"What's New in Econometrics"

Lecture 11

Discrete Choice Models

Guido Imbens

NBER Summer Institute, 2007

Outline

1. Introduction
2. Multinomial and Conditional Logit Models
3. Independence of Irrelevant Alternatives
4. Models without IIA
5. Berry-Levinsohn-Pakes
6. Models with Multiple Unobserved Choice Characteristics
7. Hedonic Models

1. Introduction

Various versions of multinomial logit models developed by McFadden in 70's.

In IO applications with substantial number of choices IIA property found to be particularly unattractive because of unrealistic implications for substitution patterns.

Random effects approach is more appealing generalization than either nested logit or unrestricted multinomial probit

Generalization by BLP to allow for endogenous choice characteristics, unobserved choice characteristics, using only aggregate choice data.

2. Multinomial and Conditional Logit Models

Models for discrete choice with more than two choices.

The choice Y_i takes on non-negative, unordered integer values between zero and J.

Examples are travel modes (bus/train/car), employment status (employed/unemployed/out-of-the-laborforce), car choices (suv, sedan, pickup truck, convertible, minivan).

We wish to model the distribution of Y in terms of covariates individual-specific, choice-invariant covariates Z_i (e.g., age) choice (and possibly individual) specific covariates X_{ij}.

2.A Multinomial Logit

Individual-specific covariates only.

$$\Pr(Y_i = j | Z_i = z) = \frac{\exp(z'\gamma_j)}{1 + \sum_{l=1}^{J} \exp(z'\gamma_l)},$$

for choices $j = 1, \ldots, J$ and for the first choice:

$$\Pr(Y_i = 0 | Z_i = z) = \frac{1}{1 + \sum_{l=1}^{J} \exp(z'\gamma_l)},$$

The γ_l here are choice-specific parameters. This multinomial logit model leads to a very well-behaved likelihood function, and it is easy to estimate using standard optimization techniques.

2.B Conditional Logit

Suppose all covariates vary by choice (and possibly also by individual). The conditional logit model specifies:

$$\Pr(Y_i = j | X_{i0}, \ldots, X_{iJ}) = \frac{\exp(X_{ij}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)},$$

for $j = 0, \ldots, J$. Now the parameter vector β is common to all choices, and the covariates are choice-specific.

Also easy to estimate.

The multinomial logit model can be viewed as a special case of the conditional logit model. Suppose we have a vector of individual characteristics Z_i of dimension K, and J vectors of coefficients γ_j, each of dimension K. Then define

$$X_{i1} = \begin{pmatrix} Z_i \\ 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots\ldots \quad X_{iJ} = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ Z_i \end{pmatrix}, \text{ and } X_{i0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

and define the common parameter vector β as $\beta' = (\gamma_1', \ldots, \gamma_J')$. Then

$$\Pr(Y_i = 0|Z_i) = \frac{1}{1 + \sum_{l=1}^{J} \exp(Z_i'\gamma_l)}$$

$$= \frac{\exp(X_{ij}'\beta)}{\sum_{l=0}^{J} \exp(X_{il}'\beta)} = \Pr(Y_i = j|X_{i0}, \ldots, X_{iJ})$$

2.D Link with Utility Maximization

Utility, for individual i, associated with choice j, is

$$U_{ij} = X'_{ij}\beta + \varepsilon_{ij}. \tag{1}$$

i choose option j if choice j provides the highest level of utility

$$Y_i = j \quad \text{if } U_{ij} \geq U_{il} \text{ for all } l = 0, \ldots, J,$$

Now suppose that the ε_{ij} are independent accross choices and individuals and have type I extreme value distributions.

$$F(\epsilon) = \exp(-\exp(-\epsilon)), \qquad f(\epsilon) = \exp(-\epsilon) \cdot \exp(-\exp(-\epsilon)).$$

(This distribution has a unique mode at zero, a mean equal to 0.58, and a a second moment of 1.99 and a variance of 1.65.)

Then the choice Y_i follows the conditional logit model.

3. Independence of Irrelevant Alternatives

The main problem with the conditional logit is the property of Independence of Irrelevant Alternative (IIA).

The conditional probability of choosing j given either j or l:

$$\Pr(Y_i = j | Y_i \in \{j, l\}) = \frac{\Pr(Y_i = j)}{\Pr(Y_i = j) + \Pr(Y_i = l)}$$

$$= \frac{\exp(X'_{ij}\beta)}{\exp(X'_{ij}\beta) + \exp(X'_{il}\beta)}.$$

This probability does not depend on the characteristics X_{im} of alternatives m.

Also unattractive implications for marginal probabilities for new choices.

Although multinomial and conditional logit models may fit well, they are not necessarily attractive as behavior/structural models. because they generates unrealistic substitution patterns.

Suppose that individuals have the choice out of three restaurants, Chez Panisse (C), Lalime's (L), and the Bongo Burger (B). Suppose we have two characteristics, price and quality

price	$P_C = 95,\ P_L = 80,\ P_B = 5,$
quality	$Q_C = 10,\ Q_L = 9,\ Q_B = 2$
market share	$S_C = 0.10,\ S_L = 0.25,\ S_B = 0.65.$

These numbers are roughly consistent with a conditional logit model where the utility associated with individual i and restaurant j is

$$U_{ij} = -0.2 \cdot P_j + 2 \cdot Q_j + \epsilon_{ij},$$

Now suppose that we raise the price at Lalime's to 1000 (or raise it to infinity, corresponding to taking it out of business).

The conditional logit model predicts that the market shares for Lalime's gets divided by Chez Panisse and the Bongo Burger, proportional to their original market share, and thus $\tilde{S}_C = 0.13$ and $\tilde{S}_B = 0.87$: most of the individuals who would have gone to Lalime's will now dine (if that is the right term) at the Bongo Burger.

That seems implausible. The people who were planning to go to Lalime's would appear to be more likely to go to Chez Panisse if Lalime's is closed than to go to the Bongo Burger, implying $\tilde{S}_C \approx 0.35$ and $\tilde{S}_B \approx 0.65$.

Recall the latent utility set up with the utility

$$U_{ij} = X'_{ij}\beta + \epsilon_{ij}. \tag{2}$$

In the conditional logit model we assume independent extreme value ϵ_{ij}. The independence is essentially what creates the IIA property. (This is not completely correct, because other distributions for the unobserved, say with normal errors, we would not get IIA exactly, but something pretty close to it.)

The solution is to allow in some fashion for correlation between the unobserved components in the latent utility representation. In particular, with a choice set that contains multiple versions of similar choices (like Chez Panisse and LaLime's), we should allow the latent utilities for these choices to be similar.

4. Models without IIA

Here we discuss 3 ways of avoiding the IIA property. All can be interpreted as relaxing the independence between the ϵ_{ij}.

The first is the nested logit model where the researcher groups together sets of choices. This allows for non-zero correlation between unobserved components of choices within a nest and maintains zero correlation across nests.

Second, the unrestricted multinomial probit model with no restrictions on the covariance between unobserved components, beyond normalizations.

Third, the mixed or random coefficients logit where the marginal utilities associated with choice characteristics vary between individuals, generating positive correlation between the unobserved components of choices that are similar in observed choice characteristics.

Nested Logit Models

Partition the set of choices $\{0, 1, \ldots, J\}$ into S sets $B_1, \ldots, B_S$

Now let the conditional probability of choice j given that your choice is in the set B_s, be equal to

$$\Pr(Y_i = j | X_i, Y_i \in B_s) = \frac{\exp(\rho_s^{-1} X_{ij}' \beta)}{\sum_{l \in B_s} \exp(\rho_s^{-1} X_{il}' \beta)},$$

for $j \in B_s$, and zero otherwise. In addition suppose the marginal probability of a choice in the set B_s is

$$\Pr(Y_i \in B_s | X_i) = \frac{\left(\sum_{l \in B_s} \exp(\rho_s^{-1} X_{il}' \beta)\right)^{\rho_s}}{\sum_{t=1}^{S} \left(\sum_{l \in B_t} \exp(\rho_t^{-1} X_{il}' \beta)\right)^{\rho_s}}.$$

If we fix $\rho_s = 1$ for all s, then

$$\Pr(Y_i = j|X_i) = \frac{\exp(X'_{ij}\beta + Z'_s\alpha)}{\sum_{t=1}^{S}\sum_{l\in B_t}\exp(X'_{il}\beta + Z_t\alpha)},$$

and we are back in the conditional logit model.

The implied joint distribution function of the ϵ_{ij} is

$$F(\epsilon_{i0},\ldots,\epsilon_{iJ}) = \exp\left(-\sum_{s=1}^{S}\left(\sum_{j\in B_s}\exp\left(-\rho_s^{-1}\epsilon_{ij}\right)\right)^{\rho_s}\right).$$

Within the sets the correlation coefficient for the ϵ_{ij} is approximately equal to $1-\rho$. Between the sets the ϵ_{ij} are independent.

The nested logit model could capture the restaurant example by having two nests, the first $B_1 = \{\text{Chez Panisse}, \text{LaLime's}\}$, and the second one $B_2 = \{\text{Bongoburger}\}$.

Estimation of Nested Logit Models

Maximization of the likelihood function is difficult.

An easier alternative is to use the nesting structure. Within a nest we have a conditional logit model with coefficients β/ρ_s. Estimates these as $\widehat{\beta/\rho_s}$.

Then the probability of a particular set B_s can be used to estimate ρ_s through

$$\Pr(Y_i \in B_s|X_i) = \frac{\left(\sum_{l\in B_s} \exp(X'_{il}\widehat{\beta/\rho_s})\right)^{\rho_s}}{\sum_{t=1}^{S}\left(\sum_{l\in B_t} \exp(X'_{il}\widehat{\beta/\rho_t})\right)^{\rho_s}} = \frac{\exp(\rho_s \hat{W}_s)}{\sum_{t=1}^{S}\exp(\rho_t \hat{W}_t)},$$

where the "inclusive values" are

$$\hat{W}_s = \ln\left(\sum_{l\in B_s} \exp(X'_{il}\widehat{\beta/\rho_s})\right).$$

These models can be extended to many layers of nests. See for an impressive example of a complex model with four layers of multiple nests Goldberg (1995). Figure 2 shows the nests in the Goldberg application.

The key concern with the nested logit models is that results may be sensitive to the specification of the nest structure.

The researcher **chooses** which choices are potentially close substitutes, with the data being used to estimate the amount of correlation.

Researcher would have to choose nest for new good to estimate market share.

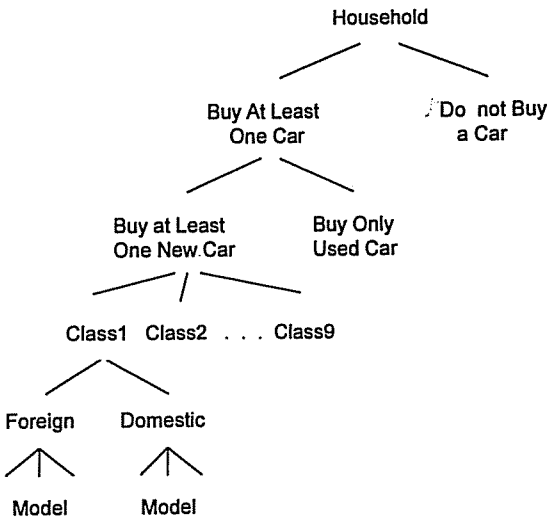

FIGURE 1.—Automobile choice model.

Multinomial Probit with Unrestricted Covariance Matrix

A second possibility is to directly free up the covariance matrix of the error terms. This is more natural to do in the multinomial probit case.

We specify:

$$U_i = \begin{pmatrix} U_{i0} \\ U_{i1} \\ \vdots \\ U_{iJ} \end{pmatrix} = \begin{pmatrix} X_{i0}'\beta + \epsilon_{i0} \\ X_{i1}'\beta + \epsilon_{i1} \\ \vdots \\ X_{iJ}'\beta + \epsilon_{iJ} \end{pmatrix} \quad \epsilon_i = \begin{pmatrix} \epsilon_{i0} \\ \epsilon_{i1} \\ \vdots \\ \epsilon_{iJ} \end{pmatrix} | X_i \sim \mathcal{N}(0, \Omega),$$

for some relatively unrestricted $(J+1) \times (J+1)$ covariance matrix Ω (beyond normalizations).

Direct maximization of the log likelihood function is infeasible for more than 3-4 choices.

Geweke, Keane, and Runkle (1994) and Hajivasilliou and McFadden (1990) proposed a way of calculating the probabilities in the multinomial probit models that allowed researchers to deal with substantially larger choice sets.

A simple attempt to estimate the probabilities would be to draw the ϵ_i from a multivariate normal distribution and calculate the probability of choice j as the number of times choice j corresponded to the highest utility.

The Geweke-Hajivasilliou-Keane (GHK) simulator uses a more complicated procedure that draws $\epsilon_{i1}, \ldots, \epsilon_{iJ}$ sequentially and combines the draws with the calculation of univariate normal integrals.

From a Bayesian perspective drawing from the posterior distribution of β and Ω is straightforward. The key is setting up the vector of unobserved random variables as

$$\theta = (\beta, \Omega, U_{i0}, \ldots, U_{iJ}) ,$$

and defining the most convenient partition of this vector.

Suppose we know the latent utilities U_i for all individuals. Then the normality makes this a standard linear model problem.

Given the parameters drawing from the unobserved utilities can be done sequentially: for each unobserved utility given the others we would have to draw from a truncated normal distribution, which is straightforward. See McCulloch, Polson, and Rossi (2000) for details.

Merits of Unrestriced Multinomial Probit

The attraction of this approach is that there are no restrictions on which choices are close substitutes.

The difficulty, however, with the unrestricted multinomial probit approach is that with a reasonable number of choices there are a large number of parameters: all elements in the $(J+1) \times (J+1)$ dimensional Ω minus some normalizations and symmetry restrictions.

Estimating all these covariance parameters precisely, based on only first choice data (as opposed to data where we know for each individual additional orderings, e.g., first and second choices), is difficult.

Prediction for new good would require specifying correlations with all other goods.

Random Effects Models

A third possibility to get around the IIA property is to allow for unobserved heterogeneity in the slope coefficients.

Why do we fundamentally think that if Lalime's price goes up, the individuals who were planning to go Lalime's go to Chez Panisse instead, rather than to the Bongo Burger? One argument is that we think individuals who have a taste for Lalime's are likely to have a taste for close substitute in terms of observable characteristics, Chez Panisse as well, rather than for the Bongo Burger.

We can model this by allowing the marginal utilities to vary at the individual level:

$$U_{ij} = X'_{ij}\beta_i + \epsilon_{ij},$$

We can also write this as

$$U_{ij} = X'_{ij}\overline{\beta} + \nu_{ij},$$

where

$$\nu_{ij} = \epsilon_{ij} + X_{ij} \cdot (\beta_i - \overline{\beta}),$$

which is no longer independent across choices.

One possibility to implement this is to assume the existence of a finite number of types of individuals, similar to the finite mixture models used by Heckman and Singer (1984) in duration settings:

$$\beta_i \in \{b_0, b_1, \ldots, b_K\},$$

with

$$\Pr(\beta_i = b_k | Z_i) = p_k, \quad \text{or} \quad \Pr(\beta_i = b_k | Z_i) = \frac{\exp(Z_i'\gamma_k)}{1 + \sum_{l=1}^{K} \exp(Z_i'\gamma_l)}.$$

Here the taste parameters take on a finite number of values, and we have a finite mixture.

Alternatively we could specify

$$\beta_i | Z_i \sim \mathcal{N}(Z_i'\gamma, \Sigma),$$

where we use a normal (continuous) mixture of taste parameters.

