Problem Set 4

Topics in Advanced Econometrics (ResEcon 703)
University of Massachusetts Amherst

Due: December 9, 11:30 am ET

Rules

Email a single .pdf file of your problem set writeup, code, and output to `mwoerman@umass.edu` by the date and time above. You may work in groups of up to three and submit one writeup for the group, and I strongly encourage you to do so. This problem set requires you to code your own estimators, rather than using R's "canned" routines (e.g., `glm()` and `mlogit()`), except where indicated in problem 2.

Data

Download the file `camping_dataset.zip` from the course website. This zipped file contains the dataset `camping.csv`, which you will use for this problem set. This dataset contains simulated data on the state park choice of 1000 visitors who camped at one of five Massachusetts State Parks. See the file `camping_description.txt` for a description of the variables in the dataset.

Problem 1: Simulation-Based Estimation

We are again estimating a mixed logit model of state park choice among campers—as in problem 2 of problem set 3—but we are now estimating the model "by hand" to better understand the simulation-based estimation method.

a. Model the camping park choice as a mixed logit model. Express the representative utility of each alternative as a linear function of its cost, time, and setting—mountain or beach—with a fixed coefficient on cost and random coefficients on time and mountain. That is, the representative utility to camper n from park j is

$$V_{nj} = \beta_1 C_{nj} + \beta_{2n} T_{nj} + \beta_{3n} M_j$$

where C_{nj} is the cost to camper n of traveling to and camping at park j, T_{nj} is the time for camper n to travel to park j, M_j is a binary indicator if park j is in the mountains. Model β_1 as a fixed coefficient and β_2 and β_3 as random with a normal distribution:

$$\beta_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$
$$\beta_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$$

Estimate the parameters of this model by maximum simulated likelihood estimation; use 100 draws for your simulation and set a seed of 703 for replication. The following steps can provide a rough guide to creating your own maximum simulated likelihood estimator:

I. Set a seed of 703 for replication.

II. Draw 200,000 standard normal random variables (2 random coefficients $\times$ 100 draws $\times$ 1000 campers).

III. Create a function to simulate choice probabilities for one camper:

   i. The function should take a set of parameters, the random draws for one camper, and the data for one camper as inputs: `function(parameters, draws, data)`.

   ii. Transform the standard normal draws into the correct distributions using the distribution parameters.

   iii. Calculate the representative utility for every alternative for each draw.

   iv. Calculate the conditional choice probability for every alternative for each draw.

   v. Calculate the simulated choice probability for every alternative as the mean over all draws.

IV. Create a function to calculate simulated log-likelihood:

   i. The function should take a set of parameters, the random draws for all campers, and the data for all campers as inputs: `function(parameters, draws, data)`.

   ii. Simulate choice probabilities for every alternative for each camper—that is, call your previous function for each camper.

   iii. Sum the log of the simulated choice probability for each camper's chosen alternative.

   iv. Return the negative of the log of simulated likelihood.

V. Maximize the simulated log-likelihood (by minimizing its negative) using `optim()`. Your call of the `optim()` function may look something like:

```
optim(par = your_starting_guesses, fn = your_second_function,
      data = your_data, draws = your_draws,
      method = 'BFGS', hessian = TRUE)
```

A few additional notes on using the `optim()` function to estimate this mixed logit model:

- This estimation could potentially take hours to converge. To speed up convergence, we can start close to our expected parameter estimates. I recommend using starting guesses that are approximately equal to the parameters we estimated in problem 2(b) from problem set 3.
- To see the progress of the optimization algorithm and receive an update on the value of the objective function at every iteration, add the argument `control = list(trace = 1, REPORT = 1)` to your call of the `optim()` function.

Report the estimated parameters and standard errors from this model. Briefly interpret these results. For example, what does each parameter mean?

b. As in problem set 3, the Massachusetts Department of Conservation and Recreation (DCR) is considering an increase to the camping fee at Mount Greylock due to the cost of maintaining those camp sites, and they want to know how this change would affect park visitation patterns. Use the model and simulation draws in part (a) to simulate the elasticity of choosing each park with respect to the cost of camping at Mount Greylock (`park_id == 1`) for each camper; that is, 5 alternatives $\times$ 1000 campers = 5000 elasticities. Then, for each park, calculate the mean of its elasticity with respect to the cost of camping at Mount Greylock. The following steps can provide a rough guide to simulating elasticities:

I. Create a function to simulate elasticities for one camper:

  i. The function should take a set of parameters, the random draws for one camper, and the data for one camper as inputs: `function(parameters, draws, data)`.
  ii. Transform the standard normal draws into the correct distributions using the distribution parameters.
  iii. Calculate the representative utility for every alternative for each draw.
  iv. Calculate the conditional choice probability for every alternative for each draw.
  v. Calculate the simulated choice probability for every alternative as the mean over all draws.
  vi. Calculate the term inside the integral of the elasticity formula for every alternative for each draw by taking products of conditional choice probabilities and the cost coefficient.
  vii. Simulate the integral in the elasticity formula by taking the mean of the previous values over all draws for every alternative.
  viii. Calculate the elasticities by multiplying these simulated integrals by the cost of camping at Mount Greylock and dividing by the simulated choice probability of the respective alternative.

II. Simulate elasticities for each camper—that is, call this function for each camper—using your parameter estimates from part (a).

III. Average the simulated elasticity of each park over all campers.

Report these five mean elasticities. Briefly interpret these results.

Problem 2: Individual-Level Coefficients

In problem set 3, we calculated the distribution of the dollar value that a camper places on each hour spent traveling and the distribution of the dollar value that a camper places on camping in the mountains (relative to camping at the beach). DCR is also interested in understanding these valuations for the campers who visit each of the five state parks in our dataset.

a. Use the model in problem 1(a) to calculate the mean coefficients for campers at each of the five parks. Use the same simulation draws as in problem 1 to simulate the mean coefficients for each camper, and then average over the campers who visit each park. The following steps can provide a rough guide to simulating mean coefficients:

  I. Create a function to simulate mean coefficients for one camper:

    i. The function should take a set of parameters, the random draws for one camper, and the data for one camper as inputs: `function(parameters, draws, data)`.
    ii. Transform the standard normal draws into the correct distributions using the distribution parameters.
    iii. Calculate the representative utility for every alternative for each draw.
    iv. Calculate the conditional choice probability of the chosen alternative for each draw.
    v. Calculate weights for each simulation draw using the conditional choice probability of the chosen alternative.
    vi. Calculate the weighted average for each coefficient using these simulation draw weights.

  II. Simulate mean coefficients for each camper—that is, call this function for each camper—using your parameter estimates from problem 1(a).

III. For each of the five parks, average the simulated mean coefficients for all campers at that park.

If you are not able to calculate these mean coefficients "by hand," you can instead use the "canned" routine. First, estimate the model in problem 1(a) using the `mlogit()` function; this model is identical to problem 2(b) from problem set 3. Then use the `fitted()` function with argument `type = 'parameters'` to calculate the mean coefficients for each camper.

i. Report the mean coefficients for each park; that is, 3 variables $\times$ 5 alternatives $=$ 15 coefficients.

ii. Calculate the mean dollar value that a camper at each park places on each hour spent traveling and the mean dollar value that a camper at each park places on camping in the mountains (relative to camping at the beach); that is, 2 variables $\times$ 5 alternatives $=$ 10 dollar values. Briefly interpret these results.