Lecture: Discrete Choice

October 31, 2024

1 Motivation and overview

Topics include

- Discrete choice and differentiated commodities
- Dynamic discrete choice
- Empirical models for auctions

Motivation

- Common theme in these three topics is the interplay of economic theory and econometrics.
- We use economic theory to formulate the models. Empirical models derived from economic theory are often called structural models.
- Major concern will be what can be estimated realistically, i.e. with available data.
- Because the models will involve functional form assumptions and assumptions on distributions, it will be important to discuss nonparametric identification, i.e. the question whether functional forms and distributions can be recovered from the available data.
- If nonparametric identification can be established, we can use either a parametric or non/semi-parametric estimation procedure.
- Other issue will be computation: Structural models can involve high-dimensional integrals or depend on functions that can only be computed recursively or as a fixed point.

- Important question: Why use models derived from economic theory in empirical research?
    - Guidance in specification (and thereby identification) of model that relates outcomes to determining variables.
    - More important: Computation of counterfactuals.

2 Discrete Choice and Differentiated Commodities

We compare two types of consumer choice problems

- Discrete choice of differentiated commodity.
    - Consider consumer who wants to buy a car.
    - There are a finite number of car types that differ by attributes as size, performance, style etc.
    - Consumer buys only one car, car is indivisible
- Continuous choice of homogeneous commodity.
    - In traditional demand analysis income is allocated to purchase of quantities of homogeneous commodities.
    - These commodities have a unit price.
    - Quantities vary continuously.

Two important differences

- Discrete choice deals with indivisible goods, continuous choice with divisible goods.
- Discrete choice deals with differentiated goods and continuous choice with homogeneous goods.
- Of course, continuous choice could also be for differentiated commodities.
- We could make discrete choice continuous by considering
    - Repeated choice by one consumer: Fraction of choices resulting in particular car type is continuous.

  - Choice of group of consumers: Fraction of choices of particular car type.
  - Both have their problems.

2.1 Additive random utility model

- Consider consumer who chooses between alternatives $i = 1, \dots, I$. We omit subscript for consumer.
- u_i is the utility associated with choice i.
- For u_i we specify a simple model, the Additive Random Utility (ARUM) model $u_i = -v_i + \varepsilon_i$
- $-v_i$ is the mean utility of alternative i . Later we make it dependent on observed (and unobserved) characteristics of the alternative and of the agent. For now think of v_i as the price of alternative i (the minus sign simplifies the notation in the sequel).
- ε_i is a random variable with $\mathrm{E}(\varepsilon_i) = 0$.
- Interpretations
  - Optimization error.
  - Unobserved attributes of alternative i, e.g. style of car.
  - Unobserved attributes of the agent.
  - Interactions of these two.
- Interpretation is important if we want to use the ARUM as a model of consumer demand for differentiated products. If we include the unobserved alternative and individual attributes in v_i only optimization error remains (but may be needed to fit the data; without it choice of alternative is predictable). For now we allow all interpretations. Repeated choice by the same agent would help in settling the interpretation.
- Interpretation also affects the nature of the joint distribution of $\varepsilon_1, \dots, \varepsilon_I$. If there are unobserved agent attributes these random variables will be

correlated. We assume

$$\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_I \end{pmatrix} \sim F(\varepsilon_1, \dots, \varepsilon_I)$$

with F left unspecified for now. We assume that F is independent of $v_1, \dots, v_I$, but we will consider cases in which this does not hold later. The joint pdf is $f(\varepsilon_1, \dots, \varepsilon_I)$.

2.1.1 ARUM and choice

- The agent choose alternative i if and only if

$$u_i > u_j \text{ for } i \neq j = 1, \dots, I \Leftrightarrow -v_i + \varepsilon_i > -v_j + \varepsilon_j \text{ for } i \neq j = 1, \dots, I$$

- We compute the probability that i is chosen, $p_i(v)$, from the joint distribution of $\varepsilon_1, \dots, \varepsilon_I$.

$$p_i(v) = \Pr(\varepsilon_1 < v_1 - v_i + \varepsilon_i, \dots, \varepsilon_I < v_I - v_i + \varepsilon_i) =$$

$$= \int_{-\infty}^{\infty} \int_{\infty}^{v_1 - v_i + \varepsilon_i} \quad \int_{-\infty}^{v_I - v_i + \varepsilon_i} f(\varepsilon_1, \dots, \varepsilon_I) \mathrm{d}\varepsilon_1 \dots \mathrm{d}\varepsilon_I \mathrm{d}\varepsilon_i =$$

$$\int_{-\infty}^{\infty} \frac{\partial F}{\partial \varepsilon_i} (v_1 - v_i + \varepsilon_i, \dots, \varepsilon_i, \dots, v_I - v_i + \varepsilon_i) \mathrm{d}\varepsilon_i$$

- For $I = 3$

$$\begin{aligned} p_1(v) &= Pr(\varepsilon_2 < v_2 - v_1 + \varepsilon_1, \varepsilon_3 < v_3 - v_1 + \varepsilon_1) \\ p_2(v) &= Pr(\varepsilon_1 < v_1 - v_2 + \varepsilon_2, \varepsilon_3 < v_3 - v_2 + \varepsilon_2) \\ p_3(v) &= Pr(\varepsilon_1 < v_1 - v_3 + \varepsilon_3, \varepsilon_2 < v_2 - v_3 + \varepsilon_3) \end{aligned}$$

or if we define $\eta_1 = \varepsilon_2 - \varepsilon_1$ and $\eta_2 = \varepsilon_3 - \varepsilon_1$ (see figure 1)

$$\begin{aligned} p_1(v) &= \Pr(\eta_1 < v_2 - v_1, \eta_2 < v_3 - v_1) \\ p_2(v) &= Pr(\eta_1 > v_2 - v_1, \eta_2 < v_3 - v_2 + \eta_1) \\ p_3(v) &= Pr(\eta_1 < v_2 - v_3 + \eta_2, \eta_2 > v_2 - v_1) \end{aligned}$$

- Note that the choice probabilities can be expressed as an event involving the random variables $\varepsilon_2 - \varepsilon_1$ and $\varepsilon_3 - \varepsilon_1$ and the average utility differences $v_2 - v_1$ and $v_3 - v_1$, i.e. they involve only utility comparisons with alternative 1 that is chosen as the reference alternative.

2.1.2 Multinomial logit

- The choice probabilities have a closed-form only for special distributions of $\varepsilon_1, \ldots, \varepsilon_I$.