Using simulation methods or Gibbs sampling with the unobserved β_i as additional unobserved random variables may be an effective way of doing inference.

The models with random coefficients can generate more realistic predictions for new choices (predictions will be dependent on presence of similar choices)

5. Berry-Levinsohn-Pakes

BLP extended the random effects logit models to allow for

1. unobserved product characteristics,

2. endogeneity of choice characteristics,

3. estimation with only aggregate choice data

4. with large numbers of choices.

Their approach has been widely used in Industrial Organization, where it is used to model demand for differentiated products.

The utility is indexed by individual, product and market:

$$U_{ijt} = \beta_i' X_{jt} + \zeta_{jt} + \epsilon_{ijt}.$$

The ζ_{jt} is a unobserved product characteristic. This component is allowed to vary by market and product.

The ϵ_{ijt} unobserved components have extreme value distributions, independent across all individuals i, products j, and markets t.

The random coefficients β_i are related to individual observable characteristics:

$$\beta_i = \beta + Z_i'\Gamma + \eta_i, \quad \text{with} \quad \eta_i | Z_i \sim \mathcal{N}(0, \Sigma).$$

The data consist of

- estimated shares $\widehat{s}_{tj}$ for each choice j in each market t,

- observations from the marginal distribution of individual characteristics (the Z_i's) for each market, often from representative data sets such as the CPS.

First write the latent utilities as

$$U_{ijt} = \delta_{jt} + \nu_{ijt} + \epsilon_{ijt},$$

where

$$\delta_{jt} = \beta' X_{jt} + \zeta_{jt}, \quad \text{and} \quad \nu_{ijt} = (Z_i' \Gamma + \eta_i)' X_{jt}.$$

Now consider for fixed Γ, Σ and δ_{jt} the implied market share for product j in market t, s_{jt}.

This can be calculated analytically in simple cases. For example with $\Gamma_{jt} = 0$ and $\Sigma = 0$, the market share is a very simple function of the δ_{jt}:

$$s_{jt}(\delta_{jt}, \Gamma = 0, \Sigma = 0) = \frac{\exp(\delta_{jt})}{\sum_{l=0}^{J} \exp(\delta_{lt})}.$$

More generally, this is a more complex relationship which we may need to calculate by simulation of choices.

Call the vector function obtained by stacking these functions for all products and markets $s(\delta, \Gamma, \Sigma)$.

Next, fix only Γ and Σ. For each value of δ_{jt} we can find the implied market share. Now find the vector of δ_{jt} such that all implied market shares are equal to the observed market shares $\widehat{s}_{jt}$.

BLP suggest using the following algorithm. Given a starting value for δ_{jt}^0, use the updating formula:

$$\delta_{jt}^{k+1} = \delta_{jt}^k + \ln s_{jt} - \ln s_{jt}(\delta^k, \Gamma, \Sigma).$$

BLP show this is a contraction mapping, and so it defines a function $\delta(s, \Gamma, \Sigma)$ expressing the δ as a function of observed market shares s, and parameters Γ and Σ.

Given this function $\delta(s, \Gamma, \Sigma)$ define the residuals

$$\omega_{jt} = \delta_{jt}(s, \Gamma, \Sigma) - \beta' X_{jt}.$$

At the true values of the parameters and the true market shares these residuals are equal to the unobserved product characteristic ζ_{jt}.

Now we can use GMM given instruments that are orthogonal to these residuals, typically things like characteristics of other products by the same firm, or average characteristics by competing products.

This step is where the method is most challenging. Finding values of the parameters that set the average moments closest to zero can be difficult.

Let us see what this does if we have, and know we have, a conditional logit model with fixed coefficients. In that case $\Gamma = 0$, and $\Sigma = 0$. Then we can invert the market share equation to get the market specific unobserved choice-characteristics

$$\delta_{jt} = \ln s_{jt} - \ln s_{0t},$$

where we set $\delta_{0t} = 0$. (this is typically the outside good, whose average utility is normalized to zero). The residual is

$$\zeta_{jt} = \delta_{jt} - \beta' X_{jt} = \ln s_{jt} - \ln s_{0t} - \beta' X_{jt}.$$

With a set of instruments W_{jt}, we run the regression

$$\ln s_{jt} - \ln s_{0t} = \beta' X_{jt} + \epsilon_{jt},$$

using W_{jt} as instrument for X_{jt}, using as the observational unit the market share for product j in market t.

6. Models with Multiple Unobserved Choice Characteristics

The BLP approach can allow only for a single unobserved choice characteristic. This is essential for their estimation strategy with aggregate data.

With individual level data one may be able to establish the presence of two unobserved product characteristics (invariant across markets). Elrod and Keane (1995), Goettler and Shachar (2001), and Athey and Imbens (2007) study such models.

These models can be viewed as freeing up the covariance matrix of unobserved components relative to the random coefficients model, but using a factor structure instead of a fully unrestricted covariance matrix as in the multinomial probit.

Athey and Imbens model the latent utility for individual i in market t for choice j as

$$U_{ijt} = X'_{it}\beta_i + \zeta'_j\gamma_i + \epsilon_{ijt},$$

with the individual-specific taste parameters for both the observed and unobserved choice characteristics normally distributed:

$$\begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix} | Z_i \sim \mathcal{N}(\Delta Z_i, \Omega).$$

Even in the case with all choice characteristics exogenous, maximum likelihood estimation would be difficult (multiple modes). Bayesian methods, and in particular markov-chain-monte-carlo methods are more effective tools for conducting inference in these settings.

7. Hedonic Models

Recently researchers have reconsidered using pure characteristics models for discrete choices, that is models with no idiosyncratic error ϵ_{ij}, instead relying solely on the presence of a small number of unobserved product characteristics and unobserved variation in taste parameters to generate stochastic choices.

Why can it still be useful to include such an ϵ_{ij}?

First, the pure characteristics model can be extremely sensitive to measurement error, because it can predict zero market shares for some products.

Consider a case where choices are generated by a pure characteristics model that implies that a particular choice j has zero market share. Now suppose that there is a single unit i for whom we observe, due to measurement error, the choice $Y_i = j$.

Irrespective of the number of correctly measured observations available that were generated by the pure characteristics model, the estimates of the latent utility function will not be close to the true values due to a **single** mismeasured observation.

Thus, one might wish to generalize the model to be more robust. One possibility is to related the observed choice Y_i to the optimal choice Y_i^*:

$$\Pr(Y_i = y | Y_i^*, X_i, \nu_i, Z_1, \ldots, Z_J, \zeta_1, \ldots, \zeta_J)$$

$$= \begin{cases} 1 - \delta & \text{if } Y = Y_i^*, \\ \delta/(J-1) & \text{if } Y \neq Y_i^*. \end{cases}$$

This nests the pure characteristics model (by setting $\delta = 0$), without the extreme sensitivity.

However, if the optimal choice Y_i^* is not observed, all of the remaining choices are equally likely.

An alternative modification of the pure characteristics model is based on adding an idiosyncratic error term to the utility function. This model will have the feature that, conditional on the optimal choice not being observed, a close-to-optimal choice is more likely than a far-from-optimal choice.

Suppose the true utility is U_{ij}^* but individuals base their choice on the maximum of mismeasured version of this utility:

$$U_{ij} = U_{ij}^* + \epsilon_{ij},$$

with an extreme value ϵ_{ij}, independent across choices and individuals. The ϵ_{ij} here can be interpreted as an error in the calculation of the utility associated with a particular choice.

Second, this model approximately nests the pure characteristics model in the following sense. If the data are generated by the pure characteristics model with the utility function $g(x, \nu, z, \zeta)$, then the model with the utility function $\lambda \cdot g(x, \nu, z, \zeta) + \epsilon_{ij}$ leads, for sufficiently large λ, to choice probabilities that are arbitrarily close to the true choice probabilities (e.g., Berry and Pakes, 2007).

Hence, even if the data were generated by a pure characteristics model, one does not lose much by using a model with an additive idiosyncratic error term, and one gains a substantial amount of robustness to measurement or optimization error.

What's New in Econometrics?
Lecture 12
Missing Data

Jeff Wooldridge
NBER Summer Institute, 2007

1. When Can Missing Data be Ignored?

2. Inverse Probability Weighting

3. Imputation

4. Heckman-Type Selection Corrections

1. **When Can Missing Data be Ignored**?

- Linear model with IVs:

$$y_i = x_i\beta + u_i, \tag{1}$$

where x_i is $1 \times K$, instruments z_i are $1 \times L$, $L \geq K$. Let s_i is the selection indicator, $s_i = 1$ if we can use observation i. With $L = K$, the "complete case" estimator is

$$\hat{\beta}_{IV} = \left(N^{-1} \sum_{i=1}^{N} s_i z_i' x_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i z_i' y_i \right) \tag{2}$$

$$= \beta + \left(N^{-1} \sum_{i=1}^{N} s_i z_i' x_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} s_i z_i' u_i \right) \tag{3}$$

For consistency, $rank\, E(z_i' x_i | s_i = 1) = K$ and

$$E(s_i z_i' u_i) = 0, \tag{4}$$

which is implied by

$$E(u_i | z_i, s_i) = 0. \tag{5}$$

Sometimes we can add a function of z_i to the equation that forces (5) to be true. Sufficient for (5) is

$$E(u_i | z_i) = 0, \ s_i = h(z_i) \tag{6}$$

for some function $h(\cdot)$.

- Zero covariance assumption in the population, $E(z_i' u_i) = 0$, is not sufficient for consistency when $s_i = h(z_i)$. Special case is when

$E(y_i|x_i) = x_i\beta$ and selection s_i is a function of x_i.

- Nonlinear models/estimation methods:

Nonlinear Least Squares: $E(y|x,s) = E(y|x)$.

Least Absolute Deviations: $Med(y|x,s) = Med(y|x)$

Maximum Likelihood: $D(y|x,s) = D(y|x)$ or $D(s|y,x) = D(s|x)$.

- All of these allow selection on x but not generally on y. For estimating $\mu = E(y_i)$, unbiasedness and consistency of the sample on the selected sample requires $E(y|s) = E(y)$.

- Panel data: if we model $D(y_t|x_t)$, and s_t is the selection indicator, the sufficient condition to ignore selection is

$$D(s_t|x_t,y_t) = D(s_t|x_t),\ t = 1,\ldots,T. \tag{7}$$

Let the true conditional density be $f_t(y_{it}|x_{it},\gamma)$. Then the partial log-likelihood function for a random draw i from the cross section can be written as

$$\sum_{t=1}^{T} s_{it} \log f_t(y_{it}|x_{it},g) \equiv \sum_{t=1}^{T} s_{it} l_{it}(g). \tag{8}$$

Can show under (7) that

$$E[s_{it} l_{it}(g)|x_{it}] = E(s_{it}|x_{it}) E[l_{it}(g)|x_{it}]. \tag{9}$$

By the Kullback-Leibler information inequality,

$E[l_{it}(\gamma)|x_{it}] \geq E[l_{it}(g)|x_{it}]$ for all $g \in \Gamma$ (parameter space). Because

$E(s_{it}|x_{it}) = P(s_{it} = 1|x_{it}) \geq 0$, it follows that

$E[s_{it} l_{it}(\gamma)|x_{it}] \geq E[s_{it} l_{it}(g)|x_{it}]$ for all $g \in \Gamma$. Apply LIE again to

conclude γ maximizes the expected value of (8). We cannot just initially appeal to genearl MLE results; (8) is not a proper log-likelihood function.

- If x_{it} includes, say, $y_{i,t-1}$, then (7) allows selection to depend on $y_{i,t-1}$, but not on "shocks" from $t-1$ to t.
- Similar findings for nonlinear least squares, quasi-MLE, quantile regression.
- Methods to remove time-constant, unobserved heterogeneity: suppose we have the linear model, written for a random draw i,

$$y_{it} = \eta_t + x_{it}\beta + c_i + u_{it}, \tag{10}$$

with instruments z_{it} for x_{it}. Random effects IV methods on the

unbalanced panel use

$$E(u_{it}|z_{i1},\ldots,z_{iT},s_{i1},\ldots,s_{iT},c_i) = 0,\ t = 1,\ldots,T \tag{11}$$

and

$$E(c_i|z_{i1},\ldots,z_{iT},s_{i1},\ldots,s_{iT}) = E(c_i) = 0. \tag{12}$$

Selection in any time period cannot depend on u_{it} or c_i.

- FE on unbalanced panel means we can get by with just the first assumption. Let $\ddot{y}_{it} = y_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir}$ and similarly for and $\ddot{x}_{it}$ and $\ddot{z}_{it}$, where $T_i = \sum_{r=1}^{T} s_{ir}$ is the number of time periods for observation i.

The FEIV estimator is

$$\hat{\beta}_{FEIV} = \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{z}_{it}' \ddot{x}_{it} \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{z}_{it}' y_{it} \right).$$

Weakest condition for consistency is $\sum_{t=1}^{T} E(s_{it} \ddot{z}_{it}' u_{it}) = 0$.

- One important violation of (11) is when units drop out of the sample in period $t + 1$ because of shocks (u_{it}) realized in time t. This generally induces correlation between $s_{i,t+1}$ and u_{it}. To test, just add $s_{i,t+1}$ to the equation at time t, estimate the model by fixed effects (or FEIV), and compute (robust) t test.
- Consistency of FE (and FEIV) on the unbalanced panel under (11) breaks down if the slope coefficients are random and one ignores this in estimation. (Earlier: FE and FEIV still can produce consistent

estimators in balanced case.) The error term contains the term $x_i d_i$ where $d_i = b_i - \beta$. Simple test based on the alternative

$$\mathrm{E}(b_i|z_{i1}, \ldots, z_{iT}, s_{i1}, \ldots, s_{iT}) = E(b_i|T_i). \tag{13}$$

Then, add interaction terms of dummies for each possible sample size (with $T_i = T$ as the base group):

$$1[T_i = 2]x_{it},\ 1[T_i = 3]x_{it},\ \ldots,\ 1[T_i = T-1]x_{it}. \tag{14}$$

Estimate equation by FE or FEIV.

- Can use FD in basic model, too, which is very useful for attrition problems (later). Generally, if

$$\Delta y_{it} = \varphi_t + \Delta x_{it} + \Delta u_{it},\ t = 2, \ldots, T \tag{15}$$

and, if z_{it} is the set of IVs at time t, we can use

$$E(\Delta u_{it}|z_{it}, s_{it}) = 0 \tag{16}$$

as being sufficient to ignore the missingess. Again, can add $s_{i,t+1}$ to test for attrition.

- Not suprisingly, nonlinear models with unosberved effects are considerably more difficult to handle, although certain conditional MLEs (logit, Poisson) can accomodate selection that is arbitrarily correlated with the unobserved effect.

2. Inverse Probability Weighting

Weighting with Cross-Sectional Data

- When selection is not on conditioning variables, can try to use

probability weights to reweight the selected sample to make it representative of the populatin. Suppose y is a random variable whose population mean $\mu = E(y)$ we would like to estimate, but some observations are missing on y. Let $\{(y_i, s_i, z_i) : i = 1, \ldots, N\}$ indicate independent, identically distributed draws from the population, where z_i is always observed (for now). "Selection on observables" assumption

$$P(s = 1|y, z) = P(s = 1|z) \equiv p(z) \tag{17}$$

where $p(z) > 0$ for all possible values of z. Consider

$$\tilde{\mu}_{IPW} = N^{-1} \sum_{i=1}^{N} \left(\frac{s_i}{p(z_i)} \right) y_i, \tag{18}$$

where s_i selects out the observed data points. Using (17) and iterated

expectations, can show $\hat{\mu}_{IPW}$ is consistent (and unbiased) for y_i. (Same kind of estimate used for treatment effects.)

