MLE, GMM, Inference, Bootstrap, Inequality Restrictions, Method of Simulated Moments, Discrete Choice Models

Whitney K. Newey

DSE Summer School, August 2022

INTRODUCTION

Here we briefly describe MLE, GMM, give a few examples, and compare them. We given inference methods including a bootstrap that gives specification robust standard errors and a bootstrap that can be used to test and impose inequality restrictions. We describe the method of simulated moments and consider an extended set of examples from discrete choice. Some references are

–MLE and GMM:

Newey, W.K. and D. McFadden (1994): Large sample estimation and hypothesis testing, Handbook of Econometrics 4, 2111-2245.

–Bootstrapping for inference with inequality restrictions:

Chernozhukov, V., W.K. Newey, and A. Santos (2022): "Constrained Conditional Moment Restriction Models," Econometrica, forthcoming.

–Method of Simulated Moments:

Pakes, A. (1986): "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica* 54, 755-784.

McFadden, D. (1989). "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica* 995-1026.

Pakes, A., & Pollard, D. (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica*, 1027-1057.

–Discrete Choice Models:

Abadie, A. (2022): "MIT Lecture Notes on Discrete Choice".

McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in Frontiers of Econometrics, ed. by P. Zarembka. Academic Press.

Hausman, J.A. and D. Wise (1978): "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica* 46, 403-426.

Berry, S., J. Levinsohn, and A. Pakes (1995), "Automobile Prices in Market Equilibrium," *Econometrica* 63, 841-890.

McFadden, D., & Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics* 15, 447-470.

DATA AND PARAMETERS

We focus here on data $(W_1, ..., W_n)$ consisting of n i.i.d. observations on a random vector W.

We allow for clustering, with each W_i being one cluster of data. For example in panel data W_i would be data on one individual over time.

The independence of observations across i will be imposed throughout.

The effect(s) of interest will be the true value θ_0 of some $p \times 1$ vector of parameters θ.

We will describe and compare maximum likelihood (MLE), generalized method of moments (GMM), the method of simulated moments (MSM), give examples, and explain how to do inference using the bootstrap, with inequality restrictions.

MLE

An estimator is some function of the data, denoted by $\hat{\theta} = \theta(W_1, ..., W_n)$ where we usually just write $\hat{\theta}$.

The MLE is based on the assumption that we know the form of the pdf $f(w|\theta)$ of a data observation; includes discrete data where $f(w|\theta)$ is a probability.

$\hat{\theta}_{MLE}$ maximizes the probablility density function (pdf) of all the data $\Pi_{i=1}^{n} f(W_i|\theta)$, at the data over some set Θ of possible values of θ; $\hat{\theta}_{MLE}$ is defined the same way when W_i is discrete and $f(w|\theta)$ is a model for $\Pr(W_i = w)$.

Since a strict monotonic transformation has the same maximizing value, taking natural log and dividing by sample size n gives (for "argmax" denoting the maximizing value)

$$\begin{aligned} \hat{\theta}_{MLE} &= \arg\max_{\theta \in \Theta} [\Pi_{i=1}^{n} f(W_i|\theta)] = \arg\max_{\theta \in \Theta} \frac{1}{n} \ln[\Pi_{i=1}^{n} f(W_i|\theta)] \\ &= \arg\max_{\theta \in \Theta} [\frac{1}{n} \sum_{i=1}^{n} \ln\{f(W_i|\theta)\}]. \end{aligned}$$

Fundamental condition for MLE to be a consistent estimator (i.e. $\hat{\theta} \xrightarrow{p} \theta_0$ whatever the true parameter value θ_0 is):

Assumption MLE: i) W_i *has probability density function* $f(w|\theta_0)$ *for some* $\theta_0 \in \Theta$; *ii)* $f(w|\theta) \neq f(w|\theta_0)$ *if* $\theta \neq \theta_0$;

Here i) is a correct specification condition and ii) is an identification condition, where identification refers to the ability to distinguish the true value θ_0 of from other θ. The properties of the MLE are summarized in the following proposition. Let J denote the information matrix,

$$J = -E[\frac{\partial^2}{\partial\theta\partial\theta'} \ln f(W_i|\theta)\Big|_{\theta=\theta_0}], \ \hat{J} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \ln f(W_i|\theta)\Big|_{\theta=\hat{\theta}_{MLE}}.$$

Define

$$V := J^{-1}, \hat{V} := \hat{J}^{-1}.$$

Proposition MLE: *If Assumption MLE and other regularity conditions are satisfied (see Newey and McFadden, 1994) then*

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, V), \ \hat{V} \xrightarrow{p} V.$$

Also $\hat{\theta}_{MLE}$ *is asymptotically efficient.*

Proposition MLE: *If Assumption MLE and other regularity conditions are satisfied (see Newey and McFadden, 1994) then*

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, V),\ \hat{V} \xrightarrow{p} V.$$

Also $\hat{\theta}_{MLE}$ *is asymptotically efficient.*

Sometimes we summarize first two conclusions as "Under correct specification, identification, and other regularity conditions MLE is consistent and asymptotically normal (CAN) with asymptotic variance V and $\hat{J}^{-1}$ is a consistent estimator of the asymptotic variance."

Efficiency means that V is smaller (in the positive semi-definite (p.s.d.) sense) than the asymptotic variance of any other estimator in certain types of estimators; it is beyond scope of this class to say more about what types of estimators.

Often $f(w|\theta)$ is actually a conditional pdf for some vector Y of outcome variables given a vector X of regressors, i.e. $w = (y, x)$ and

$$f(w|\theta) = f(y|x, \theta),$$

where $f(y|x, \theta)$ is a conditional pdf. Same properities hold for conditional MLE (CMLE) $\hat{\theta}$ and $\hat{V}$ obtained as previously with $f(w|\theta)$ replaced by $f(y|x, \theta)$.

MLE Example (Binary Choice):

$$\begin{aligned} W &= (Y, X),\ Y = 1(Y^* \geq 0), \\ Y^* &= X'\theta_0 - \varepsilon,\ \varepsilon \text{ is independent of } X \text{ with cdf } H(u) = \mathsf{Pr}(\varepsilon \leq u). \end{aligned}$$

Y^* can be interpreted at the difference of the utilities of choices $Y = 1$ and $Y = 0$. For instance Y might be mode of transportation, take bus from lodging to Tang building or walk. Here X are observed characteristics of utility, like travel time, price of the subway, and income. The ε represents unobserved individual heterogeneity in utilities.

Then

$$\mathsf{Pr}(Y = 1|X) = \mathsf{Pr}(Y^* \geq 0|X) = \mathsf{Pr}(X'\theta_0 \geq \varepsilon) = H(X'\theta_0).$$

Probit has $H(u) = \Phi(u)$ where $\Phi(u)$ is the CDF of $N(0, 1)$. Logit has $H(u) = e^u/(1 + e^u)$.

