Problem Set 2

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

Due: November 4, 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()`).

Data

Download the file `commute_datasets.zip` from the course website. This zipped file contains two datasets—`commute_binary.csv` and `commute_multinomial.csv`—that you will use for this problem set. Both datasets contain simulated data on the travel mode choice of 1000 UMass graduate students who commute to campus from more than one mile away. The `commute_binary.csv` dataset corresponds to commuting in the middle of winter when only driving a car or taking a bus are feasible options. The `commute_multinomial.csv` dataset corresponds to commuting in the spring when riding a bike and walking are feasible alternatives. See the file `commute_descriptions.txt` for descriptions of the variables in each dataset.

Problem 1: Maximum Likelihood Estimation

We are again studying how UMass graduate students choose how to commute to campus in the spring when riding a bike and walking are feasible alternatives—as in problem 3 of problem set 1—but we are now estimating the model "by hand" to better understand the maximum likelihood estimation method. Use the `commute_multinomial.csv` dataset for this question.

a. Model the commute choice during spring as a multinomial logit model. Express the representative utility of each alternative as a linear function of its cost and time with common parameters on these variables. That is, the representative utility to student n from alternative j is

$$V_{nj} = \beta C_{nj} + \gamma T_{nj}$$

where C_{nj} is the cost to student n of alternative j, T_{nj} is the time for student n of alternative j, and the β and γ parameters are to be estimated. Estimate the parameters of this model by maximum likelihood estimation. The following steps can provide a rough guide to creating your own maximum likelihood estimator:

I. Create a function that takes a set of parameters and data as inputs: `function(parameters, data)`.

II. Within that function, make the following calculations:

    i. Calculate the representative utility of every alternative for each decision maker.

    ii. Calculate the choice probability of the chosen alternative for each decision maker.

    iii. Sum the log of these choice probabilities to get the log-likelihood.

    iv. Return the negative of the log-likelihood.

III. Maximize the 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_function, data = your_data,
      method = 'BFGS', hessian = TRUE)
```

Report your parameter estimates, standard errors, z-stats, and p-values. Briefly interpret these results. For example, what does each parameter mean?

b. Again model the commute choice during spring as a multinomial logit model, but now add alternative-specific intercepts for all but one alternative. That is, the representative utility to student n from alternative j is

$$V_{nj} = \alpha_j + \beta C_{nj} + \gamma T_{nj}$$

where C_{nj} is the cost to student n of alternative j, T_{nj} is the time for student n of alternative j, and the α, β, and γ parameters are to be estimated. Estimate the parameters of this model by maximum likelihood estimation. The steps to creating your own maximum likelihood estimator are the same as in part (a), but some of the calculations will be different.

    i. Report your parameter estimates, standard errors, z-stats, and p-values. Briefly interpret these results. For example, what does each parameter mean?

    ii. Conduct a likelihood ratio test on this model to test the joint significance of the alternative-specific intercepts. That is, test the null hypothesis:

$$H_0\!: \ \alpha_{bus} = \alpha_{car} = \alpha_{walk} = 0$$

Your null hypothesis may be slightly different, depending on what you consider your "reference alternative." Do you reject this null hypothesis? What is the p-value of the test? Briefly interpret the result of this test. (Reminder: to conduct this likelihood ratio test, you need the log-likelihood value of the model in part (b) and the log-likelihood value of the restricted model that is obtained when the hypothesized restrictions are imposed.)

c. Again model the commute choice during spring as a multinomial logit model, but now allow the parameter on time to be alternative-specific. That is, the representative utility to student n from alternative j is

$$V_{nj} = \alpha_j + \beta C_{nj} + \gamma_j T_{nj}$$

where C_{nj} is the cost to student n of alternative j, T_{nj} is the time for student n of alternative j, and the α, β, and γ parameters are to be estimated. Estimate the parameters of this model by maximum likelihood estimation. The steps to creating your own maximum likelihood estimator are the same as in part (a), but some of the calculations will be different.

i. Report your parameter estimates, standard errors, z-stats, and p-values. Briefly interpret these results. For example, what does each parameter mean?

ii. Conduct a likelihood ratio test on this model to test if the alternative-specific parameters on time are equal to one another. That is, test the null hypothesis:

$$H_0: \gamma_{bike} = \gamma_{bus} = \gamma_{car} = \gamma_{walk}$$

Do you reject this null hypothesis? What is the p-value of the test? Briefly interpret the result of this test.

Problem 2: Generalized Method of Moments

We are again studying how UMass graduate students choose how to commute to campus in the winter when riding a bike and walking are infeasible—as in problem 2 of problem set 1—but we are now estimating the model "by hand" to better understand the generalized method of moments estimation method. Use the `commute_binary.csv` dataset for this question.

a. Model the choice to drive to campus during winter as a binary logit model. Express the representative utility of each alternative as a linear function of its cost and time. Include an alternative-specific intercept and allow the parameter on time to be alternative-specific. That is, the representative utility to student n from driving and taking the bus, respectively, are

$$V_{nc} = \alpha + \beta C_{nc} + \gamma_{car} T_{nc}$$
$$V_{nb} = \gamma_{bus} T_{nb}$$

where C_{nj} is the cost to student n of alternative j, T_{nj} is the time for student n of alternative j, and the α, β, and γ parameters are to be estimated. We exclude a bus-specific intercept term because only one intercept term is identified in this model, and we exclude the bus cost because it is free for all students. It may be easier to think about the difference in representative utility between driving and taking the bus for student n:

$$V_{nc} - V_{nb} = \alpha + \beta C_{nc} + \gamma_{car} T_{nc} - \gamma_{bus} T_{nb}$$

Because this is a binary logit model, we can express the choice probability of driving as a function of $V_{nc} - V_{nb}$:

$$P_{nc} = \frac{1}{1 + e^{-(V_{nc} - V_{nb})}}$$

Estimate the parameters of this model by method of moments. The following steps can provide a rough guide to creating your own method of moments estimator:

I. Write down moment conditions for this model. You should have four moment conditions for this model.

II. Create a function that takes a set of parameters and a *matrix* of data as inputs: `function(parameters, data_matrix)`.

III. Within that function, make the following calculations:

i. Calculate the difference in representative utility for each decision maker.

ii. Calculate the choice probability of driving for each decision maker.

iii. Calculate the econometric residual, or the difference between the outcome and the probability, for each decision maker.

iv. Calculate each of the L moments for each decision maker.

v. Return the $N \times L$ matrix of individual moments.

IV. Find the MM estimator using `gmm()`. Your call of the `gmm()` function may look something like:

```
gmm(g = your_function, x = your_data_matrix, t0 = your_starting_guesses,
    vcov = 'iid', method = 'Nelder-Mead'
    control = list(reltol = 1e-25, maxit = 10000))
```

Report your parameter estimates, standard errors, t-stats, and p-values. Briefly interpret these results. For example, what does each parameter mean?

b. You might be concerned that the cost and time data are exogenous; for example, a student who enjoys driving is more likely to live farther from campus because they do not mind the extra cost and time spent driving, and a student who enjoys taking the bus is more likely to live close to a bus stop so the bus commute time is less. The `commute_binary.csv` dataset includes four possible instruments that could be correlated with the cost or time of commuting: `price_gas`, `snowfall`, `construction`, and `bus_detour`. Again model the choice to drive to campus during winter as in (a). That is, the representative utility to student n from driving and taking the bus, respectively, are

$$V_{nc} = \alpha + \beta C_{nc} + \gamma_{car} T_{nc}$$
$$V_{nb} = \gamma_{bus} T_{nb}$$

where C_{nj} is the cost to student n of alternative j, T_{nj} is the time for student n of alternative j, and the α, β, and γ parameters are to be estimated. Estimate the parameters of this model by generalized method of moments, constructing moment conditions using the instruments described above. Thus, you will have five instruments: constant term, `price_gas`, `snowfall`, `construction`, and `bus_detour`. The steps to creating your own generalized method of moments estimator are the same as in part (a), but now you have four parameters and five moment conditions. (Note: this model may have challenges converging, but you can help it along by giving it different starting values. I got it to converge by using `t0 = c(0, 0, -0.3, -0.1)`.)

i. Report your parameter estimates, standard errors, t-stats, and p-values. Briefly interpret these results. For example, what does each parameter mean?

ii. Test if your model is correctly specified by performing an overidentifying restrictions test. Report the results of this test and briefly interpret these results. (Reminder: the `specTest()` function from the `gmm` package conducts this specification test.)