

Lecture Note 8: Simulated Method of Moments

(Reading for next time: Hajivassiliou, V., and McFadden, D., (1998), “The Method of Simulated Scores for Estimation of LDV Models,” *Econometrica* 66(4), 863-896; also start reading Chib and Greenberg.)

Multinomial Probit Model of Discrete Choice (McFadden, 1989)

Choices: $c = 1, \dots, C$

Measured characteristics of choices: x_1, \dots, x_C , where each x_c is a $k \times 1$ vector.

Examples of characteristics: prices, distance of choice to consumer.

Preferences are based on “weights”: $\alpha \sim N_k(\beta, \Omega)$.

Utilities:

$$\begin{aligned}u_1 &= x_1' \alpha \\u_2 &= x_2' \alpha \\&\vdots \\u_C &= x_C' \alpha\end{aligned}$$

Choice rule: pick c if $u_c \geq u_{c'} \quad \forall c' \in \{1, \dots, C\}$.

(Assume no ties, and only one object chosen.)

Let

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_C \end{pmatrix}, \quad X = \begin{pmatrix} x_1' \\ \vdots \\ x_C' \end{pmatrix}.$$

Then $u = X\alpha$, so

$$u|X, \beta, \Omega \sim N_C(X\beta, X\Omega X').$$

Let $d_c = 1$ if choice c is chosen. Then

$$\begin{aligned}E[d_c|X, \beta, \Omega] &= Pr[d_c = 1|X, \beta, \Omega] \\&= Pr[u_c \geq u_1, \dots, u_c \geq u_C|X, \beta, \Omega] \\&= \int \dots \int 1(u_c \geq u_1, \dots, u_c \geq u_C) dN(u_1, \dots, u_C|X\beta, X\Omega X'),\end{aligned}$$

where $dN(\cdot|\cdot, \cdot)$ means integration with respect to the density of the multivariate normal distribution.

Sample Data

Individuals $i = 1, \dots, n$.

Weights: $\alpha_i \stackrel{\text{iid}}{\sim} N(\beta, \Omega)$.

Now assume that θ denotes the distinct parameters of β, Ω . For example, we might restrict Ω to be diagonal. So we can write $\beta = \beta(\theta)$ and $\Omega = \Omega(\theta)$, and

$$\alpha_i \stackrel{\text{iid}}{\sim} N(\beta(\theta), \Omega(\theta)).$$

Measured characteristics: x_{i1}, \dots, x_{iC} . (For example, individuals might face different prices for the different choices.)

$$X_i = \begin{pmatrix} x'_{i1} \\ \vdots \\ x'_{iC} \end{pmatrix}.$$

Choice indicators:

$$d_i = \begin{pmatrix} d_{i1} \\ \vdots \\ d_{iC} \end{pmatrix}.$$

Likelihood for individual i :

$$\prod_{c=1}^C Pr(d_{ic} = 1 | X_i, \theta)^{d_{ic}}.$$

Full-sample likelihood:

$$\prod_{i=1}^n \prod_{c=1}^C Pr(d_{ic} = 1 | X_i, \theta)^{d_{ic}}.$$

Log likelihood:

$$L(\theta) = \sum_{i=1}^n \sum_{c=1}^C d_{ic} \log Pr(d_{ic} = 1 | X_i, \theta).$$

ML Estimator:

$$\hat{\theta}_{ML} = \arg \max_{\theta} L(\theta).$$

First order conditions:

$$0 = \frac{\partial L(\theta)}{\partial \theta} = \sum_{i=1}^n \sum_{c=1}^C d_{ic} \frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1 | X_i, \theta).$$

Note that

$$\sum_{c=1}^C Pr(d_{ic} = 1 | X_i, \theta) = 1.$$

So

$$\sum_{c=1}^C \frac{\partial}{\partial \theta} Pr(d_{ic} = 1 | X_i, \theta) = 0.$$

Also,

$$\left[\frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1 | X_i, \theta) \right] \cdot Pr(d_{ic} = 1 | X_i, \theta) = \frac{\partial}{\partial \theta} Pr(d_{ic} = 1 | X_i, \theta),$$

So

$$0 = \sum_{c=1}^C \frac{\partial}{\partial \theta} Pr(d_{ic} = 1 | X_i, \theta) \cdot Pr(d_{ic} = 1 | X_i, \theta),$$

and we can write the FOC alternatively as

$$0 = \sum_{i=1}^n \sum_{c=1}^C \left[\frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1 | X_i, \theta) \right] \{d_{ic} - Pr(d_{ic} = 1 | X_i, \theta)\}.$$

This is somewhat intuitive: the second term $\{d_{ic} - Pr(d_{ic} = 1 | X_i, \theta)\}$ is the outcome minus its conditional mean given X_i .