Here Y has conditional pdf

$$f(y|x,\theta_0) = H(x'\theta_0)^y[1 - H(x'\theta_0)]^{1-y}.$$

The (conditional) MLE is

$$\begin{aligned}\hat{\beta} &= \arg\max_{\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}\ln f(Y_i|X_i,\theta)\right\} \\ &= \arg\max_{\theta}\left\{\frac{1}{n}\sum_{i=1}^{n}(Y_i \ln H(X_i'\theta) + (1-Y_i)\ln[1 - H(X_i'\theta)])\right\}\end{aligned}$$

This estimator is CAN when $\Pr(Y = 1|X) = H(X'\theta_0)$ for some θ_0 and an identification condition is satisfied. It is asymptotically efficient if the distribution of X is unrestricted with unknown form.

GMM

GMM (generalized method of moments) is based on a $r \times 1$ vector of functions $g(W, \theta)$ that are specified to have mean zero at the true parameter, i.e. such that

$$E[g(W, \theta_0)] = 0.$$

GMM is constructed by choosing $\hat{\theta}$ so that sample averages are close to zero, i.e. close to expectation of moment functions at true value.

Define

$$\hat{g}(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(W_i, \theta).$$

Let $\hat{\mathbf{\Psi}}$ be an $r \times r$ p.s.d. matrix that can depend on the data. A GMM estimator is obtained as

$$\hat{\theta}_{GMM} = \arg \min_{\theta \in \Theta} \hat{g}(\theta)' \hat{\mathbf{\Psi}} \hat{g}(\theta).$$

That is, $\hat{\theta}_{GMM}$ sets $\hat{g}(\theta)$ to be as close to zero as possible where the p.s.d. matrix $\hat{\mathbf{\Psi}}$ is used to quantify "closeness to zero".

Fundamental condition for consistency and asymptotic normality of GMM is.

Assumption GMM: *i) The* $r \times 1$ *vector of functions* $g(w, \theta)$ *satisfies* $E[g(W, \theta_0)] = 0$; *ii)* $E[g(W, \theta)] \neq 0$ *if* $\theta \neq \theta_0$;

The properties of GMM are summarized in the following proposition. Let $g_\theta(w, \theta) = \partial g(w, \theta)/\partial \theta$ and

$$\begin{aligned} G &= E[g_\theta(W, \theta_0)],\ \hat{G} = \frac{1}{n}\sum_{i=1}^{n} g_\theta(W_i, \hat{\theta}_{GMM}),\ \hat{\Psi} \xrightarrow{p} \Psi, \\ \Omega &= E[g(W, \theta_0)g(W, \theta_0)'],\ \hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(W_i, \hat{\theta}_{GMM})g(W_i, \hat{\theta}_{GMM})'. \end{aligned}$$

Here Ψ is the probability limit of $\hat{\Psi}$. Define

$$V = (G'\Psi G)^{-1}G'\Psi\Omega\Psi G(G'\Psi G)^{-1},\ \hat{V} = (\hat{G}'\hat{\Psi}\hat{G})^{-1}\hat{G}'\hat{\Psi}\hat{\Omega}\hat{\Psi}\hat{G}(\hat{G}'\hat{\Psi}\hat{G})^{-1}.$$

Proposition GMM: *If Assumption GMM and other regularity conditions are satisfied (see Newey and McFadden, 1994) then*

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{d} N(0, V),\ \hat{V} \xrightarrow{p} V.$$

Proposition GMM: *If Assumption GMM and other regularity conditions are satisfied (see Newey and McFadden, 1994) then*

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{d} N(0, V),\ \hat{V} \xrightarrow{p} V.$$

We summarize the conclusions as "GMM is CAN with asymptotic variance V and $\hat{V}$ is a consistent estimator of the asymptotic variance."

An efficient choice of $\hat{\Psi}$, that minimizes the asymptotic variance of GMM, is a consistent estimator of Ω^{-1}, such as

$$\hat{\Psi} = [\sum_{i=1}^{n} g(W_i, \tilde{\theta}) g(W_i, \tilde{\theta})'/n]^{-1},$$

for some initial GMM estimator $\tilde{\theta}$. With this $\hat{\Psi}$ the estimator $\hat{\theta}_{GMM}$ is the oft applied two-step efficient GMM estimator.

This estimator can be quite biased when there are many moment conditions relative to the number of parameters, especially in IV settings.

GMM Example: Instrumental variables estimation (IV).

$$W = (Y, X, Z),\ Y = Z'\theta_0 + \varepsilon,\ E[X\varepsilon] = 0.$$

Here let $g(w, \theta) = x(y - z'\theta)$. The specification condition i) in Proposition GMM is

$$E[g(W, \theta_0)] = E[X\varepsilon] = 0.$$

The identification condition ii) is implied by

$$rank(G) = rank(E[XZ']) = p$$

For general GMM $rank(G) = p$ is sufficient for local identification of θ_0.

Here $\hat{g}(\theta) = \sum_{i=1}^n X_i(Y_i - Z_i'\theta)/n$ so $\hat{G} = -\sum_{i=1}^n X_i Z_i'/n$ and

$$\hat{\theta}_{GMM} = -(\hat{G}'\hat{\Psi}\hat{G})^{-1}\hat{G}'\hat{\Psi}\sum_{i=1}^n X_i Y_i/n.$$

This is an IV estimator.

COMPARING MLE AND GMM

The first order conditions for MLE makes it a special case of GMM. MLE first order conditions are

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f(W_i|\hat{\theta}_{MLE})}{\partial\theta} = 0.$$

The first order conditions for MLE are GMM with moment functions equal to the likelihood "score",

$$g(w,\theta) = \frac{\partial \ln f(w|\theta)}{\partial\theta}.$$

These moment functions minimize asymptotic variance of GMM (because MLE is efficient). Better in practice to maximize likelihood to avoid multiple solutions for first order conditions.

GMM sets $\hat{g}(\hat{\theta}_{GMM}) = 0$ in large samples whenever number of moments r is same as number of paramesters (i.e. dimension of θ).

For MLE, GMM, and any other estimator it is important to understand what the conditions for identification of θ_0 are and whether they are satisfied in an application.

Identification conditions are MLE: $f(W|\theta) \neq f(W|\theta_0)$ for any $\theta \neq \theta_0$; GMM: $E[g(W,\theta)] \neq 0$ if $\theta \neq \theta_0$.

Identification is necessary to be able to determine from the data what the value of the parameter is.

Can be difficult to understand in structural models because they are complicated but it is important to think about.