- Sometimes $p(z_i)$ is known (variable probability stratified sampling), but mostly it needs to be estimated. (And, even for VP sampling, it *should* be estimated if possible.) Let $\hat{p}(z_i)$ denote the estimated selection probability:

$$\hat{\mu}_{IPW} = N^{-1} \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(z_i)} \right) y_i. \tag{19}$$

Can also write as

$$\hat{\mu}_{IPW} = N_1^{-1} \sum_{i=1}^{N} s_i \left(\frac{\hat{\rho}}{\hat{p}(z_i)} \right) y_i \tag{20}$$

where $N_1 = \sum_{i=1}^{N} s_i$ is the number of selected observations and $\hat{\rho} = N_1/N$ is a consistent estimate of $P(s_i = 1)$. The weights reported to account for missing data are often $\hat{\rho}/\hat{p}(z_i)$.

- A different estimate is obtained by solving the least squares problem

$$\min_m \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(z_i)} \right) (y_i - m)^2.$$

- Horowitz and Manski (1998) have considered the problem of estimating population means using IPW. They focus on bounds in estimating $E[g(y)|x \in A]$ for conditioning variables, x. But they also note a problem with certain IPW estimators based on weights that estimate $P(s = 1)/P(s = 1|d = 1, z)$: the resulting estimate of the mean

can lie outside the natural bounds (when $g(y)$ is bounded). One should use $P(s = 1|x \in A)/P(s = 1|x \in A, z)$ if possible (which are not the included sampling weights). Unfortunately, cannot generally estimate the proper weights if x is sometimes missing.

- The HM problem is related to another issue. Suppose

$$E(y|x) = \alpha + x\beta. \tag{21}$$

Let z be a variables that are always observed and let $p(z)$ be the selection probability, as before. Suppose at least part of x is not always observed, so that x is not a subset of z. Consider the IPW estimator of α, β solves

$$\min_{a,b} \sum_{i=1}^{N} \left(\frac{s_i}{\hat{p}(z_i)} \right) (y_i - a - x_i b)^2. \tag{22}$$

The problem is that if

$$P(s = 1|x, y) = P(s = 1|x), \tag{23}$$

the IPW is generally inconsistent because the condition

$$P(s = 1|x, y, z) = P(s = 1|z) \tag{24}$$

is unlikely. On the other hand, if (23) holds, we can consistently estimate the parameters using OLS on the selected sample.

- If x is always observed, case for weighting is much stronger because then $x \subset z$. If selection is on x, this should be picked up in large

samples in the estimation of $P(s = 1|z)$.

- If (23) holds and x is always observed, is there a reason to use IPW?

Not if we believe (21) along with the homoskedasticity assumption $Var(y|x) = \sigma^2$. Then, OLS is efficient and IPW is less efficient. IPW can be more efficient with heteroskedasticity (but WLS with the correct heteroskedasticity function would be best).

- Still, one can argue for weighting under (23) as a way to consistently estimate the linear projection. Write

$$L(y|1, x) = \alpha^* + x\beta^* \tag{25}$$

where $L(\cdot|\cdot)$ denotes the linear projection. Under under $P(s = 1|x, y) = P(s = 1|x)$, the IPW estimator is consistent for θ^*. The

unweighted estimator has a probabilty limit that depends on $p(x)$.

- Parameters in LP show up in certain treatment effect estimators, and are the basis for the "double robustness" result of Robins and Ritov (1997) in the case of linear regression.
- The double robustness result holds for certain nonlinear models, but must choose model for $E(y|x)$ and the objective function appropriately; see Wooldridge (2007). (For binary or fractional response, use logistic function and Bernoulli quasi-log likelihood (QLL). For nonnegative response, use exponential function with Poisson QLL.)
- Return to the IPW regression estimator under $P(s = 1|y, z) = P(s = 1|z) = G(z, \gamma)$, with

$$E(u) = 0, E(x'u) = 0, \tag{26}$$

for a parametric function $G(\cdot)$ (such as flexible logit), and $\hat{\gamma}$ is the binary response MLE. As shown by Robins, Rotnitzky, and Zhou (1995) and Wooldridge (2007), the asymptotic variance of $\hat{\theta}_{IPW}$, using the estimated probability weights, is

$$Avar\sqrt{N}(\hat{\theta}_{IPW} - \theta) = [E(x_i'x_i)]^{-1}E(r_i r_i')[E(x_i'x_i)]^{-1}, \tag{27}$$

where r_i is the $P \times 1$ vector of population residuals from the regression $(s_i/p(z_i))x_i'u_i$ on d_i', where d_i is the $M \times 1$ score for the MLE used to obtain $\hat{\gamma}$. This is always smaller than the variance if we knew $p(z_i)$. Leads to a simple estimate of $Avar(\hat{\theta}_{IPW})$:

$$\left(\sum_{i=1}^{N}(s_i/\hat{G}_i)x_i'x_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{r}_i\hat{r}_i'\right)\left(\sum_{i=1}^{N}(s_i/\hat{G}_i)x_i'x_i\right)^{-1} \tag{28}$$

If selection is estimated by logit with regressors $h_i = h(z_i)$,

$$\hat{d}_i = h_i'(s_i - \Lambda(h_i\hat{\gamma})), \tag{29}$$

where $\Lambda(a) = \exp(a)/[1 + \exp(a)]$ and $h_i = h(z_i)$.

- Illustrates an interesting finding of RRZ (1995), related to the Hirano, Imbens, and Ritter (2003) efficient estimator for means using IPW estimators. Suppose for functions $h_{i1} = h_1(z_i)$, the logit model is correctly specified: $P(s_i = 1|z_i) = \Lambda(h_{i1}\gamma_1)$. Now take additional functions, $h_{i2} = h_2(z_i)$, and add them to the logit. Asymptotically, the

coefficients on h_{i2} are zero, so the adjustment to variance of $\hat{\theta}_{IPW}$ comes from regressing $[s_i/\Lambda(h_{i1}\gamma_1)] \cdot x_i'u_i$ on $[s_i - \Lambda(h_{i1}\gamma_1)] \cdot (h_{i1}, h_{i2})$. This reduces the residual variance relative to just using h_{i1}, so $\hat{\theta}_{IPW}$ using (h_{i1}, h_{i2}) generaly more efficient than using the "correct" functions, h_{i1}. HIR estimator keeps expanding h_i.

- Wooldridge (2007): adjustment in (27) carries over to general nonlinear models and estimation methods. Ignoring the estimation in $\hat{p}(z)$, as is standard, is asymptotically conservative. When selection is exogenous in the sense of $P(s = 1|x, y, z) = P(s = 1|x)$, the adjustment makes no difference.
- As a particular example, consider VP sampling. It one uses the

known sampling probabilities (probability of keeping an observation that falls into a given stratum), this is less efficient than using the frequencies estimated from the data. (These require knowing how many times each stratum was sampled.) When the latter are used, the adjustment is to subtract off within-stratum means in computing the sampling variation in the score:

$$\widehat{Avar(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})} = \left[\sum_{i=1}^{M} \mathbf{x}_i' \mathbf{x}_i / \hat{p}_{g_i} \right]^{-1}$$

$$\cdot \left\{ \sum_{g=1}^{G} \hat{p}_g^{-2} \left[\sum_{i=1}^{M_g} (\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})(\mathbf{x}_{gi}' \hat{u}_{gi} - \overline{\mathbf{x}_g' \hat{u}_g})' \right] \right\} \tag{30}$$

$$\cdot \left[\sum_{i=1}^{M} \mathbf{x}_i' \mathbf{x}_i / \hat{p}_{g_i} \right]^{-1}$$

absorbing the intercept into $\mathbf{x}_i$. If we drop $\overline{\mathbf{x}_g' \hat{u}_g}$ from the middle, we get the usual sandwich estimator for weighted least squares, which is larger than (30). Generally, the adjustment in (30) is the sourced of variance reduction using knowledge of stratum membership (with and without clustered data, too).

- Nevo studies the case where the population moments are $E[r(w_i, \theta)] = 0$ and the selection probability depends on elements of

w_i that are not always observed, and uses information on population means $E[h(w_i)]$ such that $P(s = 1|w) = P(s = 1|h(w))$ to obtain an expanded set of moment conditions for GMM estimation. So, if we use a logit model for selection,

$$E\left[\frac{s_i}{\Lambda(h(w_i)\gamma)} r(w_i, \theta) \right] = 0 \tag{31}$$

and

$$E\left[\frac{s_i h(w_i)}{\Lambda(h(w_i)\gamma)} \right] = \mu_h \tag{32}$$

where μ_h is known. Equation (32) generally identifies γ, and then this $\hat{\gamma}$ can be used in a second step to choose $\hat{\theta}$ in a weighted GMM

procedure.

- IPW can be used when data are missing due to a censored duration, t_i, where c_i is the censoring time. The needed probabilities turn out to be $G(t_i)$ where $G(t) \equiv \mathrm{P}(c_i \geq t)$ is the survivor function for the censoring values. This can be estimated using Kaplan-Meier estimator with roles of c_i and t_i are reversed. See Rotnitzky and Robins (2005) for a survey of how to obtain semiparametrically efficient estimators in linear regression. Holds for lots of nonlinear models, too.

Attrition in Panel Data

- Inverse probability weighting can be applied to the attrition problem in panel data. Many estimation methods can be used, but consider

MLE. We have a parametric density, $f_t(y_t|\mathbf{x}_t,\theta)$, and let s_{it} be the selection indicator. We already discussed just using pooled OLS on on the observed data:

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\log f_t(y_{it}|\mathbf{x}_{it},\boldsymbol{\theta}), \tag{33}$$

which is consistent if $P(s_{it}=1|y_{it},\mathbf{x}_{it}) = P(s_{it}=1|\mathbf{x}_{it})$. If not, maybe we can find variables $\mathbf{r}_{it}$, such that

$$P(s_{it}=1|\mathbf{w}_{it},\mathbf{r}_{it}) = P(s_{it}=1|\mathbf{r}_{it}) \equiv p_{it} > 0 \tag{34}$$

where $\mathbf{w}_{it} = (\mathbf{x}_{it},y_{it})$. The weighted MLE is

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N}\sum_{t=1}^{T}(s_{it}/p_{it})\log f_t(y_{it}|\mathbf{x}_{it},\boldsymbol{\theta}). \tag{35}$$

Under (34), $\hat{\theta}_{IPW}$ is generally consistent because

$$E[(s_{it}/p_{it})q_t(\mathbf{w}_{it},\boldsymbol{\theta})] = E[q_t(\mathbf{w}_{it},\boldsymbol{\theta})] \tag{36}$$

where $q_t(\mathbf{w}_{it},\boldsymbol{\theta}) = \log f_t(y_{it}|\mathbf{x}_{it},\boldsymbol{\theta})$.

- How do we choose $\mathbf{r}_{it}$ to make (34) hold (if possible)? RRZ (1995) propose a sequential strategy,

$$\pi_{it} = P(s_{it} = 1|z_{it}, s_{i,t-1} = 1), t = 1,\ldots,T. \tag{37}$$

Typically, $\mathbf{z}_{it}$ contains elements from $(\mathbf{w}_{i,t-1},\ldots,\mathbf{w}_{i1})$

- How do we obtain p_{it} from the π_{it}? Not without some strong

assumptions. Let $\mathbf{v}_{it} = (\mathbf{w}_{it}, \mathbf{z}_{it}), t = 1, \ldots, T.$ An ignorability assumption that works is

$$P(s_{it} = 1|\mathbf{v}_i, s_{i,t-1} = 1) = P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1). \tag{38}$$

That is, given the entire history $\mathbf{v}_i = (\mathbf{v}_{i1}, \ldots, \mathbf{v}_{iT}),$ selection at time t depends only on variables observed at $t - 1$. RRZ (1995) show how to relax it somewhat in a regression framework with time-constant covariates. Using this assumption, we can show that

$$p_{it} \equiv P(s_{it} = 1|\mathbf{v}_i) = \pi_{it}\pi_{i,t-1} \cdot \cdot \cdot \pi_{i1}. \tag{39}$$

So, a consistent two-step method is: (i) In each time period, estimate a binary response model for $P(s_{it} = 1|\mathbf{z}_{it}, s_{i,t-1} = 1)$, which means on the group still in the sample at $t - 1$. The fitted probabilities are the $\hat{\pi}_{it}$.

Form $\hat{p}_{it} = \hat{\pi}_{it}\hat{\pi}_{i,t-1} \cdot \cdot \cdot \hat{\pi}_{i1}$. (ii) Replace p_{it} with $\hat{p}_{it}$ in (35), and obtain the weighted pooled MLE.

- As shown by RRZ (1995) in the regression case, it is more efficient to estimate the p_{it} than to use know weights, if we could. See RRZ (1995) and Wooldridge (2002) for a simple regression method for adjusting the score.
- IPW for attrition suffers from a similar drawback as in the cross section case. Namely, if $P(s_{it} = 1|\mathbf{w}_{it}) = P(s_{it} = 1|\mathbf{x}_{it})$ then the unweighted estimator is consistent. If we use weights that are not a function of $\mathbf{x}_{it}$ in this case, the IPW estimator is generally inconsistent.
- Related to the previous point: would rarely apply IPW in the case of a

model with completely specified dynamics. Why? If we have a model for $D(y_{it}|x_{it}, y_{i,t-1}, \ldots, x_{i1}, y_{i0})$ or $E(y_{it}|x_{it}, y_{i,t-1}, \ldots, x_{i1}, y_{i0})$, then our variables affecting attrition, z_{it}, are likely to be functions of $(y_{i,t-1}, \ldots, x_{i1}, y_{i0})$. If they are, the unweighted estimator is consistent. For misspecified models, we might still want to weight.

3. Imputation

- So far, we have discussed when we can just drop missing observations (Section 1) or when the complete cases can be used in a weighting method (Section 2). A different approach to missing data is to try to fill in the missing values, and then analyze the resulting data set as a complete data set. Little and Rubin (2002) provide an

accessible treatment to *imputation* and *multiple imputation* methods, with lots of references to work by Rubin and coauthors.

- Imputing missing values cannot always be valid, of course. Most methods depend on a *missing at random* (MAR) assumption. When data are missing on only one variable – say, the response variable, y – MAR is essentially the same as $P(s = 1|y, x) = P(s = 1|x)$. The assumption *missing completely at random* (MCAR) is when s is independent of $w = (x, y)$.
- MAR can be defined for general missing data patterns. Let $w_i = (w_{i1}, w_{i2})$ be a random draw from the population, where data can be missing on either variable. Let $r_i = (r_{i1}, r_{i2})$ be the "retention"

indicators for w_{i1} and w_{i2}, so $r_{ig} = 1$ implies w_{ig} is observed. The MCAR assumption is that r_i is independent of w_i. The MAR assumption is that $P(r_{i1} = 0, r_{i2} = 0|w_i) = P(r_{i1} = 0, r_{i2} = 0) \equiv \pi_{00}$, $P(r_{i1} = 1, r_{i2} = 0|w_{i1})$, and $P(r_{i1} = 0, r_{i2} = 1|w_{i2})$. Even with just two variables, the restrictions imposed by MAR are not especially appealing, unless, of course, we have good reason to just assume MCAR.