To solve for the MLE, we need to be able to calculate $Pr(d_{ic} = 1 | X_i, \theta)$ (and possibly its derivatives) for each i, c and different possible parameter values θ .

However, the integral defining $Pr(d_{ic} = 1 | X_i, \theta)$ does not have a simple closed-form expression.

For relatively few choices ($C \leq 4$), there exist deterministic (nonrandom) numerical integration routines that are efficient, but these cannot be used when the number of choices is large.

Lerman and Manski (1981) suggest a Monte Carlo procedure:

For $s = 1, \dots, S$:

- Draw $u^s = (u_1^s, \dots, u_C^s)' \sim N(X_i \beta(\theta), X_i \Omega(\theta) X_i')$.
- Form $d^s = (d_1^s, \dots, d_C^s)'$.

Then approximate $Pr(d_{ic} = 1 | X_i, \theta)$ by

$$\tilde{P}_S(d_{ic} = 1 | X_i, \theta) = \frac{1}{S} \sum_{s=1}^S d_c^s.$$

By the LLN,

$$\tilde{P}_S(d_{ic} = 1 | X_i, \theta) \xrightarrow{p} Pr(d_{ic} = 1 | X_i, \theta) \quad \text{as } S \rightarrow \infty.$$

Simulation-based estimator:

$$\hat{\theta}_{LM} = \arg \max_{\theta} \sum_{i=1}^n \sum_{c=1}^C d_{ic} \cdot \log \tilde{P}_S(d_{ic} = 1 | X_i, \theta).$$

For this to work well, S needs to be very large so that the simulation error $\tilde{P}_S(d_{ic} = 1|X_i, \theta) - Pr(d_{ic} = 1|X_i, \theta)$ does not affect the inference. Unfortunately, this makes the procedure too computationally burdensome.

Method of Moments and Simulation

For the moment, suppose that $Pr(d_{ic} = 1|X_i, \theta)$ is not hard to calculate, and consider an alternative estimation approach.

Recall that $Pr(d_{ic} = 1|X_i, \theta)$ also had the interpretation as the conditional mean of d_{ic} given X_i . So

$$E[d_{ic} - Pr(d_{ic} = 1|X_i, \theta)|X_i] = 0.$$

This suggests moment conditions

$$E \begin{bmatrix} w_{i1}\{d_{i1} - Pr(d_{i1} = 1|X_i, \theta)\} \\ \vdots \\ w_{iC}\{d_{iC} - Pr(d_{iC} = 1|X_i, \theta)\} \end{bmatrix} = 0,$$

where the w_{ic} are vector-valued functions of X_i . (They could also be additional instrumental variables that are mean-independent of $d_{ic} - Pr(d_{ic} = 1|X_i, \theta)$.)

So we have a GMM moment function

$$g(w_i, X_i, d_i, \theta) = \begin{pmatrix} w_{i1}\{d_{i1} - Pr(d_{i1} = 1|X_i, \theta)\} \\ \vdots \\ w_{iC}\{d_{iC} - Pr(d_{iC} = 1|X_i, \theta)\} \end{pmatrix}.$$

Notice the similarity between this choice of moment function and the MLE score equation. Basically, the MLE score involves replacing w_{ic} by $\frac{\partial}{\partial \theta} \log Pr(d_{ic} = 1|X_i, \theta)$, which also depends on θ . This suggests that by choosing w_{ic} cleverly, we could achieve the same efficiency as MLE.

Assume that the instruments w are such that the dimension of g is greater than or equal to the dimension of θ . Then we could estimate θ by solving

$$\min_{\theta} \left[\sum_{i=1}^n g(w_i, X_i, d_i, \theta) \right]' \left[\sum_{i=1}^n g(w_i, X_i, d_i, \theta) \right].$$

Here, we are implicitly using an identity matrix as the GMM weighting matrix. Alternatively, we could have a different weighting matrix A , but then we could modify the instruments w to account for this weighting.