Can check $rank(\partial \hat{g}(\theta)/\partial\theta) = r$ in data for plausible θ values.

Identification requires $r \geq p$.

Which of MLE and GMM estimator uses the most information about the data distribution?

Which estimator is consistent under weaker restrictions than the other?

Should one always use the most efficient estimator?

What is the tradeoff between efficiency and robustness (consistency under misspecification) for these two types of estimators?

What is the tradeoff between efficiency and robustness from increasing the number of moment functions used in GMM?

What other reasons are there to use MLE or to use GMM?

One could (and should) think about such questions whenever there is a choice of which estimator to use.

INFERENCE

Conclusion of MLE and GMM propositions are

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V),\ \hat{V} \xrightarrow{p} V.$$

That is, $\hat{\theta}$ is consistent and asymptotically normal and $\hat{V}$ is a consistent estimator of its asymptotic variance.

Standard way to conduct inference is to treat $\hat{\theta}$ as if it was $N(\theta_0, \hat{V}/n)$ and construct confidence intervals and tests.

Asymptotic $1 - \alpha$ confidence interval for θ_{0j} is

$$\hat{\theta}_j \pm z_{1-\alpha/2}\sqrt{\hat{V}_{jj}/n},$$

where $z_{1-\alpha/2}$ is the $1 - \alpha$ quantile of standard $N(0, 1)$ random variable (i.e. $\Pr(N(0, 1) \leq z_{1-\alpha/2}) = 1 - \alpha/2$).

Can test $H_0 : h(\theta_0) = 0$ for $s \times 1$ vector of functions $h(\theta)$ using Wald test statistic

$$\hat{T} = nh(\hat{\theta})'[\hat{H}\hat{V}\hat{H}']^{-1}h(\hat{\theta}).$$

For α level test reject if $\hat{T} \geq c_{1-\alpha}$ where $c_{1-\alpha}$ is $1 - \alpha$ quantile of $\chi^2(s)$.

"All you need" is i) consistent and asymptotically normal $\hat{\theta}$ and consistent estimator $\hat{V}$ of asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$.

In structural models $\hat{\theta}$ and $\hat{V}$ may be difficult to compute. $\hat{V}$ may be particularly complicated for some estimators. Also, may want to allow for some aspects of likelihood or some moment functions to possibly be misspecified. When $\hat{\theta}$ is relatively easy to compute (some structural models) the bootstrap provides a way to get $\hat{V}$ that is robust to misspecification.

Nonparametric Bootstrap for GMM (and MLE): Algorithm is:

I) Draw n observations $W_1^b, ..., W_n^b$ at random with replacement from $W_1, ..., W_n$;

II) Compute $\hat{\theta}^b$ from $W_1^b, ..., W_n^b$ in exactly same way as $\hat{\theta}$ is computed from $W_1, ..., W_n$; e.g. for GMM compute $\hat{\mathbf{\Psi}}^b$and $\hat{g}^b(\theta) = \sum_{i=1}^n g(W_i^b, \theta)/n$ from $W_1^b, ..., W_n^b$, and then $\hat{\theta}^b = \mathbf{arg\,min}_{\theta \in \Theta} \hat{g}^b(\theta)' \hat{\mathbf{\Psi}}^b \hat{g}^b(\theta)$.

III) Repeat I) and II) B times to get $\hat{\theta}^1, ..., \hat{\theta}^B$.

IV) Use $SE(\hat{\theta}) = \sqrt{\sum_{b=1}^B (\hat{\theta}^b - \bar{\theta}^B)^2/B}$, $\bar{\theta}^B = \sum_{b=1}^B \hat{\theta}^b/B$; or use confidence interval based on distribution of $\hat{\theta} - \theta_0$ being approximated by distribution of $\hat{\theta}^b - \hat{\theta}$; that is for $\hat{q}^B(\alpha)$ the α^{th} quantile of the $\{\hat{\theta}^1, ..., \hat{\theta}^B\}$ the boostrap percentile confidence interval is $[2\hat{\theta} - \hat{q}^B(1-\alpha/2),\ 2\hat{\theta} - \hat{q}^B(\alpha/2)]$.

WARNING: Neither bootstrap nor standard asymptotics is correct if estimation imposes inequality restrictions on some of the parameters θ.

For example, some estimated parameters of structural models may be probabilities that are constrained to be between zero and one.

The bootstrap and standard asymptotic approximation are incorrect in this case (except for parameters that do not depend locally on the probabilities).

Problem is that there is a discontinuity in the limit distribution when parameter hits the boundary, where limiting distribution changes from normal to censored half normal. Andrews D.W.K. (2000, "Inconsistency of the Bootstrap When a Parameter is on the Boundary of the Parameter Space," *Econometrica* 68, 399-405.

Another bootstrap method does work, see Chernozhukov, Newey, Santos (2022).

Discuss here since inequality restrictions do show up in some structural models, this has already been used in applications, and it can be straightforward to apply.

INFERENCE WITH INEQUALITY RESTRICTIONS

We focus on linear IV estimation with equality and inequality restrictions .

The model is

$$Y_i = Z_i'\theta_0 + \varepsilon_i,\ E[X_i\varepsilon_i] = 0,\ (i = 1, \ldots, n),$$

where Y_i is a left-hand side endogenous variable, Z_i is a vector of right hand-side variables, and X_i is a vector of instrumental variables.

Here the data are $W = (Y, Z, X)$ are i.i.d..

First describe a test of null hypothesis

$$H_0 : F\theta_0 = f,\ H\theta_0 \leq h.$$

Then describe confidence interval by inverting test.

Base tests on the two step GMM estimator with optimal weighting matrix.

Let $\bar{\theta}$ be an initial IV estimator of θ_0 and $\hat{\Omega}$ be a corresponding estimator of $E[X_i X_i' \varepsilon_i^2]$, such as

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i' \hat{\varepsilon}_i^2, \ \hat{\varepsilon}_i = Y_i - Z_i' \bar{\theta}.$$

Let $\hat{g}(\theta) = (1/n) \sum_{i=1}^{n} X_i (Y_i - Z_i' \theta)$ be sample moments of products of instrumental varialbles and residuals and

$$\hat{Q}(\theta) = \sqrt{n} [\hat{g}(\theta)' \hat{\Omega}^{-1} \hat{g}(\theta)]^{1/2}.$$

Square root of usual GMM objective is convenient for asymptotic theory.

Restricted and unrestriced GMM estimators are

$$\hat{\theta} = \arg \min_{F\theta = f, \ H\theta \leq h} \hat{Q}(\theta), \ \hat{\theta}_u = \arg \min_{\theta} \hat{Q}(\theta).$$

Test statistic is

$$T = \hat{Q}(\hat{\theta}) - \hat{Q}(\hat{\theta}_u).$$

Construct critical value using the bootstrap.

