Problem Set 4

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

Due: November 22, 11:59 pm 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. You can use any "canned" routine (e.g., `lm()`, `glm()`, and `mlogit()`) for this problem set.

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: Generalized Extreme Value Models

We are studying how campers at Massachusetts State Parks choose the park where they camp, which will assist the Massachusetts Department of Conservation and Recreation (DCR) in their planning. In particular, they want to understand how campers value their time to travel to the park and the setting—mountain or beach—of the park. Additionally, 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.

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

$$V_{nj} = \beta_1 C_{nj} + \beta_2 T_{nj} + \beta_3 M_j$$

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, and the β parameters are to be estimated. Importantly, do not include alternative-specific intercepts because β_3 would not be identified. (Reminder: the `mlogit()` function from the `mlogit` package estimates a multinomial logit model, but the data must first be converted to an indexed data frame using the `dfidx()` function from the `dfidx` package. See the Week 4 slides or the `mlogit` vignettes at `cran.r-project.org/web/packages/mlogit/index.html` for information on specifying a `formula` for the `mlogit()` function.)

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

ii. Calculate the dollar value that a camper places on each hour spent traveling and the dollar value that a camper places on camping in the mountains (relative to camping at the beach).

iii. Calculate 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. For each park, report the mean of its elasticity with respect to the cost of camping at Mount Greylock. Describe how these elasticities and substitution patterns relate to an important property of the logit model. (Reminder: the `fitted()` function with argument `type = 'probabilities'` calculates the choice probabilities of each alternative for each decision maker.)

b. The multinomial logit model of part (a) is not the best model for this setting if a camper's unobserved utility includes an individual preference for the mountains or the beach, which would create correlations among parks with the same setting. Model the camping park choice as a nested logit model with two nests, one for each park setting: mountains and beach. As in part (a), model the representative utility for park j as

$$V_{nj} = \beta_1 C_{nj} + \beta_2 T_{nj} + \beta_3 M_j$$

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, and the β parameters are to be estimated. (Reminder: the `mlogit()` function from the `mlogit` package estimates a nested logit model if you use the `nests` argument to specify nests as a named list.)

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

ii. The model in part (a) is effectively imposing a restriction on the model in part (b). Write the null hypothesis that is imposed by the model in part (a) and describe this hypothesis in words. Conduct a likelihood ratio test to test this null hypothesis. Do you reject this null hypothesis? What is the p-value of the test? Briefly interpret the result of this test. (Reminder: the `lrtest()` function performs a likelihood ratio test.)

iii. Calculate the dollar value that a camper places on each hour spent traveling and the dollar value that a camper places on camping in the mountains (relative to camping at the beach).

iv. Calculate 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. For each park, report the mean of its elasticity with respect to the cost of camping at Mount Greylock. Compare these elasticities to those you found in part (a) and describe any important differences.

Problem 2: Mixed Logit Model

The models in problem 1 have common parameters for all campers in the dataset. In reality, however, some or all of these parameters are likely to vary by camper for unobserved reasons. DCR is interested in understanding this heterogeneity and how it could affect park visitation patterns.

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 random coefficients on each variable. That is, the representative utility to camper n from park j is

$$V_{nj} = \beta_{1n}C_{nj} + \beta_{2n}T_{nj} + \beta_{3n}M_j$$

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 all three β coefficients as random with a normal distribution:

$$\beta_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$\beta_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$
$$\beta_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$$

Estimate this model using 100 draws for simulation (R = 100) and set a seed of 703 for replication (seed = 703). (Reminder: the `mlogit()` function from the `mlogit` package estimates a mixed logit model if you use the `rpar` argument to specify the random coefficients as a named vector.)

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

b. It is easier to calculate how campers value their time to travel to the park and the setting—mountain or beach—of the park when cost has a fixed (not-random) coefficient. Model the camping park choice as a mixed logit model with a fixed coefficient on cost. Express the representative utility of each alternative as a linear function of its cost, time, and setting—mountain or beach—with 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 this model using 100 draws for simulation (R = 100) and set a seed of 703 for replication (seed = 703).

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

  ii. We can test if β_1 is a fixed or random coefficient. Write the null hypothesis of your test and describe this hypothesis in words. Conduct a likelihood ratio test to test this null hypothesis. Do you reject this null hypothesis? What is the p-value of the test? Briefly interpret the result of this test.

 iii. Calculate the dollar value that a camper places on each hour spent traveling and the dollar value that a camper places on camping in the mountains (relative to camping at the beach). Because we have distributions for β_2 and β_3, these dollar values will also be distributions. Report the mean and standard deviation of each of these dollar value distributions. Briefly interpret these results.

iv. Calculate the proportion of campers who have a positive value of camping in the mountains (relative to camping at the beach).

c. DCR is considering an increase to the camping fee at Mount Greylock, which would increase the cost by $20 for each camper in our dataset. Use your parameter estimates from part (b) to simulate this counterfactual.

    i. How many fewer campers—of the 1000 campers in this dataset—do you expect will camp at Mount Greylock because of this fee increase? How many more campers do you expect will camp at each of the other four parks?

    ii. How do you expect this increased camping fee at Mount Greylock will affect the economic surplus of the 1000 campers in this dataset?