Structural Estimation Exercise

Model

Consider the following location choice model. Households i choose between locations $j = 1, \dots, J$. For each household i let $\mathbf{z}_i$ be an observed $k_z \times 1$ vector of household characteristics. For each location j, let $\mathbf{x}_j$ be an observed $k_x \times 1$ vector of location characteristics, including the price to live in the location. The utility that household i gets from living in location j is:

$$\begin{aligned} V_{ij} &= \mathbf{x}_j'\mathbf{A} + \mathbf{z}_i'\mathbf{B}\mathbf{x}_j + \xi_j + \epsilon_{ij} \\ &= u_{ij} + \epsilon_{ij} \end{aligned}$$

where $\mathbf{A}$ is a $k_x \times 1$ vector of parameters and $\mathbf{B}$ is a $k_z \times k_x$ matrix of parameters. $\mathbf{A}$ measures the vertically-differentiated utility of different location characteristics, and $\mathbf{B}$ measures how the utility over different location characteristics change with household characteristics. ϵ_{ij} is an iid preference shock that we assume is distributed Type 1 Extreme Value. ξ_j is an unobserved quality component of location j. We do not impose any assumptions on ξ_j, except that we assume it is uncorrelated with any element of $\mathbf{x}_j$ other than price.

The probability that individual i chooses location j is:

$$P_{ij} = \frac{e^{u_{ij}}}{\sum_{k=1}^{J} e^{u_{ik}}}$$

We assume that the share of households that can live in any one location j is given by s_j, due to inelastic housing stock. In order for housing markets to clear, the equilibrium condition is:

$$\frac{1}{N}\sum_{i=1}^{N} P_{ij} = s_j \text{ for } j = 1, \dots, J$$

Data

Suppose for $i = 1, \dots, N$ we have data on $\mathbf{z}_i$ and j_i^*, the chosen location for household i. For $j = 1, \dots, J$ we have data on $\mathbf{x}_j$ and s_j. Moreover, for each location j we have an instrumental variable, IV_j which is correlated with price but independent of ξ_j. The log-likelihood of the data is:

$$LL = \sum_{i=1}^{N} \log\left(P_{ij_i^*}\right)$$

Our goal is to use this data to estimate $\mathbf{A}$ and $\mathbf{B}$.

Exercise

We will estimate this model using a simulated dataset. Download `household_data.csv` and `neighborhood_data.csv` from the course website. `household_data.csv` contains data on $(\mathbf{z}_i, j_i^*)$ and `neighborhood_data.csv` contains data on $(\mathbf{x}_j, s_j, IV_j)$. In $\mathbf{z}_i$, we have the household's standardized log-income and an indicator for whether the household is college-educated. In $\mathbf{x}_j$, we have the location's standardized school quality and its log-price.

1. One of the difficulties of estimating all of the parameters directly by maximum-likelihood is the presence of J nuisance parameters in ξ_j. Describe a two-step procedure in which $\mathbf{B}$ can be estimated in the first step, without having to estimate $\mathbf{A}$ or ξ_j, and $\mathbf{A}$ is estimated in the second step. (Hint: Write $u_{ij} = \mathbf{z}_i'\mathbf{B}\mathbf{x}_j + \delta_j$ and use the equilibrium condition to derive a contraction-mapping for δ_j.)

2. Download `estimate.m` from the course website. `estimate.m` is a Matlab script for estimating $\mathbf{B}$ and $\mathbf{A}$. However, it requires you to supply the objective function, `loglike.m`. Write your own version of `loglike.m` and use it to run `estimate.m`. Compare your results to `desired_output.txt`.