- Consider the case that the ε_i are independent and identically distributed with an Extreme Value distribution with pdf

$$f(\varepsilon_i) = e^{-\varepsilon_i} e^{-e^{-\varepsilon_i}}, -\infty < \varepsilon_i < \infty$$

and cdf

$$F(\varepsilon_i) = e^{-e^{-\varepsilon_i}}$$

We have $\mathrm{E}(\varepsilon_i) = \gamma = .5772$ which is Euler's constant. Also $\mathrm{Var}(\varepsilon_i) = \pi^2/6$.

- For the choice probabilities (using the final expression above)

$$p_i(v) = \int_{-\infty}^{\infty} e^{-\varepsilon_i} e^{-e^{-\varepsilon_i}} \prod_{j=1,j\neq i}^{I} e^{-e^{-(v_j - v_i + \varepsilon_i)}} \mathrm{d}\varepsilon_i = \int_{-\infty}^{\infty} e^{-\varepsilon_i} e^{-e^{-\varepsilon_i}(1+\sum_{j=1,j\neq i}^{I} e^{v_i - v_j})} \mathrm{d}\varepsilon_i$$

Define

$$\lambda_i = \ln\left(1 + \sum_{j=1,j\neq i}^{I} e^{v_i - v_j}\right)$$

so that

$$p_i(v) = \int_{-\infty}^{\infty} e^{-\varepsilon_i} e^{-e^{-(\varepsilon_i - \lambda_i)}} \mathrm{d}\varepsilon_i$$

Change of variables to $\eta_i = \varepsilon_i - \lambda_i$ so that $\varepsilon_i = \eta_i + \lambda_i$

$$p_i(v) = e^{-\lambda_i} \int_{-\infty}^{\infty} e^{-\eta_i} e^{-e^{-\eta_i}} \mathrm{d}\eta_i = e^{-\lambda_i} = \frac{1}{1 + \sum_{j=1,j\neq i}^{I} e^{v_i - v_j}} = \frac{e^{-v_i}}{\sum_{j=1}^{I} e^{-v_j}}$$

- The discrete choice model with these choice probabilities is called the Multinomial Logit (MNL) model.

Properties of MNL model

- MNL model most popular model of discrete choice, because the choice probabilities have a closed-form expression.
- MNL with scale parameters: In ARUM replace ε_i by $\tilde{\varepsilon}_i = \sigma\varepsilon_i$. The choice probabilities are

  $$p_i(v) = \frac{e^{-\frac{v_i}{\sigma}}}{\sum_{j=1}^{I} e^{-\frac{v_j}{\sigma}}} = \frac{1}{\sum_{j=1}^{I} e^{-\frac{(v_j - v_i)}{\sigma}}}$$

  - If $\sigma \to \infty$ then $p_i(v) \to \frac{1}{I}$, i.e. the choice probabilities are independent of the average utilities $v_1, \ldots, v_I$.
  - If $\sigma \to 0$, then

    $$\begin{aligned} p_i(v) &= 1 \quad \text{if } -v_i > -v_j, j \neq i \\ &= 0 \quad \text{if } -v_j > -v_i, \text{for some} j \neq i \end{aligned}$$

    i.e. the alternative with the highest average utility is chosen.
  - If we introduce a scale parameter only for alternative i, then if $\sigma_i \to \infty$, the choice probabilities are independent of the average utility of i and if $\sigma_i \to 0$, alternative i is either never or always chosen depending on whether the average utility is negative or positive, respectively.
  - The effect of v_i is proportional to $\frac{1}{\sigma_i}$, e.g. if v_i is log price, the elasticities with respect to this price are small if σ_i is large.
- It has a property that limits its usefulness (at least without changes): Independence of Irrelevant Alternatives (IIA)

  $$\frac{p_i(v)}{p_j(v)} = e^{-(v_i - v_j)}$$

  i.e. the ratio of any two choice probabilities depends only on the average utilities of these alternatives, and not on the average utilities of the other alternatives.

IIA and substitution of alternatives

- This imposes restrictions on the substitution between alternatives, if an alternative becomes less attractive, e.g. due to a price increase, or if a new alternative is introduced, e.g. a new car model.

- New alternative. We assume that initially there are 2 alternative car models, 1=Sedan, and 2=SUV that have the same average utility $-v_1 = -v_2$, so that

  $$p_1(v) = \frac{e^{-v_1}}{e^{-v_1}+e^{-v_2}} = \frac{1}{2} = p_2(v)$$

  All SUV are green. A new type of SUV is introduced, the red SUV (alternative 3). Obviously, $-v_2 = -v_3$, so that

  $$p_1(v) = \frac{e^{-v_1}}{e^{-v_1}+e^{-v_2}+e^{-v_3}} = \frac{1}{3}$$

  Conclusion: The market share (fraction of consumers that chooses the alternative) of the sedan has gone down, due to the introduction of a close substitute of the other alternative the (green) SUV. One would expect that the red SUV would only reduce the market share of the green SUV. The reason is that although the red and green SUV have the same average utilities, they have independent random utility components.

- Elasticities: We have

  $$\begin{aligned} \frac{\partial p_i}{\partial v_i}(v) &= -p_i(v)(1-p_i(v)) \\ \frac{\partial p_i}{\partial v_j}(v) &= p_i(v)p_j(v) \end{aligned}$$

  or

  $$\begin{aligned} \frac{\partial \ln p_i}{\partial v_i}(v) &= -(1-p_i(v)) \\ \frac{\partial \ln p_i}{\partial v_j}(v) &= p_j(v) \end{aligned}$$

  If v_i is log(price), we conclude that the price elasticity of alternative i only depends on the market share of the alternative whose price is increased.

- Implication for pricing if supplier is monopolist: If number of consumers is m, the demand curve is $m \cdot p_i(v)$ and with v_i the price of alternative i and c_i the constant marginal cost of an extra unit, the profit is

  $$\pi_i = (v_i - c_i)m \cdot p_i(v)$$

The first order condition is

$$v_i = c_i - \frac{p_i(v)}{\frac{\partial p_i}{\partial v_i}(v)} = c_i + \frac{1}{1 - p_i(v)}$$

Hence the mark-up over marginal cost depends only on the market share and it increases with the market share. Implication is that in niche markets without close competitors the mark-ups will be small, and that is contrary to intuition.