Let recentered moment vector be

$$\hat{g}_i = X_i\hat{\varepsilon}_i - \frac{1}{n}\sum_{j=1}^{n} X_j\hat{\varepsilon}_j,\ \hat{\varepsilon}_i = y_i - Z_i'\hat{\theta}.$$

Let $b \in \{1, ..., B\}$ index a bootstrap draws, $w_1^b, \ldots, w_n^b$ be i.i.d. $N(0,1)$ independent of the data,

$$\hat{g}^b = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i^b\hat{g}_i.$$

This is called multiplier bootstrap.

Use local version of restricted parameter space, where δ with the same dimension as θ serves as a possible value of $\sqrt{n}(\hat{\theta} - \theta_0)$.

Slackness parameter is $r_n > 0$ that makes more constraints bind as r_n increases.

Local version of restricted parameter space is

$$\hat{\Delta}_n = \left\{\delta : F\delta = 0,\ H_j\delta \leq \mathsf{max}\left\{0, -\sqrt{n}\left\{r_n + H_j\hat{\theta} - h_j\right\}\right\},\ \text{for all } j\right\}.$$

$$\hat{\Delta}_n = \left\{\delta : F\delta = 0,\ H_j\delta \leq \max\left\{0, -\sqrt{n}\left\{r_n + H_j\hat{\theta} - h_j\right\}\right\},\ \text{for all } j\right\}.$$

Let $\hat{G} = -\sum_{i=1}^n X_i Z_i'/n$.

The level α critical value is then the $1-\alpha$ quantile over bootstrap replications b of

$$\begin{aligned} T^b &= \min_{\delta\in\hat{\Delta}_n}\left\{(\hat{g}^b + \hat{G}\delta)'\hat{\Omega}^{-1}\left(\hat{g}^b + \hat{G}\delta\right)\right\}^{1/2} \\ &\quad - \min_{\delta}\left\{(\hat{g}^b + \hat{G}\delta)'\hat{\Omega}^{-1}\left(\hat{g}^b + \hat{G}\delta\right)\right\}^{1/2}. \end{aligned}$$

The main assumption required for the test to be asymptotically valid is that θ_0 be strongly identified.

Critical value depends on slackness parameter r_n; asymptotics requires r_n converges to zero slower than convergence rate of $\hat{\theta}$, i.e. slower than $1/\sqrt{n}$.

Heuristically, when r_n tends to zero any constraint that is not binding at θ_0 will also not be binding in the bootstrap with probability approaching one, so inference is not asymptotically conservative for a fixed data generating process; do "give up" a little power to make inference uniform.

$$\hat{\Delta}_n = \left\{\delta : F\delta = 0,\ H_j\delta \leq \max\left\{0, -\sqrt{n}\left\{r_n + H_j\hat{\theta} - h_j\right\}\right\},\ \text{for all } j\right\}$$

Setting $r_n = +\infty$ always theoretically valid, but may be conservative (all constraints are binding in bootstrap) and result in loss of power.

Other smaller choices give more powerful tests and tighter confidence intervals.

Can choose r_n based on data; intuitively r_n meant to quantify sampling uncertainty in $H(\hat{\theta} - \theta_0)$; cannot estimate distribution of $H(\hat{\theta} - \theta_0)$ uniformly consistently, so link r_n to sampling uncertainty $H(\hat{\theta}^u - \theta_0)$.

Choose r_n so that, for bootstrap analog $\hat{\theta}^{u*}$,

$$\Pr(\max_j\{H_j(\hat{\theta}^u - \hat{\theta}^{u*})\} \leq r_n) \approx 1 - \gamma_n.$$

for some γ_n; more interpretable as a tail probability than r_n; in simulations choice of γ_n does not matter much.

Can get confidence intervals for a linear combination $F\theta_0$ while imposing $H\theta_0 \leq h$ by inverting the test statistic.

For a given α find set of f such that test statistic is less than or equal to critical value for

$$H_0 : F\theta_0 = f,\ H\theta_0 \leq h.$$

Paper describes test and proves validity for partially identified parameters.

Here need additional tuning parameter for estimating identified set that is upper bound on objective function.

Paper describes test and proves validity for nonlinear moment, nonparametric θ models.

Need additional tuning parameter for nonlinear moments involving linearization of moment restrictions.

EMPIRICAL EXAMPLE

We illustrate the preceding discussion by revisiting the study by Angrist and Evans (1998) on the causal effect of childbearing on female labor force participation.

Use the 1980 Census Public Use Micro Sample restricted to mothers aged 21-35 with at least two children.

Outcome of interest is binary variable indicating whether mother is employed.

Treatment binary variable indicating mother has more than two children.

Instrument is indicator for whether the first two children are of the same sex.

Parameter of interest is average treatment effect for compliers (LATE).

Angrist and Evans (1998) document that the impact of childbearing on labor force participation depends on observable characteristics.

In particular, their two stage least squares (2SLS) estimates suggest a negative impact of childbearing on labor force participation across different levels of schooling, with magnitude of the impact decreasing with schooling.

Phenomenon may reflect the fact that more educated mothers have a stronger attachment to the labor force.

To examine this claim we introduce dummy variables S for each year of schooling between 9 and 16 and for the categories less than 9 and more than 16.

We test whether: (i) LATE(s) is increasing in schooling, and (ii) LATE(s) is increasing in schooling and nonpositive.

Both hypotheses fall within the framework of linear IV with inequality restrictions.

LATE(s) is identified through linear moment restrictions and the hypothesized restrictions are linear in LATE(s).

Use five thousand bootstrap replications and setting $r_n = +\infty$ or r_n as determined by $\gamma_n = .05$.

The p-value for LATE(s) being nondecreasing is 0.21.

The p-value for LATE(s) being nondecreasing and nonpositive is 0.394.