- MAR is more natural with monotone missing data problems; we just saw the case of attrition. If we order the variables so that if w_{ih} is observed the so is w_{ig}, $g < h$. Write $f(w_1, \ldots, w_G) = f(w_G|w_{G-1}, \ldots, w_1) \cdot f(w_{G-1}|w_{G-1}, \ldots, w_1) \cdots f(w_2|w_1)f(w_1)$. Given parametric models, we

can write partial log likelihood as

$$\sum_{g=1}^{G} r_{ig} \log f(w_{ig}|w_{i,g-1},\ldots,w_{i1},\theta), \tag{40}$$

where we use $r_{ig} = r_{ig}r_{i,g-1}\cdots r_{i2}$. Under MAR,

$$E(r_{ig}|w_{ig},\ldots,w_{i1}) = E(r_{ig}|w_{i,g-1},\ldots,w_{i1}). \tag{41}$$

As we showed in the attrition case, partial MLE based on (40) is consistent and $\sqrt{N}$-asymptotically normal in general. This is the basis for filling in data in monotonic MAR schemes.

- Simple example of imputation. Let $\mu_y = E(y)$, but data are missing on some y_i. Unless $P(s_i = 1|y_i) = P(s_i = 1)$, the complete-case average is not consistent for μ_y. Suppose that the selection is ignorable

conditional on $\mathbf{x}$:

$$E(y|\mathbf{x}, s) = E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}), \tag{42}$$

where $m(\mathbf{x}, \boldsymbol{\beta})$ is a parametric function. From Section 1, NLS using the selected sample is consistent for $\boldsymbol{\beta}$. Because we observe $\mathbf{x}_i$ for all i, we can obtained fitted values, $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, for any unit it the sample. Let $\hat{y}_i = s_i y_i + (1 - s_i) m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ be the imputed data. Then an imputation estimator of μ_y is

$$\hat{\mu}_y = N^{-1} \sum_{i=1}^{N} \{s_i y_i + (1 - s_i) m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})\}. \tag{43}$$

From $\text{plim}(\hat{\mu}_y) = E[s_i y_i + (1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})]$ we can show consistency of $\hat{\mu}_y$ because, by (42) and iterated expectations,

$$E[s_i y_i + (1 - s_i) m(\mathbf{x}_i, \boldsymbol{\beta})] \;\; = E[m(\mathbf{x}_i, \boldsymbol{\beta})] \;\; = \mu_y. \tag{44}$$

- Danger in using imputation methods: we might be tempted to treat the imputed data as real random draws.

Generally leads to incorrect inference because of inconsistent variance estimation. (In linear regression, easy to see that estimated variance is too small.)

- Little and Rubin (2002) call (43) the method of "conditional means." In their Table Table 4.1 the document the downward bias in variance estiimates.

- Instead, LR propose adding a random draw to $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ to impute a value – assuming, of course, that we can estimate $D(y|\mathbf{x})$. If we assume

that $D(u_i|\mathbf{x}_i) = Normal(0, \sigma_u^2)$, draw $\check{u}_i$ from a Normal$(0, \hat{\sigma}_u^2)$, distribution, where $\hat{\sigma}_u^2$ is estimated using the complete case nonlinear regression residuals, and then use $m(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) + \check{u}_i$ for the missing data. Called the "conditional draw" method of imputation, which is a special case of stochastic imputation.

- Generally difficult to quantity the uncertainty from single-imputation methods, where one imputed values is obtained for each missing variable. One possibility is to bootstrap the entire estimation/imputation steps. Can be computationally intensive because imputation needs to be done for each bootstrap sample.
- Multiple imputation is an alternative. Its theoretical justification is

Bayesian, based on obtaining the posterior distribution – in particular, mean and variance – of the parameters conditional on the observed data. For general missing data patterns, the computation required to impute missing values is quite complicated, and involves simulation methods of estimation. LR and Cameron and Trivedi (2005) provide discussion.

- General idea: rather than just impute one set of missing values to create one "complete" data set, create several imputed data sets. (Often the number is fairly small, such as five or so.) Estimate the parameters of interest using each imputed data set, and then use an averaging to obtain a final parameter estimate and sampling error.

Let $\mathbf{W}_{mis}$ denote the matrix of missing data and $\mathbf{W}_{obs}$ the matrix of observations. Assume that MAR holds. MAR used to estimate $E(\boldsymbol{\theta}|\mathbf{W}_{obs})$, the posterier mean of $\boldsymbol{\theta}$ given $\mathbf{W}_{obs}$. But by iterated expectations,

$$E(\boldsymbol{\theta}|\mathbf{W}_{obs}) = E[E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}]. \tag{45}$$

If $\hat{\boldsymbol{\theta}}_d = E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis}^{(d)})$ for imputed data set d, then approximate $E(\boldsymbol{\theta}|\mathbf{W}_{obs})$ as

$$\bar{\boldsymbol{\theta}} = D^{-1} \sum_{d=1}^{D} \hat{\boldsymbol{\theta}}_d, \tag{46}$$

Further, we can obtain a “sampling” variance by estimating $Var(\theta|\mathbf{W}_{obs})$ using

$$Var(\theta|\mathbf{W}_{obs}) = E[Var(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}] + Var[E(\boldsymbol{\theta}|\mathbf{W}_{obs}, \mathbf{W}_{mis})|\mathbf{W}_{obs}], \tag{47}$$

which suggests

$$\begin{aligned}\widehat{Var}(\theta|\mathbf{W}_{obs}) &= D^{-1}\sum_{d=1}^{D}\hat{\mathbf{V}}_d \\ &\quad + (D-1)^{-1}\sum_{d=1}^{D}(\hat{\boldsymbol{\theta}}_d - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_d - \bar{\boldsymbol{\theta}})' \\ &\equiv \bar{\mathbf{V}} + \mathbf{B},\end{aligned} \tag{48}$$

where $\bar{\mathbf{V}}$ is the average of the variance estimates across imputed samples and $\mathbf{B}$ is the between-imputation variance. For small number of imputations, a correction is usually made, namely, $\bar{\mathbf{V}} + (1 + D)^{-1}\mathbf{B}$.

assuming that one trusts the MAR assumption and the underlying distributions used to draw the imputed values, inference with multiple imputations is fairly straightforward. D need not be very large so estimation using nonlinear models is relatively easy, given the imputed data.

- Like weighting methods, imputation methods shortcomings when applied to estimation of models with missing conditioning variables. If $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, we are interested in $D(y|\mathbf{x})$, data are missing on y and $\mathbf{x}_2$ – say, for the same units – and selection is a function of $\mathbf{x}_2$. Using the complete cases will be consistent. Imputation methods would not be, as they require $D(s|y, \mathbf{x}_1, \mathbf{x}_2) = D(s|\mathbf{x}_1)$.

4. Heckman-Type Selection Corrections

● The lecture notes discuss advantages of applying IV methods when data are missing on explanatory variables in addition to the response variable. Briefly, a variable that is exogenous in the population model need not be in the selected subpopulation. (Example: wage-benefits tradeoff.)

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1 \tag{49}$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2 \tag{50}$$

$$y_3 = 1[\mathbf{z}\boldsymbol{\delta}_3 + v_3 > 0]. \tag{51}$$

Assume (a) $(\mathbf{z}, y_3)$ is always observed, (y_1, y_2) observed when $y_3 = 1$; (b) $E(u_1|\mathbf{z}, v_3) = \gamma_1 v_3$; (c) $v_3|\mathbf{z}$ ~Normal$(0,1)$; (d) $E(\mathbf{z}'v_2) = \mathbf{0}$ and

$\boldsymbol{\delta}_{22} \neq \mathbf{0}$, then we can write

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + g(\mathbf{z}, y_3) + e_1 \tag{52}$$

where $e_1 = u_1 - g(\mathbf{z}, y_3) = u_1 - E(u_1|\mathbf{z}, y_3)$. Selection is exogenous in (52) because $E(e_1|\mathbf{z}, y_3) = 0$. Because y_2 is not exogenous, we estimate (52) by IV, using the selected sample, with IVs $(\mathbf{z}, \lambda(\mathbf{z}\boldsymbol{\delta}_3))$ because $g(\mathbf{z}, 1) = \lambda(\mathbf{z}\boldsymbol{\delta}_3)$. The two-step estimator is (i) Probit of y_3 on $\mathbf{z}$ to (using all observations) to get $\hat{\lambda}_{i3} \equiv \lambda(\mathbf{z}_i\hat{\boldsymbol{\delta}}_3)$; (ii) IV (2SLS if overidentifying restrictions) of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{\lambda}_{i3}$ using instruments $(\mathbf{z}_i, \hat{\lambda}_{i3})$.

- If y_2 is always observed, tempting to obtain the fitted values $\hat{y}_{i2}$ from the reduced form y_{i2} on $\mathbf{z}_i$, and then use OLS of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}, \hat{\lambda}_{i3}$ in the

second stage. But this effectively puts $\alpha_1 v_2$ in the error term, so we would need $u_1 + \alpha_2 v_2$ to be normally (or something similar). Rules out discrete y_2. The procedure just outlined uses the linear projection $y_2 = \mathbf{z}\boldsymbol{\pi}_2 + \eta_2 \lambda(\mathbf{z}\boldsymbol{\delta}_3) + r_3$ in the selected population, and does not care whether this is a conditional expectation.

- Should have at least two elements in $\mathbf{z}$ not in $\mathbf{z}_1$: one to exogenously vary y_2, one to exogenously vary selection, y_3.
- If an explanatory variable is not always observed, ideally can find an IV for it and treat it as endogenous even if it is exogenous in the population. Generally, the usual Heckman approach (like IPW and imputation) is hard to justify in the model $E(y|\mathbf{x}) = E(y|\mathbf{x}_1)$ if $\mathbf{x}_1$ is not

always observed. The first-step would be estimation of $P(s = 1\mathbf{x}_2)$ where $\mathbf{x}_2$ is always observed. But then we would be assuming $P(s = 1|\mathbf{x}) = P(s = 1|\mathbf{x}_2)$, effectively an exclusion restriction on a reduced form.

"What's New in Econometrics"

Lecture 13

Weak Instruments and Many Instruments

Guido Imbens

NBER Summer Institute, 2007

Outline

1. Introduction

2. Motivation

3. Weak Instruments

4. Many (Weak) Instruments

1. Introduction

Standard normal asymptotic approximation to sampling distribution of IV, TSLS, and LIML estimators relies on non-zero correlation between instruments and endogenous regressors.

If correlation is close to zero, these approximations are not accurate even in fairly large samples.

In the just identified case TSLS/LIML confidence intervals will still be fairly wide in most cases, even if not valid, unless degree of endogeneity is very high. If concerned with this, alternative confidence intervals are available that are valid uniformly. No better estimators available.

In the case with large degree of overidentification TSLS has poor properties: considerable bias towards OLS, and substantial underestimation of standard errors.

LIML is much better in terms of bias, but its standard error is not correct. A simple multiplicative adjustment to conventional LIML standard errors based on Bekker asymptotics or random effects likelihood works well.

Overall: use LIML, with Bekker-adjusted standard errors.

2.A Motivation : Angrist-Krueger

AK were interested in estimating the returns to years of education. Their basic specification is:

$$Y_i = \alpha + \beta \cdot E_i + \varepsilon_i,$$

where Y_i is log (yearly) earnings and E_i is years of education.

In an attempt to address the endogeneity problem AK exploit variation in schooling levels that arise from differential impacts of compulsory schooling laws by quarter of birth and use quarter of birth as an instrument. This leads to IV estimate (using only 1st and 4th quarter data):

$$\widehat{\beta} = \frac{\overline{Y}_4 - \overline{Y}_1}{\overline{E}_4 - \overline{E}_1} = 0.089 \quad (0.011)$$

2.B AK with Many Instruments

AK also present estimates based on additional instruments. They take the basic 3 qob dummies and interact them with 50 state and 9 year of birth dummies.

Here (following Chamberlain and Imbens) we interact the single binary instrument with state times year of birth dummies to get 500 instruments. Also including the state times year of birth dummies as exogenous covariates leads to the following model:

$$Y_i = X_i'\beta + \varepsilon_i, \quad \mathbb{E}[Z_i \cdot \varepsilon_i] = 0,$$

where X_i is the 501-dimensional vector with the 500 state/year dummies and years of education, and Z_i is the vector with 500 state/year dummies and the 500 state/year dummies multiplying the indicator for the fourth quarter of birth.

The TSLS estimator for β is

$$\widehat{\beta}_{\text{TSLS}} = 0.073 \quad (0.008)$$

suggesting the extra instruments improve the standard errors a little bit.

However, LIML estimator tells a somewhat different story,

$$\widehat{\beta}_{\text{LIML}} = 0.095 \quad (0.017)$$

with an increase in the standard error.

1.C Bound-Jaeger-Baker Critique

BJB suggest that despite the large (census) samples used by AK asymptotic normal approximations may be very poor because the instruments are only very weakly correlated with the endogenous regressor.

The most striking evidence for this is based on the following calculation. Take the AK data and re-calculate their estimates after replacing the actual quarter of birth dummies by random indicators with the same marginal distribution.

In principle this means that the standard (gaussian) large sample approximations for TSLS and LIML are invalid since they rely on non-zero correlations between the instruments and the endogenous regressor.

	Single Instr		500 Instruments			
			TSLS		LIML	
Real QOB	0.089	(0.011)	0.073	(0.008)	0.095	(0.017)
Random QOB	-1.96	(18.12)	0.059	(0.009)	-0.330	(0.100)

With many random instruments the results are troubling. Although the instrument contains no information, the results suggest that the instruments can be used to infer precisely what the returns to education are.

1.D Simulations with a Single Instrument

10,000 artificial data sets, all of size 160,000, designed to mimic the AK data. In each of these data sets half the units have quarter of birth (denoted by Q_i) equal to 0 and 1 respectively.

$$\begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.446 & \rho \cdot \sqrt{0.446} \cdot \sqrt{10.071} \\ \rho \cdot \sqrt{0.446} \cdot \sqrt{10.071} & 10.071 \end{pmatrix}\right).$$

The correlation between the reduced form residuals in the AK data is $\rho = 0.318$.

$$E_i = 12.688 + 0.151 \cdot Q_i + \eta_i,$$

$$Y_i = 5.892 + 0.014 \cdot Q_i + \nu_i.$$

Now we calculate the IV estimator and its standard error, using either the actual qob variable or a random qob variable as the instrument.

We are interested in the size of tests of the null that coefficient on years of education is equal to $0.089 = 0.014/0.151$.

We base the test on the t-statistic. Thus we reject the null if the ratio of the point estimate minus 0.089 and the standard error is greater than 1.96 in absolute value.

We repeat this for 12 different values of the reduced form error correlation. In Table 3 we report the coverage rate and the median and 0.10 quantile of the width of the estimated 95% confidence intervals.

Table 3: Coverage Rates of Conv. TSLS CI by Degree of Endogeneity

ρ	0.0	0.4	0.6	0.8	0.9	0.95	0.99
implied OLS	0.00	0.08	0.13	0.17	0.19	0.20	0.21
Real QOB							
Cov rate	0.95	0.95	0.96	0.95	0.95	0.95	0.95
Med Width 95% CI	0.09	0.08	0.07	0.06	0.05	0.05	0.05
0.10 quant Width	0.08	0.07	0.06	0.05	0.04	0.04	0.04
Random QOB							
Cov rate	0.99	1.00	1.00	0.98	0.92	0.82	0.53
Med Width 95% CI	1.82	1.66	1.45	1.09	0.79	0.57	0.26
0.10 quant Width	0.55	0.51	0.42	0.33	0.24	0.17	0.08

In this example, unless the reduced form correlations are very high, e.g., at least 0.95, with irrelevant instruments the conventional confidence intervals are wide and have good coverage.

The amount of endogeneity that would be required for the conventional confidence intervals to be misleading is higher than one typically encounters in cross-section settings.