2.2 Specification of the average utilities and estimation

- We introduce a second subscript for the agents, i.e. we have subscript $i = 1, \ldots, I$ for the I alternatives and $t = 1, \ldots, T$ for T agents.
- The choice set may be different for the agents, i.e. agent t may only choose between I_t alternatives. The expressions can be easily modified to account for this.
- The ARUM model now is

  $$u_{it} = -v_{it} + \varepsilon_{it}$$

  with ε_{it} i.i.d. Extreme Value for MNL model.
- The average utility depends on
  - Attributes of the alternatives, e.g. price, quality etc. denoted by z_i.
  - Attributes of the agent, e.g. income, age etc. if the agent is an individual denoted by x_t.
- General specification of average utilities

  $$-v_{it} = \beta' z_i + \gamma_i' x_t$$

  with z_i a K-vector of alternative attributes and x_t an L-vector of agent attributes. We interpret β_k as the marginal utility of attribute z_{ik}.
- Note that β is the same for all alternatives, but z_i is alternative specific, while i is alternative specific and x_t is the same for all alternatives.

- The choice probabilities do not change if for all alternatives we add a constant to the average utilities. Hence $\gamma_i + c$ is observationally equivalent, i.e. gives the same choice probabilities, as γ_i, so that $\gamma_i, i = 1, \ldots, I$ are identified up to a vector of constants. We normalize by setting $\gamma_1 = 0$.

- A special choice for z_i is a vector of alternative specific dummies

$$\begin{aligned} d_{ij} &= 1 \quad \text{if } j = i \\ &= 0 \quad \text{if } j \neq i \end{aligned}$$

  so that (I use $\alpha_j, j = 1, \ldots, I$ as parameters for this special case)

$$\beta' z_i = \sum_{j=1}^{I} \alpha_j d_{ij}$$

  Because $\sum_{j=1}^{I} d_{ij} = 1$ we can add a constant to the α_j without changing the choice probabilities. Again we normalize by setting $\alpha_1 = 0$.

- An important question is what parameters can be identified with different types of data.

  - Aggregate fractions or market share data, i.e. fractions that choose each alternative. If we have a model with alternative specific dummies there is a one-one correspondence between these parameters and the choice probabilities

$$\ln p_i(v) - \ln p_1(v) = \alpha_i \quad i = 2, \ldots, I$$

    if is the reference alternative ($\alpha_1 = 0$). We obtain an estimator if we replace probabilities by observed fractions.

  - Aggregate fractions or market shares by subgroup. As an example assume that we have aggregate fractions by income category. We assume that there are T categories. We denote the corresponding choice probabilities by $p_i(v_t)$ with $t = 1, \ldots, T$. We have $T(I-1)$ independent choice probabilities. If we again use alternative specific dummies and in addition a set of income dummies, we have for the average utility

$$\sum_{j=2}^{I} \alpha_j d_{ij} + \sum_{t=2}^{T} \gamma_{it} x_t$$

with $\gamma_{1t} = 0$ for $t = 2, \ldots, T$ and $t = 1$ the reference income category (why not a complete set of dummies?). We have $(I-1)+(T-1)(I-1) = T(I-1)$ parameters, so that there is again a 1-1 mapping from choice probabilities to parameters (find this mapping).

If we impose restrictions on the functional form, e.g. by replacing dummies by income category averages and by specifying the average utilities as

$$\sum_{j=2}^{I} \alpha_j d_{ij} + \gamma_i x_t$$

with $(I-1)+(I-1) = 2(I-1)$ parameters for $T(I-1)$ fractions. If $T > 3$ we can in principle add $(T-2)(I-1)$ parameters. The temptation is to add alternative attributes z_i in addition to the alternative specific dummies , i.e

$$\sum_{j=2}^{I} \alpha_j d_{ij} + \beta' z_i + \gamma_i' x_t$$

Note that β is only identified by restricting the effect of income. We could also have added a polynomial in x_t or interactions $d_{ij}x_t$, i.e. make the alternative specific constant dependent on x_t. Conclusion: We cannot identify the effect of more than $I-1$ alternative attributes without making functional from assumptions.

– Individual data. If one of the x_t is continuous, it is as if we have infinitely many subgroups, so that it seems that we can have both alternative specific dummies and alternative attributes. However, there is only identification by functional form restriction. Consider $I = 2$ with choice probabilities

$$\begin{aligned} p_2(x_t) &= \frac{e^{\alpha_2 + v_2(x_t)}}{1 + e^{\alpha_2 + v_2(x_t)}} \\ p_1(x_t) &= \frac{1}{1 + e^{\alpha_2 + v_2(x_t)}} \end{aligned}$$

Then

$$\alpha_2 + v_2(x_t) = \ln \frac{p_2(x_t)}{p_1(x_t)}$$

and we can obtain α_2 by the normalization $v_2(x_0) = 0$. Again alternative specific dummies is the most unrestrictive specification for the effect of alternative attributes.

- So with individual data the preferred specifications for the average utilities are
$$\sum_{j=2}^{I} \alpha_j d_{ij} + \gamma_i' x_t$$
or imposing restrictions on the attribute effects
$$\beta' z_i + \gamma_i' x_t$$

- An important extension is to allow the βs in the last specification to depend on x_t, i.e.
$$\beta = \beta_0 + \beta_1 x_t$$
with β_1 a $K \times L$ matrix of coefficients. The average utility is
$$\beta_0' z_i + x_t' \beta_1 z_i$$
Note that I have omitted $\gamma_i' x_t$ because the current specification restricts the γ_i (same remarks apply as for the case of alternative specific dummies versus the restricted $\beta' z_i$). Note that the coefficient vector on x_t is $z_i' \beta_1'$. This involves KL parameters that should be compared to $(I-1)L$ parameters in the unrestricted $\gamma_2, \ldots, \gamma_I$. We already imposed $K \leq I - 1$.

- The resulting MNL probabilities are
$$p_i(x_t; \beta) = \frac{e^{\beta_0' z_i + x_t' \beta_1 z_i}}{\sum_{j=1}^{I} e^{\beta_0' z_j + x_t' \beta_1 z_j}}$$
Assume that there are two attributes: quality z_1 with $\beta_{01} > 0$ and log price, z_2, so that $\beta_{02} < 0$. Remember that the cross price effects in the MNL model are proportional to the product of the market shares of the product whose price is raised and the product under consideration. Let x_t be education and let the marginal utility of quality be positively related to education, i.e. $\beta_{11} > 0$ ($\beta_{12} = 0$ as the marginal utility of price does not depend on education). A high priced, high quality alternative is mainly chosen by agents with a high level of education, who will choose less low price, low quality alternatives. If we concentrate on individuals with a high level of education, then if the price of such an alternative is raised,

they will substitute to other high price, high quality alternatives. If a new low price, low quality alternative is introduced, this will not affect their choices much.

