Choice Models in Operations

Homework 2 : Multinomial logit models

Instructor: Srikanth Jagabathula *Due: Friday, Nov 2, 2012*

1. **Value of choice models.** The purpose of this homework problem is to illustrate the value of choice models.

   You are a seller selling two mutually substitutable products and you are not sure how to price them. For simplicity, assume you are a monopolist and everyone must purchase either one of your two products or nothing. In order to determine the optimal pricing, you decide to model the demand observed for each of the products.

   First you consider a simple constant elasticity model and model the demand obtained for each product as:

   $$\begin{aligned}\log d_1 &= \alpha_1 + \eta_{11}\log p_1 + \eta_{12}\log p_2 + \varepsilon_1\\ \log d_2 &= \alpha_2 + \eta_{21}\log p_1 + \eta_{22}\log p_2 + \varepsilon_2\end{aligned}$$

   where d_i denotes the demand for product i, p_i denotes the price of product i, η_{ij} denotes the elasticity of the demand of product i with respect to the price of product j. ε_1 and ε_2 are error terms.

   You then consider using the choice modeling approach to model the demands. In order to learn the parameters of your models, you only have historical data on the demand of each of the products at various price points. You don't have any other information. Given this, answer the following questions:

   (a) What is the specification of the MNL model you would use to model this problem? Give a detailed description of the attributes and coefficients in your model. *Clearly* list the assumptions you had to make.

   (b) How many parameters do you need to learn in the MNL model you specified? Can you think of a model with fewer parameters? More parameters? If so, briefly describe the specifications. If you had more than two products, how does the number of parameters scale with the number of products n.

   (c) What algorithms would you use to learn the constant-elasticity model and the MNL model?

   (d) For the case of n products, compare the number of parameters you need to learn for constant-elasticity model and the MNL model. Which model is 'simpler'?

   (e) If you had access to demand for only one price vector (p_1, p_2), which model would you use and why?

   (f) Do you see value to using choice models? Why or why not?

2. **Interpreting MNL coefficients.** The purpose of this homework problem is to work you through the interpretation of MNL coefficients.

   Fire up R and issue the following set of commands:

```
R> library("mlogit")
R> data("Train", package="mlogit")
R> Tr <- mlogit.data(Train, shape="wide", choice="choice",
varying=4:11, sep="", alt.levels=c(1,2), id="id")
R> ml.Train <- mlogit(choice~price+time+change+comfort | -1, Tr)
R> summary(ml.Train)
```

The first command above loads the mlogit library needed to fit MNL models to data. The second command loads the dataset named "Train" from the package mlogit. Type

```
R> head(Train, 30)
```

to view the first 30 rows of the dataframe "Train." The data comprises choices made between two modes of transportation named 1 and 2. The column "id" denotes the identifier for the individual: all rows with the same id correspond to the same individual. The column "choiceid" represents the virtual choice situation presented to the individual where he is asked to pick between two train tickets with different sets of attributes of price, time of travel, change (number of changes required to get to the destination), and comfort (represented as either 0, 1, or 2 with 0 being the most comfortable class). The column "choice" contains the alternative (1 or 2) chosen by the individual in each choice situation. The next four columns contain the values of attributes (price, time, change, and comfort) of the first alternative and the last four columns contain the values of attributes for the second alternative. Here, the price is measured in cents of guilders (1 guilder $\approx$ 2.20371 euros) and the time is measured in minutes.

The data is in the so-called "wide" format i.e., each row corresponds to one choice situation. The number of columns scales as the number of products times the number of features per product. The parameter "shape" in the above command specifies that the data is in the wide format. An alternative format is the so-called long-format. If the data is not already in the "long" format, the mlogit package converts it into the long format. For instance, run the following command

```
R> head(Tr, 30)
```

to see the first 30 columns of the same dataset in the long format. Specifically, for each choice instance, you will see n rows ($n = 2$ in this case) – one for each alternative. The number of columns now only scales as the number of features.

After fitting the model, answer the following questions:

(a) What are the coefficients of price, time, change, and comfort variables? Do the signs of the coefficients make intuitive sense?

(b) What is the price in euros that an individual is willing to pay for an hour decrease in travel time? What about for decreasing the number of changes by 1?