Put differently, although formally conventional confidence intervals are not valid uniformly over the parameter space (e.g., Dufour, 1997), the subsets of the parameter space where results are substantively misleading may be of limited interest.

This in contrast to the case with many weak instruments where especially TSLS can be misleading in empirically relevant settings.

3.A Single Weak Instrument

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i,$$

$$X_i = \pi_0 + \pi_1 \cdot Z_i + \eta_i,$$

with $(\varepsilon_i, \eta_i) \perp\!\!\!\perp Z_i$, and jointly normal with covariance matrix Σ. The reduced form for the first equation is

$$Y_i = \alpha_0 + \alpha_1 \cdot Z_i + \nu_i,$$

where the parameter of interest is $\beta_1 = \alpha_1/\pi_1$. Let

$$\Omega = \mathbb{E}\left[\begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix} \cdot \begin{pmatrix} \nu_i \\ \eta_i \end{pmatrix}'\right], \quad \text{and} \quad \Sigma = \mathbb{E}\left[\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \cdot \begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix}'\right],$$

Standard IV estimator:

$$\widehat{\beta}_1^{\text{IV}} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)\left(Z_i - \overline{Z}\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Z_i - \overline{Z}\right)},$$

Concentration parameter:

$$\lambda = \pi_1^2 \cdot \sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)^2 / \sigma_\eta^2.$$

Normal approximations for numerator and denominator are accurate:

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)\left(Z_i-\overline{Z}\right)-\mathsf{Cov}(Y_i,Z_i)\right)\approx\mathcal{N}\left(0,V(Y_i\cdot Z_i)\right),$$

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\left(X_i-\overline{X}\right)\left(Z_i-\overline{Z}\right)-\mathsf{Cov}(X_i,Z_I)\right)\approx\mathcal{N}\left(0,V(X_i\cdot Z_i)\right).$$

If $\pi_1 \neq 0$, as the sample size gets large, then the ratio will eventually be well approximated by a normal distribution as well.

However, if $\mathsf{Cov}(X_i, Z_i) \approx 0$, the ratio may be better approximated by a Cauchy distribution, as the ratio of two normals centered close to zero.

3.B Staiger-Stock Asymptotics and Uniformity

Staiger and Stock investigate the distribution of the standard IV estimator under an alternative asymptotic approximation.

The standard asymptotics (strong instrument asymptotics in the SS terminology) is based on fixed parameters and the sample size getting large.

In their alternative asymptotic sequence SS model π_1 as a function of the sample size, $\pi_{1N} = c/\sqrt{N}$, so that the concentration parameter converges to a constant:

$$\lambda \longrightarrow c^2 \cdot V(Z_i).$$

SS then compare coverage properties of various confidence intervals under this (weak instrument) asymptotic sequence.

The importance of the SS approach is in demonstrating for any sample size there are values of the nuisance parameters such that the actual coverage is substantially away from the nominal coverage.

More recently the issue has therefore been reformulated as requiring confidence intervals to have asymptotically the correct coverage probabilities uniformly in the parameter space. See for a discussion from this perspective Mikusheva.

Note that there cannot exist estimators that are consistent for β^* uniformly in the parameter space since if $\pi_1 = 0$, there are no consistent estimators for β_1. However, for testing there are generally confidence intervals that are uniformly valid, but they are not of the conventional form, that is, a point estimate plus or minus a constant times a standard error.

3.C Anderson-Rubin Confidence Intervals

Let the instrument $\tilde{Z}_i = Z_i - \overline{Z}$ be measured in deviations from its mean. Then define the statistic

$$S(\beta_1) = \frac{1}{N} \sum_{i=1}^{N} \tilde{Z}_i \cdot (Y_i - \beta_1 \cdot X_i) \,.$$

Then, under the null hypothesis that $\beta_1 = \beta_1^*$, and conditional on the instruments, the statistic $\sqrt{N} \cdot S(\beta_1^*)$ has an exact normal distribution

$$\sqrt{N} \cdot S(\beta_1^*) \sim \mathcal{N}\left(0, \sum_{i=1}^{N} \tilde{Z}_i^2 \cdot \sigma_\varepsilon^2\right).$$

Anderson and Rubin (1949) propose basing tests for the null hypothesis

$$H_0: \quad \beta_1 = \beta_1^0, \qquad \text{against the alternative hypothesis} \quad H_a: \quad \beta_1 \neq \beta_1^0$$

on this idea, through the statistic

$$\text{AR}\left(\beta_1^0\right) = \frac{N \cdot S(\beta_1^0)^2}{\sum_{i=1}^{N} \tilde{Z}_i^2} \cdot \left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix} \right)^{-1}.$$

A confidence interval can be based on this test statistic by inverting it:

$$\text{CI}_{0.95}^{\beta_1} = \{\beta_1 \,|\text{AR}(\beta_1) \leq 3.84\}$$

This interval can be equal to the whole real line.

3.D Anderson-Rubin with K instruments

The reduced form is

$$X_i = \pi_0 + \pi_1' Z_i + \eta_i,$$

$S\left(\beta_1^0\right)$ is now normally distributed vector.
AR statistic with associated confidence interval:

$$\mathsf{AR}\left(\beta_1^0\right) = N \cdot S\left(\beta_1^0\right)' \left(\sum_{i=1}^{N} \tilde{Z}_i \cdot \tilde{Z}_i'\right)^{-1} S\left(\beta_1^0\right) \cdot \left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix} \Omega \begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix}\right)$$

$$\mathsf{CI}_{0.95}^{\beta_1} = \left\{\beta_1 \,\middle|\, \mathsf{AR}(\beta_1) \leq \mathcal{X}_{0.95}^2(K)\right\},$$

The problem is that this confidence interval can be empty because it simultaneously tests validity of instruments.

3.E Kleibergen Test

Kleibergen modfies AR statistic through

$$\tilde{S}\left(\beta_1^0\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Z}_i'\hat{\pi}_1(\beta_1^0)\right)\cdot\left(Y_i - \beta_1^0\cdot X_i\right),$$

where $\hat{\pi}$ is the maximum likelihood estimator for π_1 under the restriction $\beta_1 = \beta_1^0$. The test is then based on the statistic

$$K\left(\beta_1^0\right) = \frac{N\cdot\tilde{S}(\beta_1^0)^2}{\sum_{i=1}^{N}\tilde{Z}_i^2}\cdot\left(\begin{pmatrix} 1 & -\beta_1^0 \end{pmatrix}\Omega\begin{pmatrix} 1 \\ -\beta_1^0 \end{pmatrix}\right)^{-1}.$$

This has an approximate chi-squared distribution, and can be used to construct a confidence interval.

3.F Moreira's Similar Tests

Moreira (2003) proposes a method for adjusting the critical values that applies to a number of tests, including the Kleibergen test. His idea is to focus on *similar* tests, test that have the same rejection probability for all values of the nuisance parameter (the π) by adjusting critical values (instead of using quantiles from the chi-squared distribution).

The way to adjust the critical values is to consider the distribution of a statistic such as the Kleibergen statistic conditional on a complete sufficient statistic for the nuisance parameter. In this setting a complete sufficient statistic is readily available in the form of the maximum likelihood estimator under the null, $\widehat{\pi}_1(\beta_1^0)$.

Moreira's preferred test is based on the likelihood ratio. Let

$$LR\left(\beta_1^0\right) = 2 \cdot \left(L\left(\widehat{\beta}_1, \widehat{\pi}\right) - L\left(\beta_1^0, \widehat{\pi}(\beta_1^0)\right)\right),$$

be the likelihood ratio.

Then let $c_{LR}(p, 0.95)$, be the 0.95 quantile of the distribution of $LR(\beta_1^0)$ under the null hypothesis, conditional on $\widehat{\pi}(\beta_1^0) = p$. The proposed test is to reject the null hypothesis at the 5% level if

$$LR\left(\beta_1^0\right) > c_{LR}(\widehat{\pi}(\beta_1^0), 0.95),$$

where conventional test would use critical values from a chi-squared distribution with a single degree of freedom. The critical values are tabulated for low values of K.

This test can then be converted to construct a 95% confidence intervals.

3.G Conditioning on the First Stage

These confidence intervals are asymptotically valid irrespective of the strength of the first stage (the value of π_1). However, they are not valid if one first inspects the first stage, and conditional on the strength of that, decides to proceed.

Specifically, if in practice one first inspects the first stage, and decide to abandon the project if the first stage F-statistic is less than some fixed value, and otherwise proceed by calculating confidence interval, the large sample coverage probabilities would not be the nominal ones.

Chioda and Jansson propose a confidence interval that is valid conditional on the strength of the first stage. A caveat is that this involves loss of information, and thus the Chioda-Jansson confidence intervals are wider than confidence intervals that are not valid conditional on the first stage.

4.A Many (Weak) Instruments

In this section we discuss the case with many weak instruments. The problem is both the bias in the standard estimators, and the misleadingly small standard errors based on conventional procedures, leading to poor coverage rates for standard confidence intervals in many situations.

Resampling methods such as bootstrapping do not solve these problems.

The literature has taken a number of approaches. Part of the literature has focused on alternative confidence intervals analogues to the single instrument case. In addition a variety of new point estimators have been proposed.

Generally LIML still does well, but standard errors need to be adjusted.

4.B Bekker Asymptotics

Bekker (1995) derives large sample approximations for TSLS and LIML based on sequences where the number of instruments increases proportionally to the sample size.

He shows that TSLS is not consistent in that case.

LIML is consistent, but the conventional LIML standard errors are not valid. Bekker then provides LIML standard errors that are valid under this asymptotic sequence. Even with relatively small numbers of instruments the differences between the Bekker and conventional asymptotics can be substantial.

For a simple case the adjustment to the variance is multiplicative.

Then one can simply multiply the standard LIML variance by

$$1 + \frac{K/N}{1 - K/N} \cdot \left(\sum_{i=1} \left(\pi_1' \tilde{Z}_i \right)^2 / N \right)^{-1} \cdot \left(\begin{pmatrix} 1 \\ \beta_1 \end{pmatrix}' \Omega^{-1} \begin{pmatrix} 1 \\ \beta_1 \end{pmatrix} \right)^{-1}$$

Recommended in practice

One can see from this expression why the adjustment can be substantial even if K is small. The second factor can be large if the instruments are weak, and the third factor can be large if the degree of endogeneity is high. If the instruments are strong, then $\sum_{i=1}(\pi_1' \tilde{Z}_i)^2/K$ will diverge, and the adjustment factor will converge to one.

4.C Random Effects Estimators

Chamberlain and Imbens propose a random effects quasi maximum likelihood (REQML) estimator. They propose modelling the first stage coefficients π_k, for $k = 1, \ldots, K$, in the regression

$$X_i = \pi_0 + \pi_1' Z_i + \eta_i = \pi_0 + \sum_{k=1}^{K} \pi_k \cdot Z_{ik} + \eta_i,$$

(after normalizing the instruments to have mean zero and unit variance,) as independent draws from a normal $\mathcal{N}(\mu_\pi, \sigma_\pi^2)$ distribution.

Assuming also joint normality for (ε_i, η_i), one can derive the likelihood function

$$\mathcal{L}(\beta_0, \beta_1, \pi_0, \mu_\pi, \sigma_\pi^2, \Omega).$$

In contrast to the likelihood function in terms of the original parameters $(\beta_0, \beta_1, \pi_0, \pi_1, \Omega)$, this likelihood function depends on a small set of parameters, and a quadratic approximation to its logarithms is more likely to be accurate.

CI discuss some connections between the REQML estimator and LIML and TSLS in the context of this parametric set up. First they show that in large samples, with a large number of instruments, the TSLS estimator corresponds to the restricted maximum likelihood estimator where the variance of the first stage coefficients is fixed at a large number, or $\sigma_\pi^2 = \infty$:

$$\widehat{\beta}_{\text{TSLS}} \approx \arg \max_{\beta_0,\beta_1,\pi_0,\mu_\pi} = L(\beta_0, \beta_1, \pi_0, \mu_\pi, \sigma_\pi^2 = \infty, \Omega).$$

From a Bayesian perspective, TSLS corresponds approximately to the posterior mode given a flat prior on all the parameters, and thus puts a large amount of prior mass on values of the parameter space where the instruments are jointly powerful.

In the special case where we fix $\mu_\pi = 0$, and Ω is known, and the random effects specification applies to all instruments, CI show that the REQML estimator is identical to LIML.

However, like the Bekker asymptotics, the REQML calculations suggests that the standard LIML variance is too small: the variance of the REQML estimator is approximately equal to the standard LIML variance times

$$1 + \sigma_\pi^{-2} \cdot \left(\begin{pmatrix} 1 \\ \beta_1 \end{pmatrix}' \Omega^{-1} \begin{pmatrix} 1 \\ \beta_1 \end{pmatrix} \right)^{-1}.$$

This is similar to the Bekker adjustment if we replace σ_π^2 by $\sum_{i=1}(\pi_1' \tilde{Z}_i)^2 (K \cdot N)$ (keeping in mind that the instruments have been normalized to have unit variance).

In practice the CI adjustment will be bigger than the Bekker adjustment because the ml estimator for σ_π^2 will take into account noise in the estimates of the $\widehat{\pi}$, and so $\widehat{\sigma}_\pi^2 < \sum_{i=1}(\widehat{\pi}_1' \tilde{Z}_i)^2 (K \cdot N)$.

4.D Choosing the Number of Instruments

Donald and Newey (2001) consider the problem of choosing a subset of an infinite sequence of instruments.

They assume the instruments are ordered, so that the choice is the number of instruments to use.

The criterion they focus on is based on an estimable approximation to the expected squared error. version of this leads to approximately the same expected squared error as using the infeasible criterion.

Although in its current form not straightforward to implement, this is a very promising approach that can apply to many related problems such as generalized method of moments settings with many moments.

4.E Flores' Simulations

In one of the more extensive simulation studies Flores-Lagunes reports results comparing TSLS, LIML, Fuller, Bias corrected versions of TSLS, LIML and Fuller, a Jacknife version of TSLS (Hahn, Hausman and Kuersteiner), and the REQML estimator, in settings with 100 and 500 observations, and 5 and 30 instruments for the single endogenous variable. Does not include LIML with Bekker standard errors.

He looks at median bias, median absolute error, inter decile range, coverage rates.

He concludes that "our evidence indicates that the random-effects quasi-maximum likelihood estimator outperforms alternative estimators in terms of median point estimates and coverage rates."

What's New in Econometrics?
Lecture 14
Quantile Methods

Jeff Wooldridge
NBER Summer Institute, 2007

1. Reminders About Means, Medians, and Quantiles
2. Some Useful Asymptotic Results
3. Quantile Regression with Endogenous Explanatory Variables
4. Quantile Regression for Panel Data

5. Quantile Methods for “Censored” Data

1. Reminders About Means, Medians, and Quantiles

- Consider the standard linear model in a population, with intercept α and $K \times 1$ slopes $\boldsymbol{\beta}$:

$$y = \alpha + \mathbf{x}\boldsymbol{\beta} + u. \tag{1}$$

Assume $E(u^2) < \infty$, so that the distribution of u is not too spread out. Given a large random sample, when should we expect ordinary least squares, which solves

$$\min_{a,\mathbf{b}} \sum_{i=1}^{N} (y_i - a - \mathbf{x}_i\mathbf{b})^2, \tag{2}$$

and least absolute deviations (LAD), which solves

$$\min_{a,\mathbf{b}} \sum_{i=1}^{N} |y_i - a - \mathbf{x}_i\mathbf{b}|, \tag{3}$$

to provide similar parameter estimates? There are two important cases. If

$$D(u|\mathbf{x}) \text{ is symmetric about zero} \tag{4}$$

then OLS and LAD both consistently estimate α and $\boldsymbol{\beta}$. If

$$u \text{ is independent of } \mathbf{x} \text{ with } E(u) = 0, \tag{5}$$

where $E(u) = 0$ is the normalization that identifies α, then OLS and LAD both consistently estimate the slopes, $\boldsymbol{\beta}$. If u has an asymmetric distribution, then $Med(u) \equiv \eta \neq 0$, and $\hat{\alpha}_{LAD}$ converges to $\alpha + \eta$ because $Med(y|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta} + Med(u|\mathbf{x}) = \alpha + \mathbf{x}\boldsymbol{\beta} + \eta$.