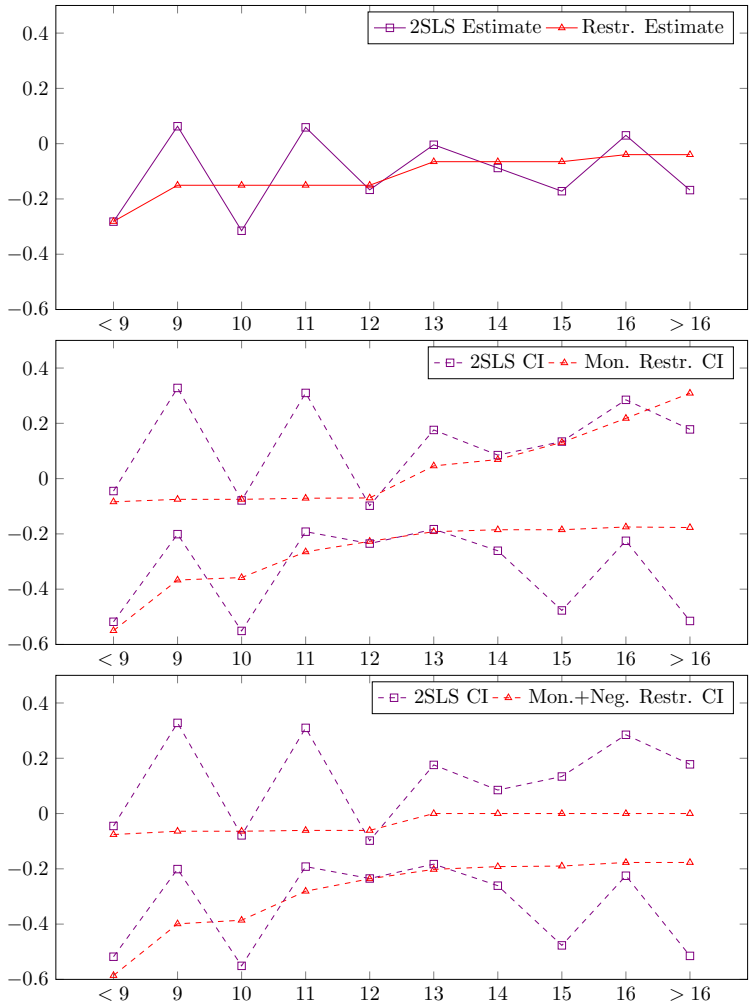


Figure 1: First Panel: Unconstrained and shape restricted LATE estimates (imposing monotonicty or monotonicity and negativity yield the same estimates). Second and Third Panels: 95% Confidence intervals for LATE at different education levels.

to mothers aged 21-35 with at least two children, and set: (i) $D \in \{0,1\}$ to indicate whether a mother has more than two children (the treatment); (ii) $Y \in \{0,1\}$ to indicate whether a mother is employed (the outcome of interest); and (iii) $Z \in \{0,1\}$ to indicate whether the first two children are of the same sex (the instrument). We further adopt the heterogeneous treatment effects model of Imbens and Angrist (1994) and let Y_d denote the potential outcome under treatment status $d \in \{0,1\}$ and employ "`C`," "`NT`," and "`AT`" to denote compliers, never takers, and always takers.

Angrist and Evans (1998) document that the impact of childbearing on labor force participation depends on observable characteristics. In particular, their two stage least squares (2SLS) estimates suggest a negative impact of childbearing on labor force participation across different levels of schooling, but that the magnitude of the impact decreases with schooling – a phenomenon that may reflect that more educated moth-

Figure 1 gives values of LATE(s) at different schooling levels.

The first panel displays the unconstrained 2SLS estimates and their monotonicity restricted counterparts.

The latter are negative and hence also requiring nonpositive effects does not change the estimates.

Second panel of Figure 1 we give 95% confidence intervals while imposing monotonicity.

Set $\gamma_n = .05$ to get r_n.

Imposing monotonicity yields confidence intervals that are sometimes substantially shorter than 2SLS counterparts.

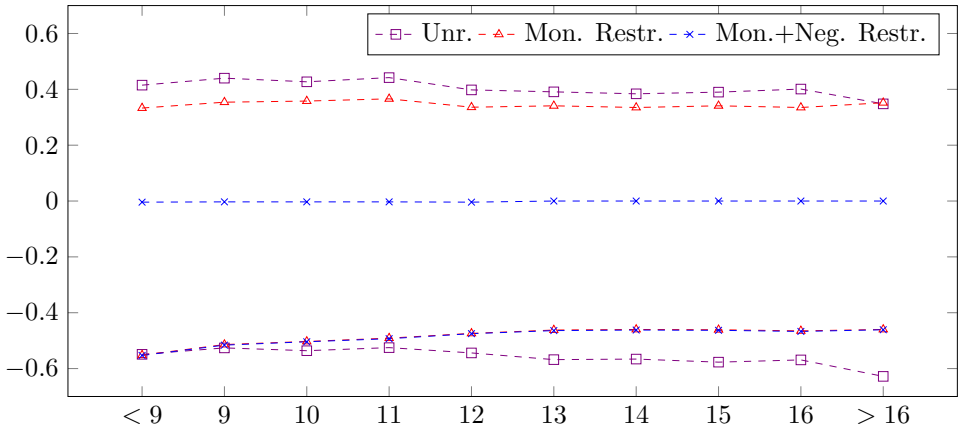


Figure 2: 95% Confidence intervals for ATE at different education levels. "Unr." uses bounds implied by $Y_d \in \{0, 1\}$; "Mon. Restr." adds that average treatment effects be increasing in education for all types; "Mon.+Neg. Restr." also requires they be negative.

obtained through the approach described in Remark 2.3 – here, the restriction $G\theta \leq g$ imposes the described shape constraints while the nonlinear restriction $\Upsilon_F(\theta) = 0$ corresponds to imposing a hypothesized value for ATE(S) through (11). In our bootstrap approximation, we let $\tau_n = 0$ and set r_n according to (7) with $\gamma_n = 0.05$ and where we used the distribution of estimators of identified parameters for their partially identified counterparts.[3] We do not report estimates of the identified sets for ATE(S) as they are very close to the obtained confidence intervals: On average the bounds of the confidence intervals exceed the bounds of the estimates by 0.011. Nonetheless, the unrestricted confidence intervals are large as the estimates for the identified set are large – a result driven by the low proportion of compliers (5% on average across S). Imposing monotonicity across types carries identifying information on the upper end of the identified set at low levels of education and on the lower end of the identified set at high levels of education. Additionally imposing nonpositivity sharpens the upper bound of the identified set at all schooling levels. The resulting confidence regions sign ATE(S) at all education levels (weakly) smaller than 12 as strictly negative, though very close to zero.

Finally, as a preview of our general analysis in Section 3, in Table 1 we employ the same shape restrictions to report estimates and 95% confidence intervals for the identified sets of the average treatment effects for: High School Dropouts (edu $\in [9, 12)$), College Dropouts (edu $\in [13, 15)$), College Graduates (edu ≥ 16) and the overall average treatment effect. These confidence regions are obtained through test inversion after noting that a hypothesized value for the average treatment effect of a subgroup can be written as a nonlinear moment restriction in θ_0 through (11) – nonlinear moment restrictions fall within our general framework but outside the scope of Section 2.2. Overall the impact of imposing shape restrictions parallels the results in Figure 2.

