Computational Economics 2020

Exercise: Unconstrained Optimization

Discrete-Choice Logit Models and Maximum Likelihood Estimation

Discrete-choice models. In a discrete choice model, each of N consumers chooses one of J mutually exclusive and totally exhaustive alternatives. Each alternative $j \in \{1, 2, \ldots, J\}$ that is considered by consumer $n \in \{1, 2, \ldots, N\}$ is characterized by K attributes. The utility u_{nj} of consumer n choosing alternative j is

$$u_{nj} = \beta^\top x_{nj} + \varepsilon_{nj},$$

where $x_{nj} \in \mathbb{R}^K$ is the vector of observed characteristics for alternative j considered by consumer n. The vector $\beta \in \mathbb{R}^K$ indicates consumers' preferences on the observed characteristics. Note that the vector of parameters β does not depend on the alternative j and does not vary over consumers n. The error term ε_{nj} summarizes the influence of unobserved characteristics and is treated as random by the researcher.

In the data, the researcher observes the characteristics x_{nj} for all n and j and each consumer's decision y_{nj}. If consumer n chooses alternative j then $y_{nj} = 1$, otherwise $y_{nj} = 0$. The researcher's objective is to infer consumers' preferences β.

Logit choice probabilities. Let x_n be the collection of x_{nj} over all products, that is, $x_n = (x_{nj})_{j=1}^J$. Assuming ε_{nj} is independent and identically distributed with the type-1 extreme value distribution, the choice probability of consumer n choosing alternative j is given by

$$P_{nj}(x_n|\beta) = \frac{e^{(\beta' x_{nj})}}{\sum_{\ell=1}^{J} e^{(\beta' x_{n\ell})}}.$$

Maximum Likelihood Estimation. Maximum Likelihood Estimation (MLE) can be applied to estimate β. Given the choice probability P_{nj}, the probability that consumer n chooses the alternative that (s)he was actually observed choosing can be expressed as

$$P_n(\beta) = \prod_{j=1}^{J} P_{nj}(x_n|\beta)^{y_{nj}}.$$

Assuming that each consumer's choice is independent of that of other consumers, the probability of each person in the sample choosing the observed alternative is

$$L(\beta) = \prod_{n=1}^{N} \prod_{j=1}^{J} P_{nj}(x_n|\beta)^{y_{nj}}.$$

The log-likelihood function is

$$LL(\beta) = \sum_{n=1}^{N} \sum_{j=1}^{J} \log\left(P_{nj}(x_n|\beta)\right) \times y_{nj}.$$

The maximum likelihood estimator is defined as

$$\hat{\beta} = \arg\max_{\beta} LL(\beta).$$

In sum, MLE of discrete-choice logit models requires the researcher to solve an unconstrained maximization problem. The objective function is $LL(\beta)$ and the decision variables are $\beta \in \mathbb{R}^K$.

Data set. The file **data_CA_houses.txt** is a text file of data for households' choices among five alternative heating systems. The observations consist of single-family houses in California that were newly built and had central air-conditioning. The data set contains 19 variables for 900 observations. Data are free format with one line per observation. Variables are in the following order: *idcase*, *depvar*, *ic1*, *ic2*, *ic3*, *ic4*, *ic5*, *oc1*, *oc2*, *oc3*, *oc4*, *oc5*, *income*, *agehed*, *rooms*, *ncost1*, *scost1*, *mountn*, *valley*. Five types of heating systems are considered to have been possible.

(1) gas central,

(2) gas room,

(3) electric central,

(4) electric room,

(5) heat pump.

The descriptions for the 19 variables in the data set are as follows.

(1) *idcase* gives the observation number (1-900)

(2) *depvar* identifies the chosen alternative (1-5)

(3) *ic1* is the installation cost for a gas central system

(4) *ic2* is the installation cost for a gas room system

(5) *ic3* is the installation cost for a electric central system

(6) *ic4* is the installation cost for a electric room system

(7) *ic5* is the installation cost for a heat pump

(8) *oc1* is the annual operating cost for a gas central system

(9) *oc2* is the annual operating cost for a gas room system

(10) *oc3* is the annual operating cost for a electric central system

(11) *oc4* is the annual operating cost for a electric room system

(12) *oc5* is the annual operating cost for a heat pump

(13) *income* is the annual income of the household

(14) *agehed* is the age of the household head

(15) *rooms* is the number of rooms in the house

(16) *ncostl* identifies whether the house is in the northern coastal region

(17) *scostl* identifies whether the house is in the southern coastal region

(18) *mountn* identifies whether the house is in the mountain region

(19) *valley* identifies whether the house is in the central valley region

The two attributes of the five alternative heating systems, namely, installation cost and annual operating cost, take different values for each alternative. Therefore, there are five installation costs (one for each of the 5 systems) and five operating costs for each data point. To estimate the logit model, the researcher needs data on the attributes of all the alternatives, not just the attributes for the chosen alternative. For example, it is insufficient for the researcher to determine how much was paid for the system that was actually installed (e.g., from the bill for the installation). The researcher needs to determine how much it would have cost to install each of the other four heating systems if they had been installed. The importance of costs in the choice process (i.e., the coefficients of installation and operating costs) is determined through comparison of the costs of the chosen system with the costs of the non-chosen systems.

For the provided data set, the costs were calculated as the amount the system would cost if it were installed in the house, given the characteristics of the house (such as size), the price of gas and electricity in the house location, and the weather conditions in the area (which determine the necessary capacity of the system and the amount of time it will be run in a year.) These cost are conditional on the house having central air-conditioning. (That's why the installation cost of gas central are lower than those for gas room: the central system can use the air-conditioning ducts that have been installed.)

Use MLE to estimate a logit model with installation cost and annual operating cost as the explanatory variables. Report running times, number of major iterations, number of function/gradient/Hessian evaluations for your implementation.