So1 Task - Efficient Computation

Description

The task at hand is to develop code for an implementation of a mathematical calculation. The equation to be calculated is described in the next section. We look for a fast and efficient solution. You can use any language and libraries you think are suitable (see output for more details).

Equation

The goal is to calculate probabilities for consumers buying certain consumer packaged goods. The Logit model is a frequently employed model to achieve this task. In the Logit model the probability that consumer n purchases product j at time t is given by

$$P_{njt} = \frac{e^{U_{njt}}}{\sum_k e^{U_{nkt}}}$$

U_{njt} is the "utility" product j offers to consumer n at time t. The purchase probability for one product j depends on the utility of all products k (including j) that are available to consumer n at time t (this set of available products is usually referred to as the "choice set").

The utility is given by

$$U_{njt} = \sum_k \alpha_{nk} \cdot Dummy_k + \beta_n \cdot Price_{jt}$$

$Dummy_k$ is a dummy variable that is 1 for product k=j, and 0 for all other products. $Price_{jt}$ is product j's price at time t. α_{nk} and β_n are the consumer specific preferences for different variables (these are estimated by the Logit model). $\alpha_{nj} \cdot Dummy_j$ is the time invariant utility that product j offers to consumer n. This part of the utility function is often referred to as the products' alternative specific constant (ASC). $\beta_n \cdot Price_{jt}$ is the time variant (dis-)utility ("dis" because beta usually is negative) induced by the products' prices.

Data

For calculating the choice probabilities you receive two files: "Coef.csv" and "Data.csv".

"Coef.csv" includes the coefficients for the utility function. For each consumer n the consumer specific preferences for the variables in the utility function α_n and β_n are provided. The first column "n" in the user index n. The other columns ("alpha_ASC_[number of alternative]" and "beta_Price") are the coefficients.

"Data.csv" includes the data for the utility function. The first three columns are indexes for user ("n"), time ("t") and product ("j"). Note that t is not a date but a counter over the consumers shopping trips that were observed in the data. Column "j" is a unique product identifier. The next

columns are the product specific dummy variables $Dummy_j$ and the products' prices $Price_{jt}$. A good starting point might be to get familiar with the indexes of the data.

Please note: The system of time invariant utilities is normalized by setting one dummy variable (in this case the dummy for alternative 9) to 0 - thus you will not find columns "alpha_ASC_9" in Coef.csv and column "ASC_9" in the Data.csv.

Output

Please provide both your (executable) code and a result file that includes all combinations of the three indices n, t, and j you find in data.csv as well as the corresponding choice probabilities (preferably ".csv"-format). Please include in your code a runtime measurement of the mathematical calculation, excluding data loading and data saving.

Literature

Train, K. (2009): Discrete Choice Methods with Simulation, 2nd Edition, Cambridge University Press, http://elsa.berkeley.edu/books/choice2.html