[3] E.g., for the constraint $E[Y_1|\texttt{NT}, S] \leq 1$ we substituted the corresponding $G_j\{\hat{\theta}_n^{\text{u}} - \hat{\theta}_n^{\text{u}^\star}\}$ term in (7)

	Unrestricted		Mon. Restr.		Mon.+Neg Restr.	
Subgroup	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
HS Drop	[-0.520,0.426]	[-0.526,0.432]	[-0.489,0.346]	[-0.500,0.356]	[-0.489,-0.008]	[-0.501,-0.003]
Coll. Drop	[-0.561,0.380]	[-0.566,0.385]	[-0.447,0.325]	[-0.460,0.337]	[-0.447,-0.004]	[-0.462,0.000]
Coll. Grad	[-0.579,0.375]	[-0.586,0.382]	[-0.446,0.328]	[-0.462,0.339]	[-0.446,-0.002]	[-0.464,0.000]
All	[-0.545,0.395]	[-0.547,0.398]	[-0.467,0.328]	[-0.477,0.338]	[-0.467,-0.008]	[-0.478,-0.003]

Table 1: Point Estimates and 95% confidence intervals for the average treatment effect at different groups defined by schooling levels under different shape restrictions.

3 General Analysis

We next develop a general inferential framework that encompasses the tests discussed in Section 2. The class of models we consider are those in which the parameter of interest $\theta_0 \in \Theta$ satisfies a finite number $\mathcal{J}$ of conditional moment restrictions

$$E_P[\rho_\jmath(X, \theta_0)|Z_\jmath] = 0 \text{ for } 1 \leq \jmath \leq \mathcal{J}$$

with $\rho_\jmath : \mathbf{X} \times \Theta \to \mathbf{R}$, $X \in \mathbf{X}$, and $Z_\jmath \in \mathbf{Z}_\jmath$. For notational simplicity, we also let $Z \equiv (Z_1, \ldots, Z_\mathcal{J})$ and $V \equiv (X, Z)$ with $V \sim P \in \mathbf{P}$. In some of the applications that motivate us, the parameter θ_0 is not identified. We therefore define the identified set

$$\Theta_0 \equiv \{\theta \in \Theta : E_P[\rho_\jmath(X, \theta)|Z_\jmath] = 0 \text{ for } 1 \leq \jmath \leq \mathcal{J}\}$$

and employ it as the basis of our statistical analysis – we emphasize that Θ_0 depends on P, but leave such dependence implicit to simplify notation. For a set R of parameters satisfying a conjectured restriction, we develop a test for the hypothesis

$$H_0 : \Theta_0 \cap R \neq \emptyset \qquad H_1 : \Theta_0 \cap R = \emptyset; \tag{12}$$

i.e. we devise a test of whether at least one element of the identified set satisfies the posited constraint. In what follows, we denote the set of distributions $P \in \mathbf{P}$ satisfying the null hypothesis in (12) by $\mathbf{P}_0$. We also note that in an identified model, a test of (12) is equivalent to a test of whether θ_0 itself satisfies the hypothesized constraint.

The defining elements determining the type of applications encompassed by (12) are the choices of Θ and R. In imposing restrictions on Θ and R we therefore aim to allow for a general framework while simultaneously ensuring enough structure for a fruitful asymptotic analysis. To this end, we require Θ to be a subset of a complete vector space $\mathbf{B}$ with norm $\|\cdot\|_\mathbf{B}$ (i.e. $(\mathbf{B}, \|\cdot\|_\mathbf{B})$ is a Banach space) and consider sets R satisfying

$$R = \{\theta \in \mathbf{B} : \Upsilon_F(\theta) = 0 \text{ and } \Upsilon_G(\theta) \leq 0\}, \tag{13}$$

with a mean zero normal distribution with the variance of the estimator for $E[Y_0|\mathtt{NT}, S]$.

Also estimate identified set of ATE.

Estimated intervals quite large due to compliers being a low proportion of population.

Estimated identified set is values where objective function is less than or equal to some τ_n.

Data based choice of τ_n is discussed in the paper.

Find wide but not uninformative confidence intervals.

Paper is at https://arxiv.org/abs/1509.06311.

METHOD OF SIMULATED MOMENTS

In some settings the moment functions are integrals, i.e.

$$g(w,\theta) = \int k(w,s,\theta) f(s) ds,$$

where $f(s)$ is **known** pdf. Here all known parameters are included in $k(w,s,\theta)$, including means and standard deviations of Gaussian or other variables in s.

Computation of the integral moment functions may be difficult when integral does not exist in closed form.

One can replace $g(w,\theta)$ by an unbiased estimator obtained by a finite number of simulations without affecting consistency of GMM; Pakes (1986), McFadden (1989), Pakes and Pollard (1989).

Let $S_1, ..., S_M$ be M random variables, independent of W, with marginal pdf $f(s)$.

Consider "augmented" data observation $W^M = (W, S_1, ..., S_M)$ where simulation draws are added to the data vector.

Simulations are obtained for each observation on W and are thereafter held fixed, i.e. do not resimulate as GMM estimator is computed; no resimulation is important for consistency and good standard errors.

Simulated moment function is

$$g_M(W^M, \theta) = \frac{1}{M} \sum_{m=1}^{M} k(W, S_m, \theta).$$

Can construct GMM estimator using this simulated moment function.

$$\hat{g}_M(\theta) = \frac{1}{n} \sum_{i=1}^{n} g_M(W_i^M, \theta) = \frac{1}{nM} \sum_{i=1}^{n} \sum_{m=1}^{M} k(W_i, S_{ji}, \theta)$$

A simulated GMM estimator is

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \hat{g}_M(\theta)' \hat{\Psi} \hat{g}_M(\theta).$$

fSimulated GMM will be CAN under Assumption GMM, i.e. under same specification and identification conditions as for GMM. By independence of $S_1, ..., S_M$ and W, iterated expectations gives

$$\begin{aligned} E[g_M(W^M, \theta)] &= E[E[g_M(W^M, \theta)|W]] = E[\frac{1}{M} \sum_{m=1}^{M} E[k(W, S_m, \theta)|W]] \\ &= E[\int k(W, s, \theta) f(s) ds] = E[g(W, \theta)]. \end{aligned}$$

Thus we have

$$E[g_M(W^M, \theta)] = 0 \iff E[g(W, \theta)] = 0,$$

Example (Binary Choice): $Y \in \{0, 1\}$, $\Pr(Y = 1|X) = H(X'\theta_0)$ where $H(u) = \int 1(s \leq u) f(s) ds$. Can form moment function $A(X)[y - H(x'\theta)]$. This moment function has the previous integral form with s having CDF $f(s)$ and

$$k(w, s, \theta) = A(x)[y - 1(s < x'\theta)].$$

The simulated moments are

$$g_M(W^M, \theta) = A(X)[Y - \frac{1}{M} \sum_{m=1}^{M} 1(S_m < X'\theta)].$$

Simulated moments replace choice probability with average of simulated choices. This is a "toy" example but is potentially important for more choices than two. With many alternatives the choice probabilities are much more complicated, so simulation could be more useful.

