

CHOICE MODELING IN R

CLASS 5

Wiktor Budziński

CHOICE MODELING

STATED PREFERENCE METHODS

REVEALED VS. STATED DATA

Choice modeling research utilizes two types of data: revealed and stated preferences

Revealed preferences refer to market data, collected either through survey or obtained from retailer etc.

- They are based on actual past behavior

Stated preferences are exclusively survey-based, based on hypothetical “markets”

- Based on hypothetical behavior, conditional on the scenarios presented in the survey

REVEALED VS. STATED DATA

Revealed preferences

- Market behavior is assumed to be truthful
 - Actual behavior has consequences
- Often incomplete data
 - Consumer may not have knowledge about the unselected alternatives
 - We don't know the full set of alternatives or attributes
- Limited to current market context
 - Behavior may not be generalizable if market changes
 - Cannot make predictions about alternatives that are not in the market
- Variables often highly correlated
 - Lead to poor identification

Stated preferences

- Hypothetical behavior may be prone to biases
 - E.g., social-desirability bias
- Choice scenario designed by researcher
 - Full control over the choice set
- Hypothetical market is flexible
 - Researcher can introduce hypothetical alternatives or attributes that are currently out of scope of the market
- Variables often orthogonal by design

DISCRETE CHOICE EXPERIMENTS

The most popular stated preference method are so-called Discrete Choice Experiments

- In marketing sometimes called conjoint analysis

Survey respondents are presented with a sequence of hypothetical scenarios (choice sets) characterized by

- Alternatives (usually policy options or some variation of a good)
- Attributes (characteristics of the good)

Respondents are asked to choose the most preferred alternative

DISCRETE CHOICE EXPERIMENTS

It is recommended that

- The good in question is a public good (e.g., public policy)
- One of the attributes should be a binding payment mechanism (e.g., tax)
- Results of the survey has no 0 chance of affecting the real-world policy outcome
- Respondents should evaluate sequence of choices independently
- Choice scenario involve status quo/opt-out option
- The hypothetical scenario is believable for the respondent

DISCRETE CHOICE EXPERIMENTS

Czajkowski, M., Budziński, W., Campbell, D., Giergiczny, M., & Hanley, N. (2017). Spatial heterogeneity of willingness to pay for forest management. *Environmental and Resource Economics*, 68(3), 705-727.

	Alternative 1	Alternative 2	Alternative 3	Alternative 4
Protection of ecologically valuable forests	 Status quo Passive protection of 50% of the most ecologically valuable forests (1.5% of all forests)	 Status quo Passive protection of 50% of the most ecologically valuable forests (1.5% of all forests)	 Status quo Passive protection of 50% of the most ecologically valuable forests (1.5% of all forests)	 Substantial improvement Passive protection of 100% of the most ecologically valuable forests (3% of all forests, 100% increase)
Litter in forests	 Status quo No change in the amount of litter in the forests	 Partial improvement Decrease the amount of litter in the forests by half (50% reduction)	 Status quo No change in the amount of litter in the forests	 Partial improvement Decrease the amount of litter in the forests by half (50% reduction)
Infrastructure	 Status quo No change in tourist infrastructure	 Status quo No change in tourist infrastructure	 Partial improvement Appropriate tourist infrastructure in an additional 50% of the forests (50% increase)	 Substantial improvement Appropriate tourist infrastructure available in twice as many forests (100% increase)
Cost	0 PLN	10 PLN	25 PLN	100 PLN
Your choice	☐	☐	☐	☐

DISCRETE CHOICE EXPERIMENTS

Czajkowski, M., Hanley, N., & Nyborg, K. (2017). Social norms, morals and self-interest as determinants of pro-environment behaviours: the case of household recycling. *Environmental and Resource Economics*, 66(4), 647-670

Choice Situation 1.	Alternative 1	Alternative 2	Alternative 3
Method of sorting in household	Into 5 categories	Into 2 categories	None
Frequency of collection	Once every 4 weeks	Once every 2 weeks	Once every week
Monthly cost for your household	75 PLN	50 PLN	100 PLN
Your choice:	☐	☐	☐

Fig. 1 Example of a choice card

CHOICE MODELING

BASIC STATISTICS

BASIC STATISTICS

When first looking at choice data it is good to look at some basic characteristics of the data

- Share of choices of each alternative
- Serial choices, e.g., serial status-quo
- Attributes' characteristics of chosen and not chosen alternatives
- Bid acceptance curve

BASIC STATISTICS

When first looking at choice data it is good to look at some basic characteristics of the data

- Share of choices of each alternative
- Serial choices, e.g., serial status-quo
- Attributes' characteristics of chosen and not chosen alternatives
- Bid acceptance curve

Let's go to  

CHOICE MODELING

LATENT CLASS MODEL

UNOBSERVED PREFERENCE HETEROGENEITY

We've already seen that we can introduce observed preference heterogeneity by interacting attributes with consumer's characteristics

In practice, there is limit to how many interactions we can introduce

- We also may simply not have all relevant factors

Models with unobserved preference heterogeneity allow us to control for differences in tastes in a straightforward fashion

The main two approaches differ in types of heterogeneity they assume

- Latent class model assume discrete segments with varying tastes
- Mixed logit model assumes continuous preference heterogeneity

LATENT CLASS MODEL

We assume that there are C distinct classes of preferences (for example, market segments).

