \ldots, N$$

Example 3

- The fact that $\boldsymbol{\beta}$ is constant for all the alternatives makes this model useful in predicting the demand for a certain new alternative that comes into existence.
- Now to see how this happens let's go back to Example 1 and suppose that we data from purchasing behavior of N customers (respondents). Each element of this data set is either 1 or 2, denoting whether an agent purchased machine 1 or 2, respectively.
- Using this data we can compute the mle $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.
- Now suppose there is another company considering to launch a new machine, machine 3, with 12 cups for 30 dollars.
- Then the probability of agent i choosing machine 3 is

$$P[y_i = 3] = \frac{e^{(x_3 - x_1)'\hat{\boldsymbol{\beta}}}}{1 + e^{(x_2 - x_1)'\hat{\boldsymbol{\beta}}} + e^{(x_3 - x_1)'\hat{\boldsymbol{\beta}}}}$$

where

$$x_1 = \begin{bmatrix} 10 \\ 25 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 15 \\ 35 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 12 \\ 30 \end{bmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}, \quad i = 1, 2, \ldots, N$$

- If the respondents are representative of those who buy coffee-makers, this probability also corresponds to the expected market share of this new product.

Example 4 (Amemiya, pp.297)

- Consider the decision of a person regarding whether he or she drives a car or travels by transit to work.
- As another example, consider the following hypothetical model of transport modal choice among three alternatives - car, bus, and train (corresponding to the subscripts 1, 2, and 3, respectively). We assume that the utilities associated with three alternatives

  $$U_{ij} = \alpha + \mathbf{z}_{ij}'\boldsymbol{\theta} + \mathbf{w}_i'\boldsymbol{\gamma} + \varepsilon_{ij}, \quad j = 1, 2, 3$$

  (so here $\mu_{ij} = \alpha + \mathbf{z}_{ij}'\boldsymbol{\theta} + \mathbf{w}_i'\gamma$. Also we could express this as $\mathbf{x}_{ij}'\boldsymbol{\beta}$ by suitably partitioning $\mathbf{x}_i$ and $\boldsymbol{\beta}$, see Greene, pp. 762.)
- We assume that the utility associated with each mode of transport is a function of the mode characteristics (attributes) **z** (mainly the time and the cost incurred by the use of the mode) and the individual's socioeconomic characteristics **w** (income, age, etc), plus an additive error term ε. It is assumed that α, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$ are constant for all i and j.
- Note that when we look at the nonstochastic utility differences $\mu_{ij} - \mu_{i1}$ only $(\mathbf{z}_{ij} - \mathbf{z}_{i1})'\boldsymbol{\theta}$ remains and therefore

  $$P[y_i = j] = \frac{e^{(\mathbf{z}_{ij} - \mathbf{z}_{i1})'\boldsymbol{\theta}}}{1 + e^{(\mathbf{z}_{i2} - \mathbf{z}_{i1})'\boldsymbol{\theta}} + e^{(\mathbf{z}_{i3} - \mathbf{z}_{i1})'\boldsymbol{\theta}}}, \quad j = 2, 3.$$

Independence of Irrelevant Alternatives (IIA)

- Despite the attractiveness of the analytical expressions given for the multinomial model, these models have one big drawback, which is due to the assumption that all ε_{ij} are independent.
- This implies that (conditional upon observed characteristics) utility levels of any two alternatives are independent.
- This is particularly troublesome if two or more alternatives are very similar. A typical example would be to decompose the category 'travel by bus' into 'travel by blue bus' and 'travel by red bus'. Clearly, we would expect that a high utility for a red bus implies a high utility for a blue bus.
- McFadden (1974) called this property of the multinomial logit model independence of irrelevant alternatives (IIA).

Ordered versus Unordered Response Models

- The use of ordered response model requires a single unobserved index variable. In another word, latent variable should have a logical interpretation when dependent variable y_i is associated with the latent variable monotonically.
- Ordered response models are not very common because economic phenomena often are complex and difficult to explain in terms of only a single unobserved index variable.
- We should be cautious in using an ordered model because if the true model is unordered, an ordered model can lead to serious biases in the estimation of the probabilities.
- On the other hand, the cost of using an unordered model when the true model is ordered is a loss of efficiency rather than consistency.

Derivation of Multinomial Logit Formula (Train, 2009, pp.74-75)

The probability that decision maker i chooses alternative j is

$$\begin{aligned} P_{ij} \equiv P[y_i = j] &= P\left[U_{ij} = \max\{U_{i1}, \ldots, U_{iM}\}\right] \\ &= P\left[\mu_{ij} + \varepsilon_{ij} > \max_{k=1,\ldots,M,\ k\neq j}\{\mu_{ik} + \varepsilon_{ik}\}\right] \\ &= P\left[\varepsilon_{ik} < \varepsilon_{ij} + \mu_{ij} - \mu_{ik}, k = 1, \ldots, M,\ k \neq j\right] \end{aligned}$$

If ε_{ij} is considered given, this expression is the cumulative distribution for each ε_{ik} evaluated at $\varepsilon_{ij} + \mu_{ij} - \mu_{ik}$, which, is $exp(-exp(-(\varepsilon_{ij} + \mu_{ij} - \mu_{ik})))$. Since the ε's are independent, this cumulative distribution over all $j \neq k$ is the product of the individual cumulative distributions:

$$P_{ij} \mid \varepsilon_{ij} = \prod_{k\neq j} e^{-e^{-(\varepsilon_{ij}+\mu_{ij}-\mu_{ik})}}$$

Then the choice probability is the integral of $P_{ij} \mid \varepsilon_{ij}$ over all values of ε_{ij} weighted by its density:

$$P_{ij} = \int_{-\infty}^{\infty} \left(\prod_{k\neq j} e^{-e^{-(s+\mu_{ij}-\mu_{ik})}}\right) e^{-s} e^{-e^{-s}}\ ds$$

where we switched to $s = \varepsilon_{ij}$ for simplicity.
Noting that $\mu_{ij} - \mu_{ij} = 0$ and then collecting terms in the exponent of e, we have

$$\begin{aligned} P_{ij} &= \int_{-\infty}^{\infty} \left(\prod_{k} e^{-e^{-(s+\mu_{ij}-\mu_{ik})}}\right) e^{-s}\ ds \\ &= \int_{-\infty}^{\infty} \left(e^{-\sum_k e^{-(s+\mu_{ij}-\mu_{ik})}}\right) e^{-s}\ ds \\ &= \int_{-\infty}^{\infty} \left(e^{-e^{-s}\sum_k e^{-(\mu_{ij}-\mu_{ik})}}\right) e^{-s}\ ds \end{aligned}$$

Define $t = e^{-s}$ such that $-e^{-s}ds = dt$. Note that as s approaches infinity, t approaches zero, and as s approaches negative infinity, t becomes infinitely large. Using this new term,

$$\begin{aligned}
P_{ij} &= \int_{\infty}^{0} \left(e^{-t \sum_k e^{-(\mu_{ij} - \mu_{ik})}} \right) (-1) dt \\
&= \int_{0}^{\infty} \left(e^{-t \sum_k e^{-(\mu_{ij} - \mu_{ik})}} \right) dt \\
&= \frac{e^{-t \sum_k e^{-(\mu_{ij} - \mu_{ik})}}}{-\sum_k e^{-(\mu_{ij} - \mu_{ik})}} \ \Big|_0^{\infty} \\
&= \frac{1}{\sum_k e^{-(\mu_{ij} - \mu_{ik})}} \\
&= \frac{e^{\mu_{ij}}}{\sum_k e^{\mu_{ik}}} \ .
\end{aligned}$$