- Conclusion: If the variation of the marginal utilities of attributes can be explained by agent characteristics x_t, then we can counter the effect of IIA by interacting the attributes with x_t. Note that this amounts to segmenting the market according to x_t in groups that value attributes similarly so that substitution occurs between products with similar attributes. In general it is hard to explain taste variation by observables (the work of marketing researchers), so that we should allow for unobservable taste differences. For economists getting the substitution pattern right is enough.

3 Estimation of parameters of MNL model with individual data

- If the choice probabilities are

$$p_i(x_t;\beta) = \frac{e^{\beta_0' z_i + x_t'\beta_1 z_i}}{\sum_{j=1}^I e^{\beta_0' z_j + x_t'\beta_1 z_j}}$$

- The data are a random sample $y_{1t},\dots,y_{It},x_t$ with $y_{it}=1$ if t chooses i and 0 if not.

- We also need $z_1,\dots,z_I$ usually obtained from product descriptions etc.

- We have

$$f(y_{1t},\dots,y_{It}|x_t,z_1,\dots,z_I;\beta) = \prod_{i=1}^I p_i(x_t;\beta)^{y_{it}}$$

  so that the log likelihood is

$$\ln L(\beta) = \sum_{t=1}^T \sum_{i=1}^I y_{it} \ln p_i(x_t;\beta)$$

- The MLE is computed in usual way and has the usual properties.

3.1 Dealing with IIA: Correlation among random utility components

3.1.1 Solution 1: Nested Multinomial Logit (NMNL)

- The key problem with IIA is that alternatives that are close, i.e. are close substitutes, have independent random components.

- A solution is to make the random components of similar alternatives dependent.

- Consider example with 4 alternatives. Choose the joint cdf of $\varepsilon_1, \ldots, \varepsilon_4$ as

$$F(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4) = e^{-\left(e^{-\frac{\varepsilon_1}{\sigma_1}} + e^{-\frac{\varepsilon_2}{\sigma_1}}\right)^{\sigma_1}} e^{-\left(e^{-\frac{\varepsilon_3}{\sigma_2}} + e^{-\frac{\varepsilon_4}{\sigma_2}}\right)^{\sigma_2}}$$

  Note that $\varepsilon_1, \varepsilon_2$ are dependent, as are $\varepsilon_3, \varepsilon_4$. We have $\sigma_1 \approx 1 - \rho(\varepsilon_1, \varepsilon_2)$.

- The choice probabilities are

$$p_{12}(v) = p_1(v) + p_2(v) = \frac{\left(e^{-\frac{v_1}{\sigma_1}} + e^{-\frac{v_2}{\sigma_1}}\right)^{\sigma_1}}{\left(e^{-\frac{v_1}{\sigma_1}} + e^{-\frac{v_2}{\sigma_1}}\right)^{\sigma_1} + \left(e^{\frac{-v_3}{\sigma_2}} + e^{-\frac{v_4}{\sigma_2}}\right)^{\sigma_2}}$$

  and

$$p_{1|1,2}(v) = \frac{e^{-\frac{v_1}{\sigma_1}}}{e^{-\frac{v_1}{\sigma_1}} + e^{-\frac{v_2}{\sigma_1}}}$$

  which has the MNL form. Hence

$$p_1(v) = p_{1|1,2}(v) p_{12}(v)$$

  This ARUM model is called the Nested Multinomial Logit (NMNL) model.

- Reconsider the auto type choice example with alternatives 1 (sedan), 2 (green SUV) and 3 (red SUV). Then

$$p_{2,3}(v) = \frac{2^\sigma e^{-v_2}}{2^\sigma e^{-v_2} + e^{-v_1}}$$

  and

$$p_{2|2,3}(v) = \frac{1}{2}$$

  If $v_1 = v_2$, then

$$p_{2,3}(v) = \frac{2^\sigma}{2^\sigma + 1}$$

Hence, if $\sigma \downarrow 0$, then $p_{2,3}(v) \to \frac{1}{2}$ which is the same as before the introduction of the red SUV.

- Specification of average utilities ($I = 4$)

$$\begin{aligned}
-v_{1t} &= \alpha' z_1 + \beta' w_1 + \gamma_1' x_t \\
-v_2 t &= \alpha' z_1 + \beta' w_2 + (\gamma_1 + \delta_1)' x_t \\
-v_{3t} &= \alpha' z_2 + \beta' w_3 + \gamma_2' x_t \\
-v_{4t} &= \alpha' z_2 + \beta' w_4 + (\gamma_2 + \delta_2)' x_t
\end{aligned}$$

with z_1, z_2 attributes of nests, w_1, w_2, w_3, w_4 attributes of the alternatives.

- Choice probabilities
  - Conditional choice probability

$$p_{1|1,2}(x_t; \beta, \delta, \sigma) = \frac{e^{\frac{\beta' w_1}{\sigma_1}}}{e^{\frac{\beta' w_1}{\sigma_1}} + e^{\frac{\beta' w_2}{\sigma_1} + \frac{\delta_1' x_t}{\sigma_1}}}$$

    Note $\frac{\alpha' z_1}{\sigma_1}$ and $\frac{\gamma' x_t}{\sigma_1}$ cancel.

  - From this MNL choice probability we can obtain MLE of $\frac{\beta}{\sigma_1}$ and $\frac{\delta_2}{\sigma_1}$.
  - Marginal choice probability

$$p_{12}(x_t; \alpha, \beta, , \delta, \sigma) =$$

$$= \frac{e^{\alpha' z_1 + \gamma_1' x_t + \sigma_1 \ln\left(e^{\frac{\beta' w_1}{\sigma_1}} + e^{\frac{\beta' w_2}{\sigma_1} + \frac{\delta_1' x_t}{\sigma_1}} \right)}}{e^{\alpha' z_1 + \gamma_1' x_t + \sigma_1 \ln\left(e^{\frac{\beta' w_1}{\sigma_1}} + e^{\frac{\beta' w_2}{\sigma_1} + \frac{\delta_1' x_t}{\sigma_1}} \right)} + e^{\alpha' z_2 + \gamma_2' x_t + \sigma_2 \ln\left(e^{\frac{\beta' w_3}{\sigma_2}} + e^{\frac{\beta' w_4}{\sigma_2} + \frac{\delta_2' x_t}{\sigma_2}} \right)}}$$

    We need to normalize by setting e.g. $\gamma_1 = 0$.

  - This is again an MNL choice probability but with an additional explanatory variable, usually called the inclusive value. Its coefficient measures the dependence (the smaller the more dependent) of the alternatives in the nest. It can be shown that

$$\mathrm{E}[\max\{u_{1t}, u_{2t}\}] = \alpha' z_1 + \gamma_1' x_t + \sigma_1 \ln\left(e^{\frac{\beta' w_1}{\sigma_1}} + e^{\frac{\beta' w_2}{\sigma_1} + \frac{\delta' 1 x_t}{\sigma_1}} \right)$$

- The expressions for the marginal and conditional choice probabilities suggest a two-step estimator, that gives consistent, but inefficient estimators of the parameters.

