Final Project

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

Due: December 10, 11:59 pm ET

Rules

Email two files—the .R file of your code and a .pdf file that combines your 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 set of code and writeup for the group, and I strongly encourage you to do so.

You have two options for answering problems 2 and 3 of this final project: writing your own estimation code or using canned routines (i.e., `mlogit()` and other functionality from the `mlogit` package, or a comparable package). If you use canned routines to answer these problems, I will deduct 10% from your score for this project. In other words, correctly writing your own estimation code will result in a score of 100%, whereas correctly using canned routines will result in a score of 90%.

Data

Download the file `final_dataset.zip` from the course website (`github.com/woerman/ResEcon703`). This zipped file contains the dataset, `nox.csv`, that you will use for the final project. See the file `data_descriptions.txt` for descriptions of the variables in the dataset.

Paper

For this final project, you will replicate the main estimation of the following paper:

> Meredith Fowlie. 2010. "Emissions Trading, Electricity Restructuring, and Investment in Pollution Abatement." *American Economic Review* 100 (3): 837–869

This paper is available for download from the AER website (`doi.org/10.1257/aer.100.3.837`). Read the paper before you begin replicating the results.

Problem 1: Summary Statistics

Before replicating the estimation of the paper, first make sure you can recreate the summary statistics presented in Figure 2 and Tables 1–3. You should be able to replicate most of these summary statistics within approximately 1% of what is reported in the paper, except as indicated in part (d).

a. The main point of this paper is to show that different regulatory regimes lead plant managers to make different choices when complying with the NOx Budget Program. Figure 2 nicely summarizes this finding by depicting, for each regulatory regime, the percentage of installed capacity with each category of compliance choices. Calculate these percentages shown in Figure 2. Do not plot your results; report the percentages in a table or well-organized R output.

b. One potential concern with this analysis is that units in different regulatory regimes have different characteristics, and these different unit characteristics drive the compliance choices. Table 1 shows this is not the case because characteristics are broadly similar across regulatory regimes. Calculate these unit summary statistics reported in Table 1 and report them in a table or well-organized R output.

c. A related concern is that the costs of compliance choices may differ for units in different regulatory regimes, which would clearly drive differences in compliance choices. Table 2 shows this is not the case because compliance costs are broadly similar across regulatory regimes. Calculate these compliance cost summary statistics reported in Table 2 and report them in a table or well-organized R output. Note that each row of this table reports costs for a specific compliance alternative (not for a category of compliance alternatives, as in Figure 2); in descending order, the rows correspond to compliance alternatives `2` (`CM`), `7` (`LNB`), `15` (`SN`), `14` (`SC`), and `10` (`N`), respectively.

d. One final concern is that all of the compliance alternatives are not feasible at every unit because of unit characteristics, and differences in choice sets could drive differences in compliance choices. Table 3 shows this is not the case because compliance choice sets are broadly similar across regulatory regimes. Calculate these choice set summary statistics reported in Table 3 and report them in a table or well-organized R output. Note that, in the bottom panel of this table, the first two rows correspond to categories of compliance alternatives—categories `cm` and `lnb`—and the final two rows correspond to individual compliance alternatives—alternatives `15` (`SN`) and `14` (`SC`). You may not be able to replicate these summary statistics as closely as the previous summary statistics, but you should be able to get within 10% and find similar patterns to what is reported in the paper.

Problem 2: Model Estimation

Table 4 reports the main estimation results of this paper. Six models are estimated, three logit models (or conditional logit models) and three mixed logit models (or random coefficient logit models); these models are described in Sections IV and V of the paper. Estimate each of these six models. Report the parameter estimates, standard errors, and log-likelihood value for each model in a table or well-organized R output. You should be able to closely replicate the parameter estimates and log-likelihood values reported in Table 4, but your standard errors may be smaller than those reported in the paper. See these additional comments to assist your estimation:

Canned routines Use the `mlogit()` function from the `mlogit` package to estimate these models. A few notes on using the `mlogit` package for these models:

- We observe choices for each unit, but the paper assumes choices are being made by the plant manager. So we observe multiple choices (units) for each individual (plant manager), which is effectively panel data. To implement panel data in the `mlogit` package, the `idx` argument in the `dfidx()` function should look something like `idx = list(c('unit_id', 'facility_id'), 'alt_id')`, and you should use the `panel` argument in the `mlogit()` function.

- Models 1 and 4 include scale parameters for deregulated and public units account for differences in the error term variances for the different regimes. To estimate scale parameters, specify a fourth "bin" in your `formula` within the `mlogit()` function. The variable in this fourth "bin" must be a factor variable; you can create a new factor variable using the `fct_infreq()` function or one of the many other factor variable functions in R. For examples of estimating scale parameters, see the `mlogit` vignettes (`cran.r-project.org/web/packages/mlogit/index.html`).
- When estimating the mixed logit models, set a seed so your results can be replicated and use at least 1000 random draws to most accurately simulate choice probabilities.
- These models will not always converge using the default optimization method. Instead, use the BHHH method by specifying `method = 'bhhh'` in the `mlogit()` function.

Hand-coded estimation Write your own functions to calculate (simulated) log-likelihood, and then use the `optim()` function to find the set of parameters that maximizes the (simulated) log-likelihood for each model. A few notes on writing your own estimation code for these models:

- You should include in your estimation only the compliance alternatives that are available to each unit.
- We observe choices for each unit, but the paper assumes choices are being made by the plant manager. So we observe multiple choices (units) for each individual (plant manager), which is effectively panel data. These panel choice probabilities are given by Equations (3) and (4) in the paper.
- Models 1 and 4 include scale parameters for deregulated and public units, as defined in Equation (6) of the paper. These parameters scale cost for all units within that regulatory regime to account for differences in the error term variances for the different regimes.
- When estimating the mixed logit models, set a seed so your results can be replicated and use at least 1000 random draws to most accurately simulate choice probabilities.
- You may have difficulty getting some of these models to converge to a global maximum. If you encounter this problem, try different starting values and convergence methods in the `optim()` function. I have found that these models tend to converge when using the BFGS method.

Problem 3: Manager-Specific Coefficient Distributions

Table 5 summarizes plant manager-specific coefficient distributions for the random coefficients in models 4–6. For each model, means and standard deviations for both the population (or unconditional) distributions and the conditional distributions are reported, and these statistics are reported separately by regulatory regime. Calculate these plant manager-specific coefficient distribution summary statistics and report them in a table or well-organized R output. You may not be able to perfectly replicate all of the values in this table, but they should generally be close; you should also find the same patterns: means are roughly equal for comparable unconditional and conditional distributions, but standard deviations are smaller for the conditional distributions. See these additional comments to assist your calculations:

Canned routines Population (or unconditional) distribution summary statistics can be calculated using only the model results. To calculate the conditional distribution summary statistics, first find plant manager-specific mean conditional coefficients using the `fitted()` function with argument `type = 'parameters'`.

Hand-coded estimation To calculate these distribution summary statistics, use the same random draws that you used when estimating the corresponding model.