Prof. Dr. S. Kranz
M. Kies

Summer term 2018
2018-05-07
Sheet 6

Exercises in Empirical Industrial Organization and Consumer Choice

(Presentation date currently unknown)

Exercise 5 - Nested Logit Model for customer data

This exercise strongly draws from Kenneth Train's training exercises for the mlogit package, in both structure and wording.[1]

Introduction and data set

The data set *HC* from the package **mlogit** contains data on the choice of heating and central cooling system for 250 single-family, newly built houses in California.
The alternatives are:

1. Gas central heat with cooling (*gcc*)
2. Electric central resistence heat with cooling (*ecc*)
3. Electric room resistence heat with cooling (*erc*)
4. Electric heat pump, which provides cooling also (*hpc*)
5. Gas central heat without cooling (*gc*)
6. Electric central resistence heat without cooling (*ec*)
7. Electric room resistence heat without cooling (*er*)

Heat pumps necessarily provide both heating and cooling such that heat pump without cooling is not an alternative.
The definition of the column variables can be found in Table 1.

[1] see Kenneth Train and Yves Croissant (2015). *Kenneth Train's exercises using the mlogit package for R.* chap. 2, pp. 10–16. URL: `http://cran.r-project.org/web/packages/mlogit/vignettes/Exercises.pdf` (visited on 06/09/2015)

column header	explanation
depvar	name of the chosen alternative
ich.alt	installation cost for the heating portion of the system.
icca	installation cost for cooling
och.alt	operating cost for the heating portion of the system
occa	operating costs for cooling
income	annual income of the household

Table 1: Explanation of the variables of the data set *HC* of the *R* package **mlogit**.

Note that the full installation cost of alternative *gcc* is *ich.gcc+icca*, and similarly for the operating cost and for the other alternatives with cooling.
The goal is to estimate a nested logit model on this data set. It is important to realize, that the definition of the nested logit model, which was given in the lectures and by Berry (1994)[2], does not quite fit to our data set. Berry (1994) used the utility function

$$u_{nj} = x_j'\beta - \alpha p_j + \xi_j + [\zeta_{ng} + (1-\sigma)\varepsilon_{nj}]$$

where n specified the customer, j the product and g the group. In particular, this implies, that the attributes of product j are identical to all customers. This makes sense in most markets, where the products are very similar (e.g. cars of the same model) or when only market data is available. In this data set however the costs for the heating systems vary by customer. We will therefore use the more general model of Wen and Koppelman (2001)[3], which is already implemented in the **mlogit** package.
Before diving into the Exercises, we will first clarify the theory. To do so, we will draw closely from Train (2015)[4].

Theory
Let the set of alternatives be partitioned into G nonoverlapping groups[5]. In contrast to the lectures, we do not need a default choice $j = g = 0$, as every family-home in our data set has exactly one heating system. The utility of customer n from product j in group g is defined as

$$u_{nj} = \underbrace{w_{ng} + y_{nj}}_{v_{nj}} + \varepsilon_{nj}$$

where w_{ng} depends only on variables that describe group g. These variables differ over groups but not over alternatives within each group. y_{nj} on the other hand depends on variables that

[2]Steven T Berry (1994). "Estimating discrete-choice models of product differentiation". In: *The RAND Journal of Economics*, pp. 242–262

[3]Chieh-Hua Wen and Frank S Koppelman (2001). "The generalized nested logit model". In: *Transportation Research Part B: Methodological* 35.7, pp. 627–641

[4]Kenneth Train (2015). "Nested Logit". In: *Discrete Choice Methods with Simulation*. Cambridge University Press. Chap. 4.2, pp. 77–88. URL: `http://eml.berkeley.edu/books/choice2.html` (visited on 06/10/2015)

[5]Sometimes called nests.

describe alternative j. These variables vary over alternatives within group g. v_{nj} on the other hand captures both kind of variables. All three variables vary (in contrast to Berry (1994)) in regard to the individual customer. All three variables are observable to the researcher.
It holds that $\varepsilon_n \coloneqq (\varepsilon_{n1}, ..., \varepsilon_{nJ})$ has the cumulative distribution

$$\exp\left(-\sum_g \left(\sum_{j\in g} e^{-\varepsilon_{nj}/\lambda_g}\right)^{\lambda_g}\right)$$