- Problem with NMNL is that one has to impose the nests. This choice may not be obvious.

3.1.2 Solution 2: Multinomial Probit (MNP)

see Section 4.

3.1.3 Solution 3: Mixed Logit

- We noted earlier that we can create realistic substitution between alternatives if attractiveness changes or new alternatives are introduced by letting the marginal utility of attributes of the alternatives, i.e. the β coefficients be dependent on observed characteristics of the agents.

- In general: If an agent places a high value on a particular attribute, the probability that he/she chooses an alternative that has much of that attribute, is relatively large (compared to alternatives that have little of the attribute). Hence, if an alternative with much of the attribute becomes less attractive, e.g. because its price increases, then the agent will substitute to other alternatives that have much of the attribute.

- Conclusion is that we should subdivide the population on the basis of the marginal utility of attributes.

- For this we can use observable agent characteristics, e.g. income, education etc. This may not explain much of the taste variation in the population and for that reason we should also consider dependence on unobservables.

- If we only allow for onobservable factors, then we obtain the random coefficient logit or mixed logit model:

$$\beta = \overline{\beta} + \nu$$

  with $\overline{\beta}$ the average marginal utility in the population and ν a mean 0 random variable that captures taste variation. The joint distribution of

ν has pdf $g(\nu;\lambda)$ with λ a vector of parameters. A popular choice is the multivariate normal or lognormal distribution.

- The resulting MNL probabilities are

$$p_i(x_t;\theta) = \int \dots \int \frac{e^{\overline{\beta}' z_i + \nu' z_i + \gamma_i' x_t}}{\sum_{j=1}^{I} e^{\overline{\beta}' z_j + \nu' z_j + \gamma_j' x_t}} g(\nu;\lambda) \mathrm{d}\nu$$

- The integral can be computed by numerical integration of by simulation with ν_r a draw from $g(\nu;\lambda_{)}$

$$\hat{p}_i(x_t;\theta) = \frac{1}{R}\sum_{r=1}^{R} \frac{e^{\overline{\beta}' z_i + \nu_r' z_i + \gamma_i' x_t}}{\sum_{j=1}^{I} e^{\overline{\beta}' z_j + \nu_r' z_j + \gamma_j' x_t}}$$

and these can be used in MSM or in simulated ML.

4 Multinomial Probit (MNP)

4.1 Setup

- If nests are not obvious it may be better to let the data decide on the correlation structure of the random components.

- Consider ARUM

$$u_i = -v_i + \varepsilon_i \quad , \; i = 1,\dots,I$$

with

$$\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_I \end{pmatrix} \sim N(0,\Sigma)$$

- Define the $I-1$ vectors

$$\eta_{(i)} = \begin{pmatrix} \varepsilon_1 - \varepsilon_i \\ \vdots \\ \varepsilon_I - \varepsilon_i \end{pmatrix} \qquad v_{(i)} = \begin{pmatrix} v_1 - v_i \\ \vdots \\ v_I - v_i \end{pmatrix}$$

- Note

$$\eta_{(i)} = A_i \varepsilon$$

with

$$A_i = \begin{pmatrix} 1 & 0 & \cdots & -1 & \cdots & 0 \\ 0 & 1 & \cdots & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & -1 & 0 & \vdots \\ \vdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & -1 & 0 & 1 \end{pmatrix}$$

so that

$$\eta_{(i)} \sim N(0, \tilde{\Sigma})$$

with

$$\tilde{\Sigma} = A_i \Sigma A_i'$$

- The choice probability is an orthant probability in the multivariate normal distribution
$$p_i(v) = Pr(\eta_{(i)} < v_{(i)}) = \Phi(v_{(i)}; \tilde{\Sigma})$$

- Example for $I = 3$

$$\begin{aligned} p_1(v) &= \Pr(\underbrace{\varepsilon_2 - \varepsilon_1}_{\eta_1} < \underbrace{v_2 - v_1}_{\tilde{v}_1}, \underbrace{\varepsilon_3 - \varepsilon_1}_{\eta_2} < \underbrace{v_3 - v_1}_{\tilde{v}_2}) \\ p_2(v) &= \Pr(\underbrace{\varepsilon_2 - \varepsilon_1}_{\eta_1} > \underbrace{v_2 - v_1}_{\tilde{v}_1}, \underbrace{\varepsilon_3 - \varepsilon_1}_{\eta_2} < \underbrace{v_3 - v_2}_{\tilde{v}_2 - \tilde{v}_1} + \underbrace{\varepsilon_2 - \varepsilon_1}_{\eta_1}) \\ p_3(v) &= \Pr(\underbrace{\varepsilon_3 - \varepsilon_1}_{\eta_2} > \underbrace{v_3 - v_1}_{\tilde{v}_2}, \underbrace{\varepsilon_2 - \varepsilon_1}_{\eta_1} < \underbrace{v_2 - v_3}_{\tilde{v}_1 - \tilde{v}_2} + \underbrace{\varepsilon_3 - \varepsilon_1}_{\eta_2}) \end{aligned}$$

- Hence we can identify

  - $\tilde{v}_1 = v_2 - v_1$, $\tilde{v}_2 = v_3 - v_1$
  - Variance matrix of η_1, η_2.

- Because we can divide all inequalities by any positive constant we divide by the standard deviation of η_1 or equivalently we set this equal to 1 or in terms of original parameters
$$\sigma_{11} + \sigma_{22} - 2\sigma_{12} = 1$$

- The other two components of $\tilde{\Sigma}$ are

$$\begin{aligned}\tilde{\sigma}_{12} &= \sigma_{23} - \sigma_{12} - \sigma_{13} + \sigma_{11} \\ \tilde{\sigma}_{22} &= \sigma_{33} + \sigma_{11} - 2\sigma_{13}\end{aligned}$$

- We have 3 equations (identified variances/covariances) with 6 unknowns. Hence, we must fix 3 parameters.

- Option 1: $\sigma_{11} = \sigma_{22} = \sigma_{33} = 1$, i.e. the random components have same variance.