- In many applications, neither (4) nor (5) is likely to be true. For example, y may be a measure of wealth, in which case the error distribution is probably asymmetric and $Var(u|\mathbf{x})$ not constant.
- Therefore, it is important to remember that if $D(u|\mathbf{x})$ is asymmetric and changes with $\mathbf{x}$, then we should not expect OLS and LAD to deliver similar estimates of $\boldsymbol{\beta}$, even for "thin-tailed" distributions. It is important to separate discussions of resiliency to outliers from the

different quantities identified by least squares ($E(y|\mathbf{x})$) and least absolution deviations ($Med(y|\mathbf{x})$).

- Of course, LAD is much more resilient to changes in extreme values because, as a measure of central tendency, the median is much less sensitive than the mean to changes in extreme values. But it does not follow that a large difference in OLS and LAD estimates means something is "wrong" with OLS.
- Big advantage for median over mean: the median passes through monotonic functions. For example, if $\log(y) = \alpha + \mathbf{x}\boldsymbol{\beta} + u$ and $Med(u|\mathbf{x}) = 0$, then $\text{Med}(y|\mathbf{x}) = \exp(\text{Med}[\log(y)|\mathbf{x}]) = \exp(\alpha + \mathbf{x}\boldsymbol{\beta})$. By contrast, we cannot generally find $\text{E}(y|\mathbf{x}) = \exp(\alpha + \mathbf{x}\boldsymbol{\beta})\text{E}[\exp(u)|\mathbf{x}]$.

- But the expectation operator has useful properties that the median does not: linearity and the law of iterated expectations. Suppose we begin with a random coefficient model

$$y_i = a_i + \mathbf{x}_i \mathbf{b}_i, \tag{6}$$

If $(a_i, \mathbf{b}_i)$ is independent of $\mathbf{x}_i$, then

$$\mathrm{E}(y_i|\mathbf{x}_i) = \mathrm{E}(a_i|\mathbf{x}_i) + \mathbf{x}_i \mathrm{E}(\mathbf{b}_i|\mathbf{x}_i) \equiv \alpha + \mathbf{x}_i \boldsymbol{\beta}, \tag{7}$$

where $\alpha = \mathrm{E}(a_i)$ and $\boldsymbol{\beta} = \mathrm{E}(\mathbf{b}_i)$. So OLS consistently estimates α and $\boldsymbol{\beta}$. By contrast, no way to derive $\mathrm{Med}(y_i|\mathbf{x}_i)$ without imposing more restrictions.

- What can we add so that LAD estimates something of interest in (7)? If $\mathbf{u}_i$ is a vector, then its distribution conditional on $\mathbf{x}_i$ is centrally

symmetric if $D(\mathbf{u}_i|\mathbf{x}_i) = D(-\mathbf{u}_i|\mathbf{x}_i)$, which implies that, if $\mathbf{g}_i$ is any vector function of $\mathbf{x}_i$, $D(\mathbf{g}_i'\mathbf{u}_i|\mathbf{x}_i)$ has a univariate distribution that is symmetric about zero. This implies $E(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{0}$.

- Apply central symmetry to random coefficient model by writing $\mathbf{c}_i = (a_i, \mathbf{b}_i)$ with $\boldsymbol{\gamma} = E(\mathbf{c}_i)$, and let $\mathbf{d}_i = \mathbf{c}_i - \boldsymbol{\gamma}$. Then

$$y_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + (a_i - \alpha) + \mathbf{x}_i(\mathbf{b}_i - \boldsymbol{\beta}) \tag{8}$$

with $\mathbf{g}_i = (1, \mathbf{x}_i)$. If $\mathbf{c}_i$ given $\mathbf{x}_i$ is centrally symmetric about $\boldsymbol{\gamma}$, then $Med(\mathbf{g}_i'(\mathbf{c}_i - \boldsymbol{\gamma})|\mathbf{x}_i) = 0$, and LAD applied to the usual model $y_i = \alpha + \mathbf{x}_i\boldsymbol{\beta} + u_i$ consistently estimates α and $\boldsymbol{\beta}$.

- For $0 < \tau < 1$, $q(\tau)$ is the τ^{th} quantile of y_i if $P(y_i \leq q(\tau)) \geq \tau$ and $P(y_i \geq q(\tau)) \geq 1 - \tau$.

- Usually, we are interested in how covariates affect quantiles (of which the median is the special case with $\tau = 1/2$). Under linearity,

$$\text{Quant}_\tau(y_i|\mathbf{x}_i) = \alpha(\tau) + \mathbf{x}_i\boldsymbol{\beta}(\tau). \tag{9}$$

Under (9), consistent estimators of $\alpha(\tau)$ and $\boldsymbol{\beta}(\tau)$ are obtained by minimizing the "check" function:

$$\min_{\alpha\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^K} \sum_{i=1}^{N} c_\tau(y_i - \alpha - \mathbf{x}_i\boldsymbol{\beta}), \tag{10}$$

where $c_\tau(u) = (\tau 1[u \geq 0] + (1-\tau)1[u < 0])|u| = (\tau - 1[u < 0])u$ and $1[\cdot]$ is the "indicator function." Consistency is relatively easy to establish because $(\alpha(\tau), \boldsymbol{\beta}(\tau))$ are known to minimize $E[c_\tau(y_i - \alpha - \mathbf{x}_i\boldsymbol{\beta})]$ (for example, Manski (1988)). Asymptotic

normality is more difficult because any sensible definition of the Hessian of the objective function, away from the nondifferentiable kink, is identically zero. But it has been worked out under a variety of conditions; see Koenker (2005) for a recent treatment.

2. Some Useful Asymptotic Results

What Happens if the Quantile Function is Misspecified?

- Property of OLS: if α^* and $\boldsymbol{\beta}^*$ are the plims from the OLS regression y_i on $1, \mathbf{x}_i$ then these provide the smallest mean squared error approximation to $E(y|\mathbf{x}) = \mu(\mathbf{x})$ in that $(\alpha^*, \boldsymbol{\beta}^*)$ solve

$$\min_{\alpha,\boldsymbol{\beta}} E[(\mu(\mathbf{x}) - \alpha - \mathbf{x}\boldsymbol{\beta})^2]. \tag{11}$$

Under restrictive assumptions on distribution of $\mathbf{x}$, β_j^* can be equal to or

proportionl to average partial effects.

- Linear quantile formulation has been viewed by several authors as an approximation (Buchinsky (1991), Chamberlain (1991), Abadie, Angrist, Imbens (2002)). Recently, Angrist, Chernozhukov, and Fernandez-Val (2006) characterized the probability limit of the quantile regression estimator. Absorb the intercept into $\mathbf{x}$ and let $\boldsymbol{\beta}(\tau)$ be the solution to the population quantile regression problem. ACF show that $\boldsymbol{\beta}(\tau)$ solves

$$\min_{\boldsymbol{\beta}} E\{w_\tau(\mathbf{x},\boldsymbol{\beta})[q_\tau(\mathbf{x}) - \mathbf{x}\boldsymbol{\beta})]^2\}, \tag{12}$$

where the weight function $w_\tau(\mathbf{x},\boldsymbol{\beta})$ is

$$w_\tau(\mathbf{x}, \boldsymbol{\beta}) = \int_0^1 (1-u) f_{y|x}(u\mathbf{x}\boldsymbol{\beta} + (1-u) q_\tau(\mathbf{x})|\mathbf{x}) du. \tag{13}$$

In other words, $\boldsymbol{\beta}(\tau)$ is the best weighted mean square approximation to the true quantile function, where the weights depend on average of the conditional density of y_i over a line from $\mathbf{x}\boldsymbol{\beta}$, to the true quantile function, $q_\tau(\mathbf{x})$.

Computing Standard Errors

- For given τ, write

$$y_i = \mathbf{x}_i\boldsymbol{\theta} + u_i, \text{ Quant}_\tau(u_i|\mathbf{x}_i) = 0, \tag{14}$$

and let $\hat{\boldsymbol{\theta}}$ be the quantile estimator. Define quantile residuals $\hat{u}_i = y_i - \mathbf{x}_i\hat{\boldsymbol{\theta}}$. Under weak conditions (see, for example, Koenker

(2005)), $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normal with asymptotic variance $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, where

$$\mathbf{A} \equiv \mathrm{E}[f_u(0|\mathbf{x}_i)\mathbf{x}_i'\mathbf{x}_i] \tag{15}$$

and

$$\mathbf{B} \equiv \tau(1-\tau)\mathrm{E}(\mathbf{x}_i'\mathbf{x}_i). \tag{16}$$

When we assume the quantile function is actually linear, a consistent estimator of $\mathbf{B}$ is

$$\hat{\mathbf{B}} = \tau(1-\tau)\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right). \tag{17}$$

Generally, a consistent estimator of $\mathbf{A}$ is (Powell (1986, 1991))

$$\hat{\mathbf{A}} = (2Nh_N)^{-1} \sum_{i=1}^{N} 1[|\hat{u}_i| \leq h_N] \mathbf{x}_i' \mathbf{x}_i, \tag{18}$$

where $\{h_N > 0\}$ is a nonrandom sequence shrinking to zero as $N \to \infty$ with $\sqrt{N} h_N \to \infty$. For example, $h_N = aN^{-1/3}$ for any $a > 0$. Might use a smoothed version so that all residuals contribute.

- Works for reasons similar to heteroskedasticity-robust standard errors.
- If u_i and $\mathbf{x}_i$ are independent,

$$Avar\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \frac{\tau(1-\tau)}{[f_u(0)]^2} [\mathrm{E}(\mathbf{x}_i' \mathbf{x}_i)]^{-1}, \tag{19}$$

and $Avar(\hat{\boldsymbol{\theta}})$ is estimated as

$$\widehat{Avar(\hat{\boldsymbol{\theta}})} = \frac{\tau(1-\tau)}{[\hat{f}_u(0)]^2}\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}, \tag{20}$$

where, say, $\hat{f}_u(0)$ is the histogram estimator

$$\hat{f}_u(0) = (2Nh_N)^{-1}\sum_{i=1}^{N}1[|\hat{u}_i| \leq h_N]. \tag{21}$$

Estimate in (20) is commonly reported (by, say, Stata).

- If the quantile function is misspecified, even the "robust" form of the variance matrix, based on the estimate in (20), is not valid. In the generalized linear models literature, the distinction is sometimes made between a "fully robust" variance estimator and a "semi-robust"

variance estimator. If mean is correctly specified and estimator allows unspecified variance, it is semi-robust. If the mean is allowed to be misspecified, fully robust.

- For quantile regression, a fully robust variance requires a different estimator of **B**. Kim and White (2002) and Angrist, Chernozhukov, and Fernández-Val (2006) show

$$\hat{\mathbf{B}} = \left(N^{-1} \sum_{i=1}^{N} (\tau - 1[\hat{u}_i < 0])^2 \mathbf{x}_i' \mathbf{x}_i \right) \tag{22}$$

is generally consistent, and then $\widehat{Avar(\hat{\boldsymbol{\theta}})} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$ with $\hat{\mathbf{A}}$ given by (18).

- Hahn (1995, 1997) shows that the nonparametric bootstrap and the

Bayesian bootstrap generally provide consistent estimates of the fully robust variance without claims about the conditional quantile being correct. Bootstrap does not provide “asymptotic refinements” for testing and confidence intervals.

- ACF provide the covariance function for the process $\{\hat{\boldsymbol{\theta}}(\tau) : \varepsilon \leq \tau \leq 1 - \varepsilon\}$ for some $\varepsilon > 0$, which can be used to test hypotheses jointly across multiple quantiles (including all quantiles at once).
- Example using Abadie (2003). These are nonrobust standard errors. *nettfa* is net total financial assets.

Dependent Variable:	*nettfa*			
Explanatory Variable	Mean (OLS)	.25 Quantile	Median (LAD)	.75 Quantile
inc	.783	.0713	.324	.798
	(.104)	(.0072)	(.012)	(.025)
age	−1.568	.0336	−.244	−1.386
	(1.076)	(.0955)	(.146)	(.287)
age^2	.0284	.0004	.0048	.0242
	(.0138)	(.0011)	(.0017)	(.0034)
*e*401*k*	6.837	1.281	2.598	4.460
	(2.173)	(.263)	(.404)	(.801)
N	2,017	2,017	2,017	2,017

3. Quantile Regression with Endogenous Explanatory Variables

- Suppose

$$y_1 = \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + u_1, \tag{23}$$

where $\mathbf{z}$ is exogenous and y_2 is endogenous – whatever that means in the context of quantile regression.

- First, LAD. Amemiya's (1982) two-stage LAD estimator adds a reduced form for y_2, say

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2. \tag{24}$$

First step applies OLS or LAD to (24), and gets fitted values, $y_{i2} = \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$. These are inserted for y_{i2} to give LAD of y_{i1} on $\mathbf{z}_{i1}, \hat{y}_{i2}$. The 2SLAD estimator relies on symmetry of the composite error $\alpha_1 v_2 + u_1$

given $\mathbf{z}$.

- If $D(u_1, v_2|\mathbf{z})$ is centrally symmetric, can use a control function approach. Write

$$u_1 = \rho_1 v_2 + e_1, \tag{25}$$

where e_1 given $\mathbf{z}$ would have a symmetric distribution. Get LAD residuals $\hat{v}_{i2} = y_{i2} - \mathbf{z}_i\hat{\boldsymbol{\pi}}_2$ and do LAD of y_{i1} on $\mathbf{z}_{i1}, y_{i2}, \hat{v}_{i2}$. Use t test on $\hat{v}_{i2}$ to test null that y_2 is exogenous.

- Interpretation of LAD in context of omitted variables is difficult unless lots of symmetry assumed.

- Abadie (2003) and Abadie, Angrist, and Imbens (2002) define and estimate policy parameters with a binary endogenous treatment, say D,

and binary instrumental variable, say Z. The potential outcomes are Y_d, $d = 0, 1$ – that is, without treatment and with treatment, respectively. The counterfactuals for treatment are D_z, $z = 0, 1$. Observed are $X, Z, D = (1 - Z)D_0 + ZD_1$, and $Y = (1 - D)Y_0 + DY_1$. AAI study treatment effects for *compliers*, that is, the (unobserved) subpopulation with $D_1 > D_0$. The assumptions are

$$(Y_1, Y_0, D_1, D_0) \text{ independent of } Z \text{ conditional on } X \tag{26}$$

$$0 < P(Z = 1|X) < 1 \tag{27}$$

$$P(D_1 = 1|X) \neq P(D_0 = 1|X) \tag{28}$$

$$P(D_1 \geq D_0|X) = 1. \tag{29}$$

Under these assumptions, treatment is unconfounded for compliers:

$$D(Y_0, Y_1|D, X, D_1 > D_0) = D(Y_0, Y_1|X, D_1 > D_0) \tag{30}$$

and so treatment effects can be defined based on $D(Y|X, D, D_1 > D_0)$, where Y is the observed outcome. AAI focus on *quantile treatment effects* (Abadie looks at other distributional features):

$$Quant_\tau(Y|X, D, D_1 > D_0) = \alpha_\tau D + X\beta_\tau. \tag{31}$$

(This results in estimated differences for the quantiles of Y_1 and Y_0, not the quantile of the difference $Y_1 - Y_0$.)