This distribution is a type of generalized extreme value distribution, which is a generalized form of the extreme value distribution used in class. The ε_{nj} are marginally univariate extreme value distributed, but correlated within each group. In other words, two products within a group are correlated, but two products from two different groups are uncorrelated, just as all products in the Berry Logit Model[6].
The parameter λ_g is a measure of the degree of independence in unobserved utility among the alternatives in group g. A higher value of λ_g means greater independence and less correlation. A value of $\lambda_g = 1$ indicates complete independence within group g, that is, no correlation. If this holds for all λ_g, then this model is reduced to the Standard Discrete Choice Model. Note that λ_g can therefore be seen similar to the σ of Berrys Nested Logit Model, even though we generalize by allowing each group to have differing degrees of interdependence.
Calculating the results of this model leads to the following choice probability for j of group g^*:

$$P_{nj} = \frac{e^{v_{nj}/\lambda_{g^*}}\left(\sum_{h\in g^*} e^{v_{nh}/\lambda_{g^*}}\right)^{\lambda_{g^*}-1}}{\sum_g \left(\sum_{h\in g} e^{v_{nh}/\lambda_g}\right)^{\lambda_g - 1}}$$

Additionally it holds for the probability that a given customer n choses some product of group g^* that

$$P_{ng^*} = \frac{e^{w_{ng^*}+\lambda_{g^*} I_{ng^*}}}{\sum_g e^{w_{ng}+\lambda_g I_{ng}}} \qquad \text{where} \qquad I_{ng} = \log \sum_{h\in g} e^{y_{nh}/\lambda_g}$$

Stated in words, the probability of choosing an alternative in g^* takes the form of the logit formula, as if it resulted from a model for a choice among groups.
The quantity I_{ng} is often called *inclusive value* or *inclusive utility* of group g. Sometimes the wording *log-sum term* is used. The inclusive value is important, as it captures the information of the "in group" model. Indeed it holds that $\lambda_g I_{ng}$ is the expected utility that customer n receives from the choice of alternatives within the group g. Keep in mind, that the Nested Logit Model can be seen as two Standard Logit Models after each other. First a decision regarding the groups is taken, then within the group a decision regarding the specific choice. The inclusive value captures information about the latter.

[6]To avoid confusion: Both, the "Berry Logit Model" and "Nested Logit Model on Market Level' (which is defined in Berry (1994)) are from the same author, but are different models. The Berry Logit Model as defined in the lectures is the basic one without groups.

The estimation of this model can be done with Maximum Likelihood Estimation. The R-Package *mlogit* provides a useful framework to do so. For more information regarding the Nested Logit Model on individual data please have a look at Train (2015).

Tasks

(a) Cooling as nesting parameter:

1. Run a nested logit model on the data for two groups $g_1 = \{gcc, ecc, erc, hpc\}$ [with cooling] and $g_2 = \{gc, ec, er\}$ [without cooling]. Assume that both groups have the same degree of interdependence, i.e. $\lambda_1 = \lambda_2$ and that the intercept is 0. Your regressors should be *ich*, *och*, *icca* and *occa* of the respective alternative, the two interaction variables *inc.room* and *inc.cooling* [which should be $income \cdot \mathbb{1}_{Room}$ and $income \cdot \mathbb{1}_{Cooling}$] and a dummy variable capturing $\mathbb{1}_{Cooling}$. Here $\mathbb{1}_{Cooling}$ is defined as the Indicator function of the "cooling alternatives" which should be 1 if it is a cooling alternative and 0 otherwise. $\mathbb{1}_{Room}$ is defined accordingly.