- Option 2: $\sigma_{12} = \sigma_{13} = \sigma_{23} = 0$, i.e. random components are independent. This is not attractive. Reconsider example in which a red SUV was introduced. Then for alternative 1 (sedan)

$$p_{1|1,2}(v) = \Pr(\varepsilon_2 < v_2 - v_1 + \varepsilon_1)$$

and

$$p_{1|1,2,3}(v) = \Pr(\varepsilon_2 < v_2 - v_1 + \varepsilon_1, \varepsilon_3 < v_3 - v_1 + \varepsilon_1)$$

Given ε_1 under independence

$$\Pr(\varepsilon_2 < v_2 - v_1 + \varepsilon_1, \varepsilon_3 < v_3 - v_1 + \varepsilon_1) = \Pr(\varepsilon_2 < v_2 - v_1 + \varepsilon_1)\Pr(\varepsilon_3 < v_3 - v_1 + \varepsilon_1) <$$

$$< \Pr(\varepsilon_2 < v_2 - v_1 + \varepsilon_1)$$

If we integrate w.r.t. the pdf of ε_1, we have

$$p_{1|1,2,3}(v) < p_{1|1,2}(v)$$

i.e. the market share of the sedan decreases. Conclusion: It is the independence and not the extreme value distribution that causes the problem.

4.2 Estimation of MNP model

- We specify the average utility as in the MNL model, e.g.

$$-v_{it} = \beta' z_i + \gamma_i' x_t$$

with the normalization $\gamma_1 = 0$.

- The data are as in the MNL case.
- The conditional density that enters the likelihood is

$$f(y_{1t},\ldots,y_{It}|x_t,z_1,\ldots,z_I;\beta,\gamma) =$$

$$= \prod_{i=1}^{I} \Phi(\beta'(z_1-z_i)+(\gamma_1-\gamma_i)'x_t,\ldots,\beta'(z_I-z_i)+(\gamma_I-\gamma_i)'x_t;\tilde{\Sigma})^{y_{it}}$$

- The identified parameters were discussed earlier.
- Likelihood involves computation of $I-1$ dimensional normal cdf, i.e. an $I-1$ dimensional numerical integral.
- Consider $T=1000$, $I=10$ and number of alternative attributes is 5 and that of agent attributes is 10.
- The number of parameters is $\frac{1}{2}I(I-1)-1$ ($\tilde{\Sigma}$) plus 5 (β) plus $10(I-1)$ (γ_i-s) is $\frac{1}{2}I(I-1)+4+10(I-1)=139$
- Most numerical search algorithms require numerical first partial derivatives

$$\frac{\partial \ln L}{\partial \theta_j} \approx \frac{\ln L(\theta+\delta e_j)-\ln L(\theta-\delta e_j)}{2\delta}$$

  with e_j the j-th unit vector and δ a small number.
- Derivative requires 2*139=278 log likelihood evaluations. Hence iteration in algorithm requires 279000 9-dimensional numerical integrals. With 20 iterations we need 5580000 numerical integrals. If 1 integral requires 1 second the estimation will require 1550 hours or almost 65 days!

4.3 Simulation estimation of MNP

- Alternative to numerical integration is simulation. Choice probability

$$\Phi(\beta'(z_1-z_i)+(\gamma_1-\gamma_i)'x_t,\ldots,\beta'(z_I-z_1)+(\gamma_I-\gamma_i)'x_t;\tilde{\Sigma}) =$$

$$= \Pr(\eta_1 \le \beta'(z_1-z_i)+(\gamma_1-\gamma_i)'x_t,\ldots,\eta_{I-1} \le \beta'(z_I-z_1)+(\gamma_I-\gamma_i)'x_t;\tilde{\Sigma})$$

- Simulation algorithm

1. Draw R vectors
$$\begin{pmatrix} \eta_{1r} \\ \vdots \\ \eta_{I-1,r} \end{pmatrix} \quad r = 1, \ldots, R$$
from $N(0, \tilde{\Sigma})$. Do this once.

2. Simulate an estimate of the choice probability for observation t
$$\hat{p}_i(x_t; \beta, \gamma) =$$
$$\frac{1}{R}\sum_{r=1}^{R} I\left(\eta_{1r} \leq \beta'(z_1 - z_i) + (\gamma_1 - \gamma_i)'x_t, \ldots, \eta_{I-1,r} \leq \beta'(z_I - z_1) + (\gamma_I - \gamma_i)'x_t\right)$$

3. Simulate log likelihood
$$\widehat{lnL(\beta, \gamma)} = \sum_{t=1}^{T}\sum_{i=1}^{I} y_{it} \ln \hat{p}_i(x_t; \beta, \gamma))$$
and maximize with respect to the parameters.

4.3.1 Issues with this simple simulation approach

(i) The simulated probability $\hat{p}_i(x_t; \beta, \gamma)$ may be 0 (problem with log).

(ii) Because the simulated probability is a frequency it is a discontinuous function of the parameters. This makes numerical maximization of the log likelihood difficult.

(iii) The simulated MLE is not consistent, unless the number of simulations increases with the number of observations. Reason
$$\hat{p}_i(x_t; \beta, \gamma) = p_i(x_t; \beta, \gamma) + u_{it}$$
with u_{it} the simulation error with $\mathrm{E}(u_{it}) = 0$ and $\mathrm{Var}(u_{it}) = \frac{1}{R} p_i(x_t; \beta, \gamma)(1 - p_i(x_t; \beta, \gamma))$. Remember MLE is consistent because average log likelihood
$$\frac{1}{T}\ln L(\beta, \gamma) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{I} y_{it} \ln p_i(x_t; \beta, \gamma)) \xrightarrow{p}$$
$$\mathrm{E}_x\left[\sum_{i=1}^{I} p_i(x; \beta_0, \gamma_0) \ln p_i(x; \beta, \gamma))\right]$$

and the limit is uniquely maximized at the population value β_0, γ_0 of the parameters.

With simulated probabilities this becomes

$$\frac{1}{T}\ln \widehat{L(\beta,\gamma)} = \frac{1}{T}T\sum_{t=1}^{T}\sum_{i=1}^{I} y_{it}\ln(p_i(x_t;\beta,\gamma)+u_{it}) \xrightarrow{p}$$

$$\mathrm{E}_{x,u}\left[\sum_{i=1}^{I} p_i(x;\beta_0,\gamma_0)\ln(p_i(x;\beta,\gamma)+u_i)\right]$$

Only if u_i very small, i.e. if R is very large will the maximum of the limit be at the population parameters: the simulated MLE is biased, even in large samples.

4.3.2 Solutions

- Problems (i) and (ii) require smarter simulation. In the Appendix, we describe the Geweke-Hajivassiliou-Keane simulator that gives simulated choice probabilities that are continuous in the parameters and never 0 or 1.

- For problem (iii) we change the way we estimate the model: instead of ML we use the Generalized Method of Moments (GMM).