- If the dummy variable $C = 1[D_1 > D_0]$ could be observed, problem would be straightforward. Would like to use linear quantile estimation for the subpopulation $C = 1$ because the parameters solve

$$\min_{\alpha,\beta} E[C \cdot g(Y,X,D,\alpha,\beta)] \tag{32}$$

where $g(Y,X,D,\alpha,\beta) = c_\tau(Y - \alpha D - X\beta)]$ is the check function for a linear quantile estimation. Instead, can solve

$$\min_{\alpha,\beta} E[\kappa(U) \cdot g(Y,X,D,\alpha,\beta)], \tag{33}$$

where $U = (Y,X,D)$ and $\kappa(U) = P(C = 1|U)$. AAI show

$$\kappa_v(U) = 1 - \frac{D(1 - v(U))}{1 - \pi(X)} - \frac{(1 - D)v(U)}{\pi(X)}, \tag{34}$$

where $v(U) = P(Z = 1|U)$, and $\pi(X) = P(Z = 1|X)$, which can both be estimated using observed data.

- Two-step estimator solves

$$\min_{\delta} \sum_{i=1}^{N} 1[\hat{\kappa}_v(U_i) \geq 0]\hat{\kappa}_v(U_i)c_\tau(Y_i - W_i\delta). \tag{35}$$

where $W_i = (D_i, X_i)$ and δ contains α and β. The indicator function $1[\hat{\kappa}_v(U_i) \geq 0]$ ensures that only observations with nonnegative weights are used. Can use flexible parametric models (series) estimators for $\hat{v}(u)$ and $\hat{\pi}(x)$.

- Chernozhukov and Hansen (2005, 2006) consider identification and estimation of QTEs in a model with endogenous treatment. Let $q(d, x, \tau)$ denote the τ^{th} quantile function for treatment level $D = d$ and covariates x. In the binary case, CH define the QTE as

$$QTE_\tau(x) = q(1, x, \tau) - q(0, x, \tau). \tag{36}$$

- CH use the representation that Y_d, conditional on $X = x$, can be expressed as

$$Y_d = q(d, x, U_d) \tag{37}$$

where

$$U_d | Z \sim \text{Uniform}(0, 1), \tag{38}$$

and Z is the instrumental variable for treatment assignment, D. Key assumptions are that $q(d, x, u)$ is strictly increasing in u and a "rank invariance" condition, whose simplest form is conditional on $X = x$ and $Z = z$, U_d does not depend on d. CH show that, with the observed Y defined as $Y = q(D, X, U_D)$,

$$P[Y \leq q(D,X,\tau)|X,Z] = P[Y < q(D,X,\tau)|X,Z] = \tau. \quad (39)$$

If we could take $Z = D$, (39) would define the quantile $\text{Quant}_\tau(Y|D,X)$. Generally, it defines conditional moment conditions

$$E(\{1[Y \leq q(D,X,\tau)] - \tau\}|X,Z) = 0, \quad (40)$$

which is analogous to conditional moment conditions in models with additive errors.

- Chernozhukov and Hansen (2006) assume a linear functional form and obtain the *quantile regression instrumental variables estimator*.

4. Quantile Regression for Panel Data

- Without unobserved effects, easy to use quantile regression methods on panel data:

$$\text{Quant}_{\tau}(y_{it}|\mathbf{x}_{it}) = \mathbf{x}_{it}\boldsymbol{\theta},\ t = 1,\ldots,T. \tag{41}$$

Use pooled quantile regression. But need to generally account for serial correlation in the "scores,

$$\mathbf{s}_{it}(\boldsymbol{\theta}) = -\mathbf{x}_{it}'\{\tau 1[y_{it} - \mathbf{x}_{it}\boldsymbol{\theta} \geq 0] - (1-\tau)1[y_{it} - \mathbf{x}_{it}\boldsymbol{\theta} < 0]\}.$$

Use

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{T}\mathbf{s}_{it}(\hat{\boldsymbol{\theta}})\mathbf{s}_{ir}(\hat{\boldsymbol{\theta}})' \tag{42}$$

and then

$$\hat{\mathbf{A}} = (2Nh_N)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}1[|\hat{u}_{it}| \leq h_N]\mathbf{x}_{it}'\mathbf{x}_{it}. \tag{43}$$

- Explicitly allowing unobserved effects is harder.

$$\text{Quant}_\tau(y_{it}|\mathbf{x}_i, c_i) = \text{Quant}_\tau(y_{it}|\mathbf{x}_{it}, c_i) = \mathbf{x}_{it}\boldsymbol{\theta} + c_i. \tag{44}$$

- "Fixed effects" approach, where do not restrict $\text{D}(c_i|\mathbf{x}_i)$, is attractive. From Honoré (1992) applied to the uncensored case, LAD on the first differences is consistent when $\{u_{it} : t = 1, \ldots, T\}$ is an iid. sequence conditional on $(\mathbf{x}_i, c_i)$, even if the common distribution is not symmetric. But this is a fairly strong assumption. When $T = 2$, applying LAD on the first differences is equivalent to estimating the c_i along with $\boldsymbol{\theta}$. Generally, an incidental parameters problem with small T.
- Alternative suggested by Abrevaya and Dahl (2006) for $T = 2$. In

Chamberlain's correlated random effects linear model,

$$E(y_t|\mathbf{x}_1,\mathbf{x}_2) = \psi_t + \mathbf{x}_t\boldsymbol{\beta} + \mathbf{x}_1\boldsymbol{\xi}_1 + \mathbf{x}_2\boldsymbol{\xi}_2, t = 1, \tag{45}$$

$$\boldsymbol{\beta} = \frac{\partial E(y_1|\mathbf{x})}{\partial \mathbf{x}_1} - \frac{\partial E(y_2|\mathbf{x})}{\partial \mathbf{x}_1}. \tag{46}$$

Abrevaya and Dahl suggest modeling $\text{Quant}_\tau(y_t|\mathbf{x}_1,\mathbf{x}_2)$ as in (46) and then defining the partial effect as

$$\boldsymbol{\beta}_\tau = \frac{\partial \text{Quant}_\tau(y_1|\mathbf{x})}{\partial \mathbf{x}_1} - \frac{\partial \text{Quant}_\tau(y_2|\mathbf{x})}{\partial \mathbf{x}_1}. \tag{47}$$

- Generally, correlated random effects approaches are hampered because finding quantiles of sums of random variables is difficult. Suppose we write $c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i$ and then

$$y_{it} = \psi + \mathbf{x}_{it}\boldsymbol{\theta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it}. \tag{48}$$

Generally, $v_{it} = a_i + u_{it}$ will not have zero conditional quantile. Could just estimate (48) by pooled quantile regression for different quantiles and use the ACF results on approximating quantiles.

- A little more flexibility if we start with median,

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\theta} + c_i + u_{it}, \text{Med}(u_{it}|\mathbf{x}_i, c_i) = 0, \tag{49}$$

and make symmetry assumptions. If $\text{D}(\mathbf{u}_i|\mathbf{x}_i) = \text{D}(-\mathbf{u}_i|\mathbf{x}_i)$ then all linear combinations of the errors have a symmetric distribution, and so we can apply LAD to the time-demeaned equation $\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\theta} + \ddot{u}_{it}$, being sure to obtain fully robust standard errors for pooled LAD.

- If we impose the Chamberlain-Mundlak device as in (48), we can get

by with central symmetry of $\mathrm{D}(a_i, u_{it}|\mathbf{x}_i)$ has a symmetric distribution around zero then $\mathrm{D}(a_i + u_{it}|\mathbf{x}_i)$ is symmetric about zero, and, if this holds for each t, pooled LAD of y_{it} on $1, \mathbf{x}_{it}$, and $\bar{\mathbf{x}}_i$ consistently estimates $(\psi_t, \boldsymbol{\theta}, \boldsymbol{\xi})$. (If we use pooled OLS with $\bar{\mathbf{x}}_i$ included, we obtain the FE estimate.) Should use robust inference.

5. Quantile Methods for "Censored" Data

- Censored LAD applicable to data censoring and and corner solutions.

Very useful for true data censoring, where parameters of underlying linear model are of interest. w_i is the response variable (say, wealth or log of a duration) following

$$w_i = \mathbf{x}_i\boldsymbol{\beta} + u_i, \tag{50}$$

but it is top coded or right censored at r_i, then we can estimate $\boldsymbol{\beta}$ under the assumption

$$Med(u_i|\mathbf{x}_i, r_i) = 0 \tag{51}$$

because $Med(y_i|\mathbf{x}_i, r_i) = \min(\mathbf{x}_i\boldsymbol{\beta}, r_i)$ where $y_i = \min(y_i^*, r_i)$. Leads to Powell's (1986) CLAD estimator. (Need to always observe r_i; see

Honoré, Khan, and Powell (2002) to relax.)

- Less clear that CLAD is "better" than parametric models for corner solution responses. CLAD identifies a single feature of $D(y|\mathbf{x})$, namely, $Med(y|\mathbf{x})$. Models such as Tobit assume more but deliver more. Not just enough to estimate parameters. Common model for corner at zero:

$$y = \max(0, \mathbf{x}\boldsymbol{\beta} + u), \; Med(u|\mathbf{x}) = 0. \tag{52}$$

β_j measures the partial effects on $Med(y|\mathbf{x}) = \max(0, \mathbf{x}\boldsymbol{\beta})$ once $Med(y|\mathbf{x}) > 0$.

- A model no more or less restrictive than (52) is

$$y = a \cdot \exp(\mathbf{x}\boldsymbol{\beta}), \; E(a|\mathbf{x}) = 1, \tag{53}$$

in which case $E(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta})$ is identified. Allows for corner because

$P(a = 0|\mathbf{x}) > 0$ is allowed.

- How to interpret panel data applications of CLAD for corner solutions?

$$Med(y_{it}|\mathbf{x}_i, c_i) = \max(0, \mathbf{x}_{it}\boldsymbol{\beta} + c_i). \tag{54}$$

Honoré (1992), Honoré and Hu (2004) show how to estimate $\boldsymbol{\beta}$ under exchangeability assumptions on the idiosyncratic errors in the latent variable model. The partial effect of x_{tj} on $Med(y_{it}|\mathbf{x}_{it} = \mathbf{x}_t, c_i = c)$ is

$$\theta_{tj}(\mathbf{x}_t, c) = 1[\mathbf{x}_t\boldsymbol{\beta} + c > 0]\beta_j. \tag{55}$$

What values should we insert for c? We need to know something about $D(c_i)$. The average of (55) across the distribution of unobserved heterogeneity would be average partial effects (on the median). Again,

we need to identify $D(c_i)$. The β_j give us the sign and relative effects of the APEs. If c_i has a $Normal(\mu_c, \sigma_c^2)$ distribution, then it is easy to show $E_{c_i}[\theta_{tj}(\mathbf{x}_t, c_i)] = \Phi[(\mu_c - \mathbf{x}_t\boldsymbol{\beta})/\sigma_c]\beta_j$.

"What's New in Econometrics"

Lecture 15

Generalized Method of Moments

and Empirical Likelihood

Guido Imbens

NBER Summer Institute, 2007

Outline

1. Introduction

2. Generalized Method of Moments Estimation

3. Empirical Likelihood

4. Computational Issues

5. A Dynamic Panel Data Model

1. Introduction

GMM has provided a very influential framework for estimation since Hansen (1982). Many models and estimators fit in.

In the case with over-identification the traditional approach is to use a two-step method with estimated weight matrix.

For this case Empirical Likelihood provides attractive alternative with higher order bias properties, and liml-like advantages in settings with high degrees of over-identification.

The choice between various EL-type estimators is less important than the choice between the class and two-step gmm.

Computationally the estimators are only marginally more demanding. Most effective seems to be to concentrate out Lagrange multipliers.

2. Generalized Method of Moments Estimation

Generic form of the GMM estimation problem: The parameter vector θ^* is a K dimensional vector, an element of Θ, which is a subset of $\mathbb{R}^K$. The random vector Z has dimension P, with its support $\mathcal{Z}$ a subset of $\mathbb{R}^P$.

The moment function, $\psi : \mathcal{Z} \times \Theta \to \mathbb{R}^M$, is a known vector valued function such that

$$\mathbb{E}\left[\psi(Z, \theta^*)\right] = 0, \quad \text{and} \quad \mathbb{E}\left[\psi(Z, \theta)\right] \neq 0, \text{ for all } \theta \neq \theta^*$$

The researcher has available an independent and identically distributed random sample $Z_1, Z_2, \ldots, Z_N$. We are interested in the properties of estimators for θ^* in large samples.

Example I: Maximum Likelihood

If one specifies the conditional distribution of a variable Y given another variable X as $f_{Y|X}(y|x,\theta)$, the score function satisfies these conditions for the moment function:

$$\psi(Y,X,\theta) = \frac{\partial \ln f}{\partial \theta}(Y|X,\theta).$$

By standard likelihood theory the score function has expectation zero only at the true value of the parameter.

Interpreting maximum likelihood estimators as generalized method of moments estimators suggests a way of deriving the covariance matrix under misspecification (e.g., White, 1982), as well as an interpretation of the estimand in that case.

Example II: Linear Instrumental Variables

Suppose one has a linear model

$$Y = X'\theta^* + \varepsilon,$$

with a vector of instruments Z. In that case the moment function is

$$\psi(Y, X, Z, \theta) = Z' \cdot (Y - X'\theta).$$

The validity of Z as an instrument, together with a rank condition implies that θ^* is the unique solution to $E[\psi(Y, X, Z, \theta)] = 0$.

Example III: A Dynamic Panel Data Model

Consider the following panel data model with fixed effects:

$$Y_{it} = \eta_i + \theta \cdot Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean zero given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. We have observations Y_{it} for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, with N large relative to T.

This is a stylized version of the type of panel data models studied in Keane and Runkle (1992), Chamberlain (1992), and Blundell and Bond (1998). This specific model has previously been studied by Bond, Bowsher, and Windmeijer (2001).

One can construct moment functions by differencing and using lags as instruments, as in Arellano and Bond (1991), and Ahn and Schmidt, (1995):

$$\psi_{1t}(Y_{i1},\ldots,Y_{iT},\theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot \Big((Y_{it} - Y_{it-1} - \theta \cdot (Y_{it-1} - Y_{it-2}) \Big).$$

This leads to $t-2$ moment functions for each value of $t = 3,\ldots,T$, leading to a total of $(T-1)\cdot(T-2)/2$ moments, with only a single parameter (θ).

In addition, under the assumption that the initial condition is drawn from the stationary long-run distribution, the following additional $T-2$ moments are valid:

$$\psi_{2t}(Y_{i1},\ldots,Y_{iT},\theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta \cdot Y_{it-1}).$$

GMM: Estimation

In the just-identified case where M, the dimension of ψ, and K, the dimension of θ are identical, one can generally estimate θ^* by solving

$$0 = \frac{1}{N}\sum_{i=1}^{N} \psi(Z_i, \widehat{\theta}_{\text{gmm}}). \tag{1}$$

Under regularity conditions solutions will be unique in large samples and consistent for θ^*. If $M > K$ there is in general there will be no solution to (1).

Hansen's solution was to minimize the quadratic form

$$Q_{C,N}(\theta) = \frac{1}{N}\left[\sum_{i=1}^{N} \psi(z_i, \theta)\right]' \cdot C \cdot \left[\sum_{i=1}^{N} \psi(z_i, \theta)\right],$$

for some positive definite $M \times M$ symmetric matrix C (which if $M = K$ still leads to a $\widehat{\theta}$ that solves the equation (1).