Computation is challenging due to discontinuity in parameters of moment functios; this makes asymptotic theory hard too; see Pakes and Pollard (1989) and Newey and McFadden (1994).

$$\Omega_M = \Omega + E[Var(g_M(W^M, \theta_0)|W)].$$

Consequently, asymptotic variance of simulated GMM is larger than asymptotic variance of GMM. Asymptotic variance of simulated GMM is

$$V = (G'\Psi G)^{-1} G'\Psi \Omega_M \Psi G (G'\Psi G)^{-1}.$$

This V is equal to GMM formula with Ω_M replacing Ω, so larger than GMM formula by Ω_M larger than Ω.

Can estimate V in usual way for GMM using simulated moment functions. Let

$$\hat{\Omega}_M = \frac{1}{n} \sum_{i=1}^{n} g_M(W_i^M, \hat{\theta}) g_M(W_i^M, \hat{\theta})'.$$

Let $\hat{G}$ be as for GMM or a numerical derivative of $\hat{g}_M(\theta)$ if simulated moments not differentiable; see Pakes and Pollard (1989). Then an estimator of V is

$$\hat{V} = (\hat{G}'\hat{\Psi}\hat{G})^{-1} \hat{G}'\hat{\Psi}\hat{\Omega}_M \hat{\Psi}\hat{G} (\hat{G}'\hat{\Psi}\hat{G})^{-1}.$$

Ω_M has a special form in the binary choice example when $S_1, ..., S_M$ are i.i.d.

$$\begin{aligned}\Omega &= E[A(X)A(X)'(Y - H(X'\theta_0))^2] \\ &= E[A(X)A(X)'Var(Y|X)] \\ &= E[A(X)A(X)'H(X'\theta_0)\{1 - H(X'\theta_0)\}]\end{aligned}$$

Also, since $S_1, ..., S_M$ i.i.d. conditional on W,

$$\begin{aligned}E[Var(g_M(W_M, \theta_0)|W)] &= E[\frac{1}{M}Var(A(X)[Y - 1(S_m < X'\theta_0)]|W)] \\ &= \frac{1}{M}E[A(X)A(X)'Var(1(S_m < X'\theta_0)|W)] \\ &= \frac{1}{M}\Omega.\end{aligned}$$

Thus for binary choice and i.i.d. simulations the asymptotic variance of the simulated GMM estimator is $1 + (1/M)$ times the asymptotic variance of the integral GMM estimator. Doing simulation increases asymptotic variance.

The simulation draws $S_1, ..., S_M$ are ancillary, meaning their distribution does not depend on the parameter of interest. An important statistical tradition, motivated by interesting examples, suggests that we should do inference conditional on ancillary objects. If we condition on the simulation draws, there is no noise from them, but there is bias! Will need $M \longrightarrow \infty$ to get consistency. Current econometric practice treats simulation draws as random.

DISCRETE CHOICE

Multinomial Choice: Each individual chooses one of J alternatives to maximize utility

$$V_j(X, \theta_0) + \varepsilon_j,$$

where ε_j represents preference heterogeneity and $\varepsilon_1, ..., \varepsilon_J$ is independent of X with known pdf.

The probability that alternative j is chosen is

$$\begin{aligned} P_j(X, \theta_0) &= \mathsf{Pr}(V_j(X, \theta_0) + \varepsilon_j \geq V_k(X, \theta_0) + \varepsilon_k \text{ for all } k|X) \\ &= \int [\Pi_{j=1}^{J} 1(s_j - s_k \geq V_k(X, \theta_0) - V_j(X, \theta_0))] f(s_1, ..., s_J) ds_1 ... ds_J. \end{aligned}$$

For J more than a few this probability is an integral that is very difficult to compute is most cases.

Could use MSM but turns out to not work super well in this setting, and needs much adjustment.

Multinomial Logit: For $\varepsilon_1, ..., \varepsilon_J$ independent Type I extreme value random variables $P_j(X, \theta)$ has a closed form

$$P_j(X, \theta_0) = \frac{\exp(V_j(X, \theta))}{\sum_{k=1}^{J} \exp(V_k(X, \theta))}$$

In addition to closed form the likelihood has nice numerical properties. For Y_j equal to 1 if choice j is made and zero otherwise, the log likelihood

$$\ln f(Y|X, \theta) = \sum_{j=1}^{J} Y_j \ln P_j(X, \theta)$$

is concave in $V_1(X, \theta), ..., V_J(X, \theta)$. Hence if each $V_j(X, \theta)$ is linear in parameters, say $V_j(X, \theta) = X_j'\theta$, then the log-likelihood is concave in parameters. Makes MLE straightforward.

Multinomial Logit has an important limitation that

$$P_j(X, \theta)/P_k(X, \theta) = \exp(V_j(X, \theta) - V_k(X, \theta)),$$

i.e. ratio of probabilities only depends on utility of choices j and k. This puts sever limitations on substitution patterns across choices. For example, if $V_j(X, \theta)$ depends only on the price of alternative j then the ratio of probabilities of any two goods only depends on the prices of those two alternatives. This limitation of multinomial logit is often referrred to as Independence from Irrelevant Alternatives (IIA).

Mixed Multinomial Logit: This refers to a way to relax IIA that has become quite important in practice. One way to describe this is to specify that the utility of choice j is

$$V_j(X, \theta_0, s_{J+1}) + \varepsilon_j, \ (j = 1, ..., J),$$

where s_{J+1} is a vector of unobserved preference variables that is independent of $\varepsilon_1, ..., \varepsilon_J$, which are independent Type I extreme value. By iterated expectations the choice probability is now

$$\begin{aligned} P_j(X, \theta_0) &= E[\mathsf{Pr}(Y_j = 1|X, s_{J+1})|X] \\ &= \int \frac{\mathsf{exp}(V_j(X, \theta_0, s_{J+1}))}{\sum_{k=1}^{J} \mathsf{exp}(V_k(X, \theta_0, s_{J+1}))} f(s_{J+1}) ds_{J+1}. \end{aligned}$$

MSM works well here. Simulated choice proabilities are

$$\hat{P}_j^M(X, \theta) = \frac{1}{M} \sum_{m=1}^{M} \left\{ \frac{\mathsf{exp}(V_j(X, \theta, s_{J+1,m}))}{\sum_{k=1}^{J} \mathsf{exp}(V_k(X, \theta, s_{J+1,m}))} \right\}.$$

This works well in practice; here estimated choice probability $\hat{P}_j^M(X, \theta)$ is nice smooth function of θ; that seems to help.