- Starting point is the conditional moment restriction (E_0 means that the expectation is over the population distribution)

$$\mathrm{E}_0[y_{it} - p_i(x_t;\beta_0,\gamma_0)|x_t] = 0$$

  so that for any function $w(x_t)$ we have the unconditional moment restriction

$$\mathrm{E}[(y_{it} - p_i(x_t;\beta_0,\gamma_0))w(x_t)] = 0$$

  The $w(x_t)$ are called the instruments.

- The corresponding sample moment condition is

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{I}(y_{it} - p_i(x_t;\beta,\gamma))w(x_t) = \frac{1}{T}\sum_{t=1}^{T} m(x_t;\beta,\gamma) = m_T(\beta,\gamma)$$

and the GMM estimator minimizes

$$m_T(\beta,\gamma)V_T^{-1}m_T(\beta,\gamma)$$

with V_T^{-1} a weighting matrix (chosen by you).

- How to choose $w(x_t)$?
  - Optimal (smallest variance of GMM estimator) instruments are with $\theta = (\beta'\gamma')'$ and the conditional moment function $cm(y_{it}, x_t; \theta) = y_{it} - p_i(x_t;\theta)$
    $$E\left[\frac{\partial cm}{\partial \theta}(x_t;\theta)|x_t\right] = \frac{\partial p_i}{\partial \theta}(x_t;\theta)$$
    This requires the choice probabilities.
  - One approach is to simulate the instruments (independently of the simulation of the choice probabilities), e.g. use GHK and take derivatives.
  - Other approach is to approximate the optimal instruments by polynomials in x and z.

- In simulated GMM or Method of Simulated Moments (MSM) replace choice probabilities by their simulators

$$\widehat{m_T(\theta)} = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{I}(y_{it} - \hat{p}_i(x_t;\theta))w(x_t) =$$

$$= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{I}(y_{it} - p_i(x_t;\theta))w(x_t) - \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{I} w(x_t)u_{it}$$

- If simulation is unbiased then $\mathrm{E}(u_{it}|x_t) = 0$ with expectation over simulations. Hence by the Law of Large numbers

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{I} w(x_t)u_{it} \xrightarrow{p} 0$$

so that

$$\widehat{m_T(\theta)} \xrightarrow{p} m_T(\theta)$$

Intuition: The simulation errors average out in estimation.

- Conclusion: MSM estimator is consistent if $T \to \infty$ even for finite R.
- Effect of simulation is only on variance of MSM. If we use R draws then the variance of the MSM estimator $\hat{\theta}_R$ is

$$V(\hat{\theta}_R) = (1 + \frac{1}{R})V(\hat{\theta}_\infty)$$

- In general the simulation errors average out of the estimating equations, i.e. the equation to which the estimator is the solution, if it is linear in the simulated function.

4.3.3 Appendix: Geweke-Hajivassiliou-Keane (GHK) simulator

- Write $\tilde{\Sigma}$ as the product of two lower triangular matrices (Choleski decomposition)

$$\tilde{\Sigma} = \Delta\Delta'$$

with

$$\begin{pmatrix} \delta_{11} & 0 & \cdots & \cdots & 0 \\ \delta_{21} & \delta_{22} & \cdots & 0 & 0 \\ \delta_{31} & \delta_{32} & \delta_{33} & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \delta_{I-1,1} & \cdots & \cdots & \cdots & \delta_{I-1,I-1} \end{pmatrix}$$

- Hence we have ($\overset{d}{=}$ means is distributed as)

$$\eta \overset{d}{=} \Delta\zeta \quad , \ \zeta \sim N(0, I)$$

because $\text{Var}(\eta) = \text{E}(\eta\eta') = \text{E}[\Delta\zeta\zeta'\Delta'] = \tilde{\Sigma}$.

- In sequel we consider $I = 4$ and the computation of $p_{1t}(\beta, \gamma)$. We define

$$c_i = \beta'(z_i - z_1) + (\gamma_i - \gamma_1)'x_t \quad , \ i = 2, 3, 4$$

- The integration region for the computation of $p_{1t}(\beta, \gamma)$ is

$$\begin{aligned} \eta_1 &\leq c_1 \\ \eta_2 &\leq c_2 \\ \eta_3 &= c_3 \end{aligned}$$

or

$$\begin{aligned}\zeta_1 &\leq \frac{c_1}{\delta_{11}}\\ \zeta_2 &\leq \frac{c_2-\delta_{32}\zeta_1}{\delta_{22}}\\ \zeta_3 &\leq \frac{c_3-\delta_{31}\zeta_1-\delta_{32}\zeta_2}{\delta_{33}}\end{aligned}$$

Hence

$$p_1(x_t;\beta,\gamma)=\int_{-\infty}^{\frac{c_1}{\delta_{11}}}\int_{-\infty}^{\frac{c_2-\delta_{32}\zeta_1}{\delta_{22}}}\int_{-\infty}^{\frac{c_3-\delta_{31}\zeta_1-\delta_{32}\zeta_2}{\delta_{33}}}\phi(\zeta_1)\phi(\zeta_2)\phi(\zeta_3)\mathrm{d}\zeta_3\mathrm{d}\zeta_2\mathrm{d}\zeta_1$$

We can also write the integral using an indicator function

$$p_1(x_t;\beta,\gamma)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}I\left(\zeta_1\leq\frac{c_1}{\delta_{11}},\zeta_2\leq\frac{c_2-\delta_{32}\zeta_1}{\delta_{22}},\zeta_3\leq\frac{c_3-\delta_{31}\zeta_1-\delta_{32}\zeta_2}{\delta_{33}}\right)\Phi\left(\frac{c_1}{\delta_{11}}\right)\cdot$$

$$\cdot\Phi\left(\frac{c_2-\delta_{32}\zeta_1}{\delta_{22}}\right)\Phi\left(\frac{c_3-\delta_{31}\zeta_1-\delta_{32}\zeta_2}{\delta_{33}}\right)\frac{\phi(\zeta_1)}{\Phi\left(\frac{c_1}{\delta_{11}}\right)}\frac{\phi(\zeta_2)}{\Phi\left(\frac{c_2-\delta_{32}\zeta_1}{\delta_{22}}\right)}\frac{\phi(\zeta_3)}{\Phi\left(\frac{c_3-\delta_{31}\zeta_1-\delta_{32}\zeta_2}{\delta_{33}}\right)}\mathrm{d}\zeta_3\mathrm{d}\zeta_2\mathrm{d}\zeta_1=$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Phi\left(\frac{c_1}{\delta_{11}}\right)\Phi\left(\frac{c_2-\delta_{32}\zeta_1}{\delta_{22}}\right)\Phi\left(\frac{c_3-\delta_{31}\zeta_1-\delta_{32}\zeta_2}{\delta_{33}}\right)\cdot$$

$$\cdot\phi\left(\zeta_1\middle|\zeta_1\leq\frac{c_1}{\delta_{11}}\right)\phi\left(\zeta_2\middle|\zeta_2\leq\frac{c_2-\delta_{32}\zeta_1}{\delta_{22}}\right)\mathrm{d}\zeta_2\mathrm{d}\zeta_1$$

where you should note that on the support of the truncated distributions the indicator is always 1.