(c) What are the price and time elasticities of the alternatives?

3. **Simulation.** The purpose of this problem is to illustrate the use of simulation to test various model specifications. This problem requires you to write computer code. You can use the language of your choice. Please attach to your solutions the code you write.

   (a) First write an MNL model simulator. The inputs to your program must be the number of alternatives n, an $n \times m_1$ matrix of variables in which each row corresponds to the vector of m_1 alternative specific variables with generic coefficients, an $n \times m_2$ matrix of alternative specific variables with alternative specific coefficients, a length m_1 vector of generic coefficients, and an $n \times m_2$ matrix of alternative specific coefficients, and finally a length n vector of intercepts. The program must also take the number of samples N to be generated as an input.

   The output of the program should be a matrix with N rows, where each row corresponds one of the N samples. The first column will be just a row identifier with the ith row taking value i. The second column should be the choice made the individual. The individual should choose alternative j with probability equal to the MNL choice probability. The remaining columns list the attributes of the alternatives. In particular, there would be $n(m_1 + m_2)$ columns with $m_1 + m_2$ columns for each alternative containing the values of the attributes.

   (b) Use your MNL model simulator to perform the following tasks. Take $n = 2$ and consider the MNL model $V_j = \beta p_j$, where p_j is the price of alternative j and β is the generic price coefficient. Take $p_1 = 5$ and $p_2 = 6$. Start with $\beta = -1$ and use these parameter values in your MNL simulator to generate $N = 1000$ samples. Store the resultant matrix in a text file with the columns separated by tabs. If you are working in MATLAB, you may use the following command:

```
dlmwrite('data_frame.txt', m, 'delimiter', '\t')
```

   where "data_frame.txt" is the name of the text file I am using and "m" is the matrix output from the MNL simulator. Using the R code I have posted on the course website along with this homework, fit the following MNL models to the simulated data:

   - $V_j = \beta p_j$
   - $V_j = \alpha_j + \beta p_j$ with one of the α_js set to zero
   - $V_j = \delta_j p_j$.

   Repeat the above procedure with $\beta = -2$. Repeat it again with $\beta = -2$ but this time adding the no-purchase option ($p_0 = 0$) to the mix. Please explain the results you've obtained.

   (c) Now fix the following model: $n = 3$ and $V_{\text{blue}} = 3.584 - 0.75 p_{\text{blue}}$, $V_{\text{red}} = 1.99 - 0.5 p_{\text{red}}$, and $V_0 = 0$, where as discussed in the class, p_{red} and p_{blue} are respectively the prices of red and blue cars. Verify that when the price vector is $(p_{\text{red}}, p_{\text{blue}}) = (5, 6)$ the choice probabilities are $\Pr(\text{red}) = 0.3$, $\Pr(\text{blue}) = 0.2$, and $\Pr(0) = 0.5$. This is the same example we used in class. We will use this model as the ground-truth to simulate purchase probabilities.

Given any simulated data, we will fit two different models: $V_j = \beta p_j$ and $V_j = \delta_j p_j$ and compare their performance. We will test the performance of each model as follows: generate 100 different test price vectors $(p_{\text{red}}, p_{\text{blue}})$ by sampling each of the prices independently and uniformly at random from the interval $[0, 10]$. For each price vector, compute the true sales $S^{\text{true}}$of the blue car when only the blue car and the no-purchase option are available. Compute the sales of the blue car under the test model $\hat{S}$. Compute the relative error $|S^{\text{true}} - \hat{S}|/S^{\text{true}}$. Finally, compute the average relative error over the 100 different price vectors.

With the above description, perform the following tasks:

i. Generate 100 different *training* price vectors $(p_{\text{red}}, p_{\text{blue}})$ by sampling each price independently and uniformly at random from the interval $[0, 10]$.

ii. For each training price vector,
   - use the ground-truth model to simulate 1000 samples of choices.
   - Fit the two models $V_j = \beta p_j$ and $V_j = \delta_j p_j$ to this data. In order to fit the models, use the closed-form expressions for the parameters derived in class (available in the writeup posted).
   - Compute average relative errors over 100 test price vectors as described above for the two models.

iii. Plot a histogram of the 100 relative errors for the two models used above. Compare the histograms of the two models. What do you conclude?