Does break IIA; how much this frees up cross-price effects is a matter of debate.

$$\hat{P}_j^M(X,\theta) = \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{\exp(V_j(X,\theta,s_{J+1,m}))}{\sum_{k=1}^{J}\exp(V_k(X,\theta,s_{J+1,m}))}\right\}$$

Here $s_{J+1,m}$ has known distribution so mean and standard deviation of simulation contained in θ.

In estimation take $s_{J+1,m}$ to be fixed as change θ; otherwise asymptotic theory does not work.

Example:

$$V_j(X,\theta,s_{J+1}) = X_j'(\mu + Bs_{J+1}).$$

This is heterogenous coefficient specification where μ is a vector of means, B a lower triangular matrix, s_{j+1} is a vector of i.i.d. standard normal variables.

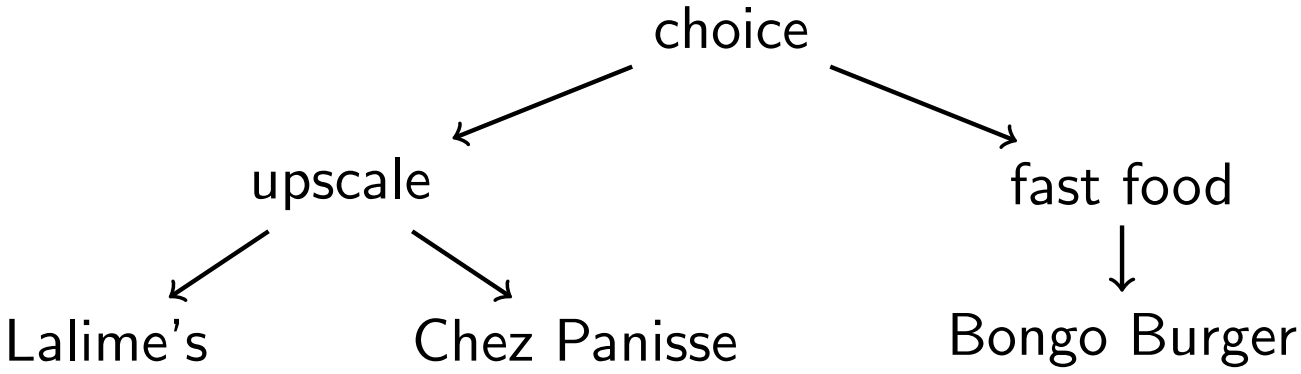
Here θ is nonzero elements of $(\mu', vec(B)')'$. For each i obtain $s_{J+1,m,i}$, $m = 1, ..., M$; then

$$\hat{P}_j^M(X,\theta) = \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{\exp(X_{ij}'(\mu + Bs_{J+1,m,i}))}{\sum_{k=1}^{J}\exp(X_{ik}'(\mu + Bs_{J+1,m,i}))}\right\}.$$

Estimate θ by doing multinomial choice MLE on for these choice probabilities holding all $M \times N$ simulations $s_{J+1,m,i}$ $(i = 1, ..., n; m = 1, ..., M)$ fixed (i.e. treating simulations as data).

The Nested Logit Model

A simple way to relax IIA is to nest similar alternatives.



Choices are nested into "upscale" and "fast food", in order to make the following distinction:

- Choices within nests: Among upscale restaurants p_{ij}/p_{ik} may not depend on the attributes of the fast food restaurant (IIA within nests). E.g., if prices at Bongo Burger increase, Lalime's and Chez Panisse gain clients in a way proportional to their previous market shares.
- Choices between nests: The ratio of choice probabilities, p_{ij}/p_{ik}, between Chez Panisse and Bongo Burger may very well depend on the attributes (e.g., prices) at Lalime's (no IIA between nests).

The Nested Logit Model

Suppose that alternatives are grouped into S nests: $B_1, \ldots, B_S$.

Assume that alternative j belongs to nest s. Conditional on the choice being in nest s, the probability of choosing j is:

$$\frac{e^{X_{ij}'\beta/\rho_s}}{\sum_{k \in B_s} e^{X_{ik}'\beta/\rho_s}}.$$

The probability of choosing an alternative that belongs to nest s is:

$$\frac{e^{Z_s'\alpha}\left(\sum_{k \in B_s} e^{X_{ik}'\beta/\rho_s}\right)^{\rho_s}}{\sum_{r=1}^{S} e^{Z_r'\alpha}\left(\sum_{k \in B_r} e^{X_{ik}'\beta/\rho_r}\right)^{\rho_r}}.$$

The variables Z_s are nest-specific characteristics, often nest-specific constants, in which case $e^{Z_s'\alpha} = e^{\alpha_s}$ (usually normalizing $\alpha_1 = 0$).

The Nested Logit Model

These probabilities can be derived as the solution of a model with a utility-maximizing agent that derives utility from choice j equal to:

$$U_{ij} = X'_{ij}\beta + Z'_s\alpha_s + u_{ij},$$

where the joint distribution function of $(u_{i0}, \ldots, u_{im})$ has a generalized extreme value (GEV) distribution:

$$F(u_{i0}, \ldots, u_{im}) = e^{-\sum_{s=1}^{S}\left(\sum_{j\in B_s} e^{-u_{ij}/\rho_s}\right)^{\rho_s}},$$

where $0 < \rho_s \leq 1$. Between nests u_{ij} and u_{ik} are independent. Within a nest the correlation between u_{ij} and u_{ik} is equal to $1 - \rho_s^2$, where s is the index of the nest.

The Nested Logit Model

Multiplying the probability of j given B_s times the probability of B_s, we obtain the probability of choice j:

$$p_{ij} = \frac{e^{X_{ij}'\beta/\rho_s + Z_s'\alpha}\left(\sum_{k\in B_s} e^{X_{ik}'\beta/\rho_s}\right)^{\rho_s - 1}}{\sum_{r=1}^{S} e^{Z_r'\alpha}\left(\sum_{k\in B_r} e^{X_{ik}'\beta/\rho_r}\right)^{\rho_r}}.$$

Using this formula, it is easy to see that:

- If both j and k belong to the same nest, then p_{ij}/p_{ik} depends only on $X_{ij} - X_{ik}$, so IIA holds.
- If j and k belong to different nests, B_s and B_r, and it is not the case that $\rho_s = \rho_r = 1$, then p_{ij}/p_{ik} may depend on the attributes of other alternatives in B_s and B_r, so IIA does not hold.
- If $\rho_1 = \cdots = \rho_S = 1$ we go back to the conditional logit model.