- Simulation scheme

  1. Obtain draw from $\phi\left(\zeta_1\middle|\zeta_1\leq\frac{c_1}{\delta_{11}}\right)$ (see below how) and compute $\frac{c_2-\delta_{32}\zeta_1}{\delta_{22}}$. Use this result to obtain draw from $\phi\left(\zeta_2\middle|\zeta_2\leq\frac{c_2-\delta_{32}\zeta_1}{\delta_{22}}\right)$ and compute $\frac{c_3-\delta_{31}\zeta_1-\delta_{32}\zeta_2}{\delta_{33}}$. Do this step R times.
  2. Compute the simulated choice probability

  $$\hat{p}_i(x_t;\beta,\gamma)=\frac{1}{R}\sum_{r=1}^{R}\Phi\left(\frac{c_1}{\delta_{11}}\right)\Phi\left(\frac{c_2-\delta_{32}\zeta_{1r}}{\delta_{22}}\right)\Phi\left(\frac{c_3-\delta_{31}\zeta_{1r}-\delta_{32}\zeta_{2r}}{\delta_{33}}\right)$$

- Note that $\hat{p}_i(x_t;\beta,\gamma)$ is continuous in the parameters and strictly between 0 and 1.

- How to draw from $\phi\left(\zeta 1 \middle| \zeta_1 \leq \frac{c_1}{\delta_{11}}\right)$?

  - Use fact that if X has a uniform distribution on $[0,1]$, then for any cdf F, $Y = F^{-1}(X) \sim F$, because $F_Y(y) = \Pr(Y \leq y) = \Pr(F^{-1}(X) \leq y) = \Pr(X \leq F(y)) = F(y)$.

  - The cdf of the distribution of ζ_1 given $\zeta_1 \leq \frac{c_1}{\delta_{11}}$ is

$$\frac{\Phi(\zeta_1)}{\Phi\left(\frac{c_1}{\delta_{11}}\right)}$$

    for $\zeta_1 \leq \frac{c_1}{\delta_{11}}$.

  - The inverse cdf is

$$\Phi^{-1}\left(u\Phi\left(\frac{c_1}{\delta_{11}}\right)\right)$$

  - Hence if U is a draw from the uniform $[0,1]$ distribution, then we compute ζ_1 as

$$\zeta_1 = \Phi^{-1}\left(U\Phi\left(\frac{c_1}{\delta_{11}}\right)\right)$$

Figure 1: Integration regions of choice probabilities $I = 3$

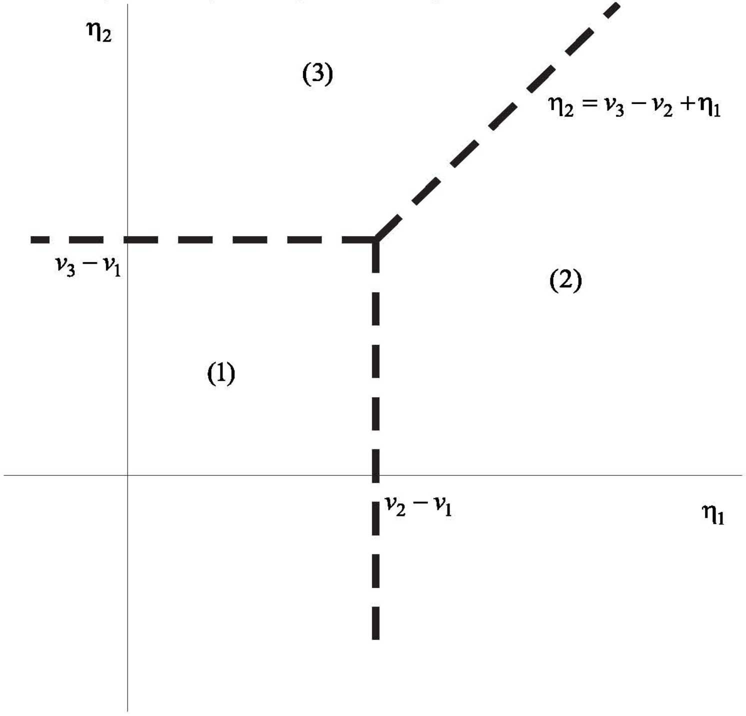
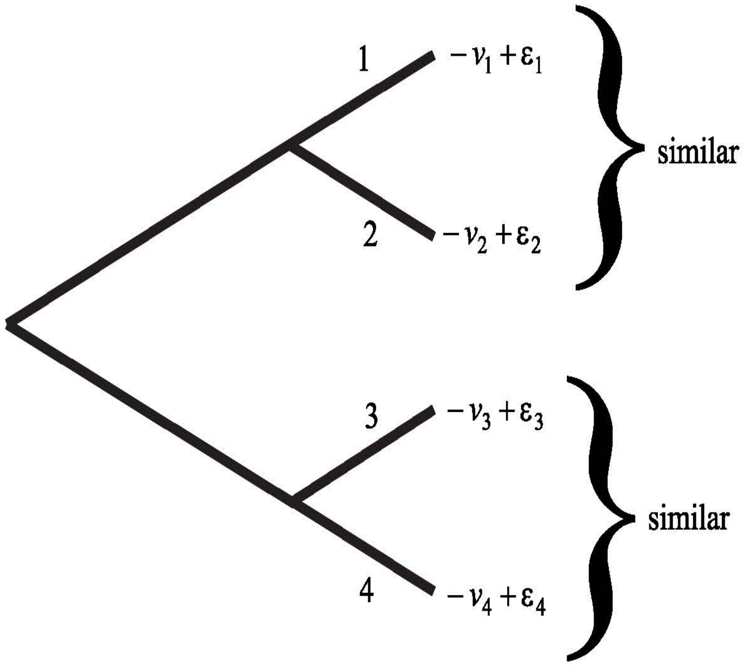


Figure 2: Nested choices


1
$-v_1 + \varepsilon_1$
2
$-v_2 + \varepsilon_2$
similar
3
$-v_3 + \varepsilon_3$
4
$-v_4 + \varepsilon_4$
similar