By stacking the observations, we can rewrite this estimator as

$$\hat{\theta}_{MM} = \arg \min_{\theta} [d - P(\theta)]' W' W [d - P(\theta)],$$

where

$$d = \begin{pmatrix} d_{11} \\ \vdots \\ d_{1C} \\ \vdots \\ d_{n1} \\ \vdots \\ d_{nC} \end{pmatrix}, \quad P(\theta) = \begin{pmatrix} Pr(d_{11} = 1|X_1, \theta) \\ \vdots \\ Pr(d_{1C} = 1|X_1, \theta) \\ \vdots \\ Pr(d_{n1} = 1|X_n, \theta) \\ \vdots \\ Pr(d_{nC} = 1|X_n, \theta) \end{pmatrix}.$$

Now, consider simulation of the choice probabilities. For a finite, possibly small value of S , the simulation error will not be negligible. However, the simulations are unbiased in the sense that

$$E[\tilde{P}_S(d_{ic} = 1|X_i, \theta) - Pr(d_{ic} = 1|X_i, \theta)|X_i] = 0.$$

Also $\tilde{P}_S(d_{ic} = 1|X_i, \theta)$ is independent of the actual outcomes d_{ic} conditional on X_i . So it is true that

$$E[d_{ic} - \tilde{P}_S(d_{ic} = 1|X_i, \theta)|X_i] = 0.$$

The variance of $d_{ic} - \tilde{P}_S(d_{ic} = 1|X_i, \theta)$ will differ from the variance of $d_{ic} - Pr(d_{ic} = 1|X_i, \theta)$, due to the simulation error, but the conditional means are both 0.

So, letting $\tilde{P}(\theta)$ denote the stacked set of simulated response probabilities, we could use the simulated method of moments estimator

$$\hat{\theta}_{SMM} = \arg \min_{\theta} [d - \tilde{P}(\theta)]' W' W [d - \tilde{P}(\theta)],$$

This estimator does not require $S \rightarrow \infty$; instead the simulation error is averaged out across individual observations. It will have a different asymptotic variance than the MM estimator, since the moment equation has a different variance.

Implementation Notes

1. Notice that we need to construct the simulated probability vector $\tilde{P}(\theta)$ for all $\theta \in \Theta$ (or at least enough values so that we can maximize the criterion function). Exactly how we do this turns out to matter.

In particular, suppose we generate new random draws for each different value of θ . Then it turns out that the SMM estimator is no longer consistent and asymptotically normal.

Instead, we have to use the same random draws to construct $\tilde{P}(\theta)$ for different θ . Here is one way to do this:

For $s = 1, \dots, S$, we want to draw $u_i^s(\theta) \sim N(X_i, \beta(\theta), X_i \Omega(\theta) X_i')$. This can be done by drawing

$$\eta_{i1}^s, \dots, \eta_{iC}^s \stackrel{\text{iid}}{\sim} N(0, 1),$$

and forming

$$u_i^s(\theta) = X_i \beta(\theta) + \Gamma_i(\theta)' (\eta_{i1}^s, \dots, \eta_{iC}^s)',$$

where $\Gamma_i(\theta)$ is the Cholesky factor of $X_i \Omega(\theta) X_i'$, an upper triangular matrix such that $\Gamma_i(\theta)' \Gamma_i(\theta) = X_i \Omega(\theta) X_i'$.

Then we form $d_{ic}^s(\theta)$ by seeing which choice gives highest utility among the elements of u_i^s , and construct

$$\tilde{P}_S(d_{ic} = 1 | X_i, \theta) = \frac{1}{S} \sum_{s=1}^S d_{ic}^s(\theta).$$

The key here is that $\eta_{i1}^s, \dots, \eta_{iC}^s$ do not change with θ . We draw them once, and only change $\beta(\theta)$ and $\Gamma_i(\theta)$ when calculating $u_i^s(\theta)$.

2. Another issue is that $\tilde{P}(\theta)$ is not continuous in θ . Take a typical element

$$\tilde{P}_S(d_{ic} = 1 | X_i, \theta) = \frac{1}{S} \sum_{s=1}^S d_{ic}^s(\theta).$$

For fixed S , this will be a number in the set $\{0, 1/S, 2/S, \dots, S/S\}$. As θ is varied, $\tilde{P}_S(d_{ic} = 1 | X_i, \theta)$ will jump between fractions of S .

As a consequence, $\tilde{P}(\theta)$ is discontinuous in θ , and so is the criterion function $[d - \tilde{P}(\theta)]' W' W [d - \tilde{P}(\theta)]$.

This has the practical problem that simple gradient-based numerical optimization routines will not work for solving the GMM minimization problem. We need to use some more sophisticated optimization routines, such as simulated annealing or simplex methods. (Amoeba is a simplex method.)

The discontinuity of the criterion function also means that our standard proofs of consistency and asymptotic normality of method of moments estimators do not work. More powerful asymptotic techniques are needed – see McFadden (1989) and Pakes and Pollard (1989).

3. Since the discontinuity of $\tilde{P}(\theta)$ can make solving the estimator problematic, one alternative is to modify the simulated probabilities to make them smooth in θ . McFadden discusses some different ways to do this. Some versions lead to simulated probabilities that are slightly biased. In this case, we need the bias to be small asymptotically.