- Implicit assumption is that within each class individuals have the same preference profile

We do not know to which class each individual belongs, so we model it as a discrete latent variable

Each class has its own vector of coefficients. Specifically, we may assume that utility function within each class is given as:

$$U_{ij}^c = \beta^c \mathbf{X}_{ij} + \varepsilon_{ij}$$

β^c is just a vector of coefficients to be estimated similarly as in MNL

LATENT CLASS MODEL

Within each class probability of choosing alternative j is given by the simple MNL formula

$$P(y_i = j | \mathbf{X}_i, C_i = c) = \frac{\exp(\boldsymbol{\beta}^c \mathbf{X}_{ij})}{\sum_l \exp(\boldsymbol{\beta}^c \mathbf{X}_{il})}$$

As class membership is not observed we need to take this into account and integrate it out. For discrete distribution it is quite easy, as it only requires taking a weighted sum

$$P(y_i = j | \mathbf{X}_i) = \mathbf{E}[P(y_i = j | \mathbf{X}_i, C_i = c)] = \sum_{c=1}^C \pi_c P(y_i = j | \mathbf{X}_i, C_i = c)$$

$0 < \pi_c < 1$ is a probability of belonging to a given class. The probability for the last class does not have to be estimated.

LATENT CLASS MODEL

If the data has a (quasi-) panel structure then we need to take this into account as well

$$P(y_i | \mathbf{X}_i) = \sum_{c=1}^C \pi_c \left[\prod_t P(y_{it} | \mathbf{X}_{it}, C_i = c) \right]$$

Probabilities of class membership are usually also modelled as MNL, namely

$$\pi_c = \frac{\exp(\gamma_c)}{\sum_s \exp(\gamma_s)}$$

With γ 's being parameters to be estimated and $\gamma_C = 0$.

LATENT CLASS MODEL

We can also include some socio-demographic variables to account for the fact that different individuals have different probabilities of belonging to a given class

- Observed preference heterogeneity

$$\pi_c = \frac{\exp(\gamma_c + \gamma_c^{SD} SD_i)}{\sum_s \exp(\gamma_l + \gamma_l^{SD} SD_i)}$$

LATENT CLASS MODEL

Advantages

- Easy to estimate (no simulation required)
- Easy to interpret
- Can be used to combine different models (not only MNLs)

Disadvantages

- No. of coefficients can quickly become very high
- No. of classes needs to be prespecified
- During estimation local maxima are very likely. It is recommended to try several different starting values
- If no. of classes is high then usually some important coefficients become not significant for some classes (for example, cost)

Choice Situation 1.	Alternative 1	Alternative 2	Alternative 3
Method of sorting in household	Into 5 categories	Into 2 categories	None
Frequency of collection	Once every 4 weeks	Once every 2 weeks	Once every week
Monthly cost for your household	75 PLN	50 PLN	100 PLN
Your choice:	☐	☐	☐

Estimate latent class model on recycling data

- Consider models with 2 and 3 classes
- Add variables explaining class membership probability

EXERCISE

Choice Situation 1.	Alternative 1	Alternative 2	Alternative 3
Method of sorting in household	Into 5 categories	Into 2 categories	None
Frequency of collection	Once every 4 weeks	Once every 2 weeks	Once every week
Monthly cost for your household	75 PLN	50 PLN	100 PLN
Your choice:	☐	☐	☐

Estimate latent class model on recycling data

- Consider models with 2 and 3 classes
- Add variables explaining class membership probability

EXERCISE

Let's go to



CHOICE MODELING

POSTERIOR PROBABILITIES

POSTERIOR PROBABILITIES

We can use Bayes theorem to obtain conditional estimates of class probabilities

- By taking into account choices that individuals made we can gain more insight regarding to which class they belong

$$\tilde{\pi}_c^i = P(C_i = c | y_i) = \frac{\pi_c P(y_i = j | \mathbf{X}_i, C_i = c)}{\sum_{k=1}^C \pi_c P(y_i = j | \mathbf{X}_i, C_i = k)}$$

POSTERIOR PROBABILITIES

We can use Bayes theorem to obtain conditional estimates of class probabilities

- By taking into account choices that individuals made we can gain more insight regarding to which class they belong

$$\tilde{\pi}_c^i = P(C_i = c | y_i) = \frac{\pi_c P(y_i = j | \mathbf{X}_i, C_i = c)}{\sum_c \pi_c P(y_i = j | \mathbf{X}_i, C_i = k)}$$

Let's go to



POSTERIOR PROBABILITIES

If socio-demographic variables are used in the model, one can use conditional probabilities to obtain the expected profile of the respondent in the given class

$$SD_c = \frac{\sum_i \tilde{\pi}_c^i SD_i^c}{\sum_i \tilde{\pi}_c^i}$$

One can also obtain the conditional expected WTP by weighting WTP in each class with conditional probabilities

$$WTP_i^{cond} = \sum_c \tilde{\pi}_c^i WTP_c$$

POSTERIOR PROBABILITIES

If socio-demographic variables are used in the model, one can use conditional probabilities to obtain the expected profile of the respondent in the given class

$$SD_c = \frac{\sum_i \tilde{\pi}_c^i SD_i^c}{\sum_i \tilde{\pi}_c^i}$$

One can also obtain the conditional expected WTP by weighting WTP in each class with conditional probabilities

Let's go to  