   To make this task easier, we will break it up into sub-tasks.

   i. Import your data with the *data()* command. The data is provided in wide-format, so you have to transform it into long-format, so that it resembles Exercise 4 of the R-Tutor data set *ps_2a_logit*, i.e.

```
> head(l.HC)
      n   j     y  icca occa   ich  och income
1.ec  1  ec FALSE 27.28 2.95 24.50 4.09     20
1.ecc 1 ecc FALSE 27.28 2.95  7.86 4.09     20
1.er  1  er FALSE 27.28 2.95  7.37 3.85     20
1.erc 1 erc  TRUE 27.28 2.95  8.79 3.85     20
1.gc  1  gc FALSE 27.28 2.95 24.08 2.26     20
1.gcc 1 gcc FALSE 27.28 2.95  9.70 2.26     20
```

   ii. Add two columns, one for the cooling modes and one for the room.modes (i.e. *erc* and *er*).

   iii. Installation and operating costs of cooling do not capture that they are only relevant if indeed cooling is chosen. Set those costs (i.e. *icca* and *occa*) to 0 for non-cooling systems.

   iv. Create two additional variables, *inc.cooling* and *inc.room*, which should be the income if the row corresponds to the "cooling-group" and "room-group", respectively and 0 if not.

Your final table should look like this:

```
> head(l.HC)
      n   j     y  icca  occa   ich  och income cooling.mode room.mode inc.cooling inc.room
1.ec  1  ec FALSE  0.00  0.00 24.50 4.09     20         FALSE      FALSE           0        0
1.ecc 1 ecc FALSE 27.28  2.95  7.86 4.09     20          TRUE      FALSE          20        0
1.er  1  er FALSE  0.00  0.00  7.37 3.85     20         FALSE       TRUE           0       20
1.erc 1 erc  TRUE 27.28  2.95  8.79 3.85     20          TRUE       TRUE          20       20
1.gc  1  gc FALSE  0.00  0.00 24.08 2.26     20         FALSE      FALSE           0        0
1.gcc 1 gcc FALSE 27.28  2.95  9.70 2.26     20          TRUE      FALSE          20        0
```

v. Using the *mlogit.data()* function, shape the data set into a data set, which can be recognized by the *mlogit()* function.[7]

vi. Estimate the wanted model and show a summary of the result.

2. The estimated coefficient of the inclusive value, the value denoted with *iv* in the summary, is 0.59.[8] What does this estimate tell you about the degree of correlation in unobserved factors over alternatives within each group?

3. Test the hypothesis that the coefficient of the inclusive value is 1.0 (the value that it takes for a standard logit model.) Can the hypothesis that the true model is standard logit be rejected?

(b) Are room alternatives in one group and central alternatives in the other group a better model?

1. Re-estimate the model with the room alternatives in one group and the central alternatives in another group. (Note that a heat pump is a central system.)

2. What does the estimate imply about the substitution patterns across alternatives? Do you think the estimate is plausible?

3. Is the λ significantly different from 1?

4. How does the value of the log-likelihood function compare for this model relative to the model in exercise (a), where the cooling alternatives are in one group and the heating alternatives in the other group?

(c) What happens, when we allow different interdependence values for the groups?

1. Rerun the model that has the cooling alternatives in one group and the non-cooling alternatives in the other group (like for exercise (a)), with a separate interdependence variable λ_g for each group. Show a summary of the results.

2. Which group is estimated to have the higher correlation in unobserved factors? Can you think of a real-world reason for this group to have a higher correlation?

3. Are the two interdependence variables significantly different from each other? That is, can you reject the hypothesis that the model in exercise (a) is the true model?

[7] We could have used *mlogit.data()* on the wide data set, but for didactic reasons we wanted to reshape the data set manually.

[8] Keep in mind, that for the "group model" $w_{ng} + \lambda_g I_{ng}$ is estimated. The coefficient of I_{ng} is therefore λ_g.

(d) Rewrite the code to allow three groups. For simplicity, estimate only one interdependence variable which is applied to all three groups. Estimate a model with alternatives *gcc*, *ecc* and *erc* in a group, *hpc* in a group alone, and alternatives *gc*, *ec* and *er* in a group. Does this model seem better or worse than the model in exercise (a), which puts alternative *hpc* in the same group as alternatives *gcc*, *ecc* and *erc*?