GMM: Large Sample Properties

Under regularity conditions the minimand $\widehat{\theta}_{\mathsf{gmm}}$ has the following large sample properties:

$$\widehat{\theta}_{\mathsf{gmm}} \xrightarrow{p} \theta^*,$$

$$\sqrt{N}(\widehat{\theta}_{\mathsf{gmm}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma' C \Gamma)^{-1} \Gamma' C \Delta C \Gamma (\Gamma' C \Gamma)^{-1}),$$

where

$$\Delta = \mathbb{E}\left[\psi(Z_i, \theta^*)\psi(Z_i, \theta^*)'\right] \quad \text{and} \quad \Gamma = \mathbb{E}\left[\frac{\partial}{\partial \theta'}\psi(Z_i, \theta^*)\right].$$

In the just–identified case with the number of parameters K equal to the number of moments M, the choice of weight matrix C is immaterial.

In that case Γ is a square matrix, and because it is full rank by assumption, Γ is invertible and the asymptotic covariance matrix reduces to $(\Gamma' \Delta^{-1} \Gamma)^{-1}$, irrespective of the choice of C.

GMM: Optimal Weight Matrix

In the overidentified case with $M > K$ the choice of the weight matrix C is important.

The optimal choice for C in terms of minimizing the asymptotic variance is in this case the inverse of the covariance of the moments, Δ^{-1}.

Then:

$$\sqrt{N}(\widehat{\theta}_{\text{gmm}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, (\Gamma' \Delta^{-1} \Gamma)^{-1}). \tag{2}$$

This estimator is not feasible because Δ^{-1} is unknown.

The feasible solution is to obtain an initial consistent, but generally inefficient, estimate of θ^* and then can estimate the optimal weight matrix as

$$\hat{\Delta}^{-1} = \left[\frac{1}{N}\sum_{i=1}^{N} \psi(z_i, \tilde{\theta}) \cdot \psi(z_i, \tilde{\theta})'\right]^{-1}.$$

In the second step one estimates θ^* by minimizing $Q_{\hat{\Delta}^{-1},N}(\theta)$.

The resulting estimator $\widehat{\theta}_{\text{gmm}}$ has the same first order asymptotic distribution as the minimand of the quadratic form with the true, rather than estimated, optimal weight matrix, $Q_{\Delta^{-1},N}(\theta)$.

Compare to TSLS having the same asymptotic distribution as estimator with optimal instrument.

GMM: Specification Testing

If the number of moments exceeds the number of free parameters, not all average moments can be set equal to zero, and their deviation from zero forms the basis of a test.Formally, the test statistic is

$$T = Q_{\widehat{\Delta},N}(\widehat{\theta}_{\text{gmm}}).$$

Under the null hypothesis that all moments have expectation equal to zero at the true value of the parameter the distribution of the test statistic converges to a chi-squared distribution with degrees of freedom equal to the number of over-identifying restrictions, $M - K$.

Interpreting Over-identified GMM as a Just-identified Moment Estimator

One can also interpret the two–step estimator for over–identified GMM models as a just–identified GMM estimator with an augmented parameter vector. Fix an arbitrary $M \times M$ postitive definite matrix C. Then:

$$h(x,\delta) = h(x,\theta,\Gamma,\Delta,\beta,\Lambda) = \begin{pmatrix} \Lambda - \frac{\partial \psi}{\partial \theta'}(x,\beta) \\ \Lambda' C \psi(x,\beta) \\ \Delta - \psi(x,\beta)\psi(x,\beta)' \\ \Gamma - \frac{\partial \psi}{\partial \theta'}(x,\theta) \\ \Gamma' \Delta^{-1} \psi(x,\theta) \end{pmatrix}. \quad (3)$$

This interpretation emphasizes that results for just–identified GMM estimators such as the validity of the bootstrap can directly be translated into results for over–identified GMM estimators.

For example, one can use the just-identified representation to find the covariance matrix for the over–identified GMM estimator that is robust against misspecification: the appropriate submatrix of

$$\left(E\left[\frac{\partial h}{\partial \delta}(X,\delta^*)\right]\right)^{-1} E[h(Z,\delta^*)h(Z,\delta^*)'] \left(E\left[\frac{\partial h}{\partial \delta}(Z,\delta^*)\right]\right)^{-1},$$

estimated by averaging at the estimated values. This is the GMM analogue of the White (1982) covariance matrix for the maximum likelihood estimator under misspecification.

Efficiency

Chamberlain (1987) demonstrated that Hansen's (1982) estimator is efficient, not just in the class of estimators based on minimizing the quadratic form $Q_{N,C}(\theta)$, but in the larger class of semiparametric estimators exploiting the full set of moment conditions.

Chamberlain assumes that the data are discrete with finite support $\{\lambda_1, \ldots, \lambda_L\}$, and unknown probabilities $\pi_1, \ldots, \pi_L$. The parameters of interest are then implicitly defined as functions of these points of support and probabilities. With only the probabilities unknown, the Cramér-Rao variance bound is conceptually straightforward to calculate.

It turns out this is equal to variance of GMM estimator with optimal weight matrix.

3. Empirical Likelihood

Consider a random sample $Z_1, Z_2, \ldots, Z_N$, of size N from some unknown distribution. The natural choice for estimating the dsitribution function is the empirical distribution, that puts weight $1/N$ on each of the N sample points.

Suppose we also know that $\mathbb{E}[Z] = 0$. The empirical distribution function with weights $1/N$ does not satisfy the restriction $E_F[Z] = 0$ as $E_{\widehat{F}_{emp}}[Z] = \sum z_i/N \neq 0$.

The idea behind empirical likelihood is to modify the weights to ensure that the estimated distribution $\widehat{F}$ does satisfy the restriction.

The empirical likelihood is

$$\mathcal{L}(\pi_1, \ldots, \pi_N) = \prod_{i=1}^{N} \pi_i, \quad \text{for } 0 \leq \pi_i \leq 1, \ \sum_{i=1}^{N} \pi_i = 1$$

The empirical likelihood estimator for the distribution function is, given $\mathbb{E}[Z] = 0$,

$$\max_{\pi} \sum_{i=1}^{N} \pi_i \quad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \text{ and } \sum_{i=1}^{N} \pi_i \cdot z_i = 0.$$

Without the second restriction the π's would be estimated to be $1/N$, but the second restriction forces them slightly away from $1/N$ in a way that ensures the restriction is satisfied.

This leads to

$$\widehat{\pi}_i = 1/(1 + t \cdot z_i) \quad \text{where t solves } \textstyle\sum_{i=1}^{N} \frac{z_i}{1+t \cdot z_i} = 0,$$

EL: The General Case

More generally, in the over-identified case a major focus is on obtaining point estimates through the following estimator for θ:

$$\max_{\theta,\pi} \sum_{i=1}^{N} \ln \pi, \quad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \ \sum_{i=1}^{N} \pi_i \cdot \psi(z_i, \theta) = 0.$$

This is equivalent, to first order asymptotics, to the two-step GMM estimator.

For many purposes the empirical likelihood has the same properties as a parametric likelihood function. (Qin and Lawless, 1994; Imbens, 1997; Kitamura and Stutzer, 1997).

EL: Cressie-Read Discrepancy Statistics

Define

$$I_\lambda(p,q) = \frac{1}{\lambda \cdot (1+\lambda)} \sum_{i=1}^{N} p_i \left[\left(\frac{p_i}{q_i} \right)^\lambda - 1 \right].$$

and solve

$$\min_{\pi,\theta} I_\lambda(\iota/N, \pi) \quad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \text{ and } \sum_{i=1}^{N} \pi_i \cdot \psi(z_i, \theta) = 0.$$

The precise way in which the notion "as close as possible" is implemented is reflected in the choice of metric through λ.

Empirical Likelihood is special case with $\lambda =\longrightarrow 0$.

EL: Generalized Empirical Likelihood

Smith (1997), Newey and Smith (1994) considers a more general class of estimators. For a given function $g(\cdot)$, normalized so that it satisfied $g(0) = 1$, $g'(0) = 1$, consider the saddle point problem

$$\max_{\theta} \min_{t} \sum_{i=1}^{N} g(t'\psi(z_i, \theta)).$$

This representation is attractive from a computational perspective, as it reduces the dimension of the optimization problem to $M + K$ rather than a constrained optimization problem of dimension $K + N$ with $M + 1$ restrictions.

There is a direct link between the t parameter in the GEL representation and the Lagrange multipliers in the Cressie-Read representation. NS show how to choose $g(\cdot)$ for a given λ so that the corresponding GEL and Cressie-Read estimators agree.

EL: Special cases, Continuously Updating Estimator

$\lambda = -2$.

This case was originally proposed by Hansen, Heaton and Yaron (1996) as the solution to

$$\min_{\theta} \frac{1}{N}\left[\sum_{i=1}^{N} \psi(z_i, \theta)\right]' \cdot \left[\frac{1}{N}\sum_{i=1}^{N} \psi(z_i, \theta)\psi(z_i, \theta)'\right]^{-1} \cdot \left[\sum_{i=1}^{N} \psi(z_i, \theta)\right],$$

where the GMM objective function is minimized over the θ in the weight matrix as well as the θ in the average moments.

Newey and Smith (2004) pointed out that this estimator fits in the Cressie-Read class.

EL: Special cases, Exponential Tilting Estimator

$\lambda \longrightarrow -1$.

The second case is the exponential tilting estimator with $\lambda \rightarrow -1$ (Imbens, Spady and Johnson, 1998), whose objective function is equal to the empirical likelihood objective funtion with the role of π and ι/N reversed.

It can also be written as

$$\min_{\pi,\theta} \sum_{i=1}^{N} \pi_i \cdot \ln \pi_i \quad \text{subject to } \sum_{i=1}^{N} \pi_i = 1, \text{ and } \sum_{i=1}^{N} \pi_i \cdot \psi(z_i, \theta) = 0.$$

Comparison of GEL Estimators

Little known in general.

EL ($\lambda = 0$) has higher order bias properties (NS), but implicit probabilities can get large.

CUE ($\lambda = -2$) tends to have more outliers

ET ($\lambda = -1$) computationally stable.

Testing

Likelihood Ratio test:

$$LR = 2 \cdot (L(\iota/N) - L(\hat{\pi})), \quad \text{where} \quad L(\pi) = \sum_{i=1}^{N} \ln \pi_i.$$

$$\text{WALD} = \frac{1}{N} \left[\sum_{i=1}^{N} \psi(z_i, \hat{\theta}) \right]' \hat{\Delta}^{-1} \left[\sum_{i=1}^{N} \psi(z_i, \hat{\theta}) \right],$$

where $\hat{\Delta}$ is some estimate of the covariance matrix of the moments.

Lagrange Multiplier test, based on estimated lagrange multipliers $\hat{t}$

$$LM = \hat{t}' \hat{\Delta} \hat{t}.$$

4. Computational Issues

In principle the EL estimator has many parameters (π_i and θ), which could lead to computational difficulties.

Solving the First Order Conditions the first order conditions does not work well.

Imbens, Spady and Johnson suggest penalty function approaches which work better, but not great.

Concentrating out the Lagrange Multipliers

Mittelhammer, Judge and Schoenberg (2001) suggest concentrating out both probabilities and Lagrange multipliers and then maximizing over θ without any constraints. This appears to work well.

Concentrating out the probabilities π_i can be done analytically.

Although it is not in general possible to solve for the Lagrange multipliers t analytically for given θ it is easy to numerically solve for t. E.g., in the exponential tilting case, solve

$$\min_t \sum_{i=1}^{N} \exp(t'\psi(z_i, \theta)).$$

This function is strictly convex as a function of t, with easy-to-calculate first and second derivatives.

After solving for $t(\theta)$, one can solve

$$\max_{\theta} \sum_{i=1}^{N} \exp(t(\theta)'\psi(z_i, \theta)).$$

Calculating first derivatives of the concentrated objective function only requires first derivatives of the moment functions, both directly and indirectly through the derivatives of $t(\theta)$ with respect to θ.

The function $t(\theta)$ has analytic derivatives with respect to θ equal to:

$$\frac{\partial t}{\partial \theta'}(\theta) = -\left(\frac{1}{N}\sum_{i=1}^{N} \psi(z_i, \theta)\psi(z_i, \theta)' \exp(t(\theta)'\psi(z_i, \theta)\right)^{-1}$$

$$\cdot \left(\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \psi}{\partial \theta'}(z_i, \theta) \exp(t(\theta)'\psi(z_i, \theta) + \psi(z_i, \theta)t(\theta)'\frac{\partial \psi}{\partial \theta'}(z_i, \theta) \exp(t(\theta)'\psi(z_i\right.$$

5. A Dynamic Panel Data Model

To get a sense of the finite sample properties of the empirical likelihood estimators we compare two-step GMM and one of the EL estimators (exponential tilting) in the context of a panel data model

The model is

$$Y_{it} = \eta_i + \theta \cdot Y_{it-1} + \varepsilon_{it},$$

where ε_{it} has mean zero given $\{Y_{it-1}, Y_{it-2}, \ldots\}$. We have observations Y_{it} for $t = 1, \ldots, T$ and $i = 1, \ldots, N$.

Moments:

$$\psi_{1t}(Y_{i1},\ldots,Y_{iT},\theta) = \begin{pmatrix} Y_{it-2} \\ Y_{it-3} \\ \vdots \\ Y_{i1} \end{pmatrix} \cdot \Big((Y_{it} - Y_{it-1} - \theta \cdot (Y_{it-1} - Y_{it-2}) \Big).$$

This leads to $(T-1)\cdot(T-2)/2$ moments.

Additional $T-2$ moments:

$$\psi_{2t}(Y_{i1},\ldots,Y_{iT},\theta) = (Y_{it-1} - Y_{it-2}) \cdot (Y_{it} - \theta \cdot Y_{it-1}).$$

Note that the derivatives of these moments are stochastic and potentially correlated with the moments themselves. So, potentially substantial difference between estimators.

We report some simulations for a data generating process with parameter values estimated on data from Abowd and Card (1989) taken from the PSID. See also Card (1994).

This data set contains earnings data for 1434 individuals for 11 years. The individuals are selected on having positive earnings in each of the eleven years, and we model their earnings in logarithms. We focus on estimation of the autoregressive coefficient θ.

Using the Abowd-Card data we estimate θ and the variance of the fixed effect and the idiosyncratic error term. The latter two are estimated to be around 0.3. We use $\theta = 0.5$ and $\theta = 0.9$ in the simulations. The first is comparable to the value estimated from the Abowd-Card data.

$\theta = 0.5$	Number of time periods					
	3	4	6	7	9	11
Two-Step GMM						
median bias	-0.00	0.00	-0.00	-0.00	0.00	0.00
relative median bias	-0.07	0.01	-0.06	-0.08	0.09	0.14
median absolute error	0.05	0.03	0.01	0.01	0.01	0.01
coverage rate 90% ci	0.91	0.88	0.91	0.91	0.89	0.90
covarage rate 95% ci	0.95	0.94	0.95	0.96	0.95	0.94
Exponential Tilting						
median bias	-0.00	-0.00	-0.00	-0.00	0.00	0.00
relative median bias	-0.04	-0.02	-0.09	-0.07	0.02	0.10
median absolute error	0.05	0.03	0.01	0.01	0.01	0.01
coverage rate 90% ci	0.90	0.87	0.90	0.92	0.90	0.91
covarage rate 95% ci	0.95	0.94	0.96	0.95	0.95	0.95

$\theta = 0.9$	Number of time periods					
	3	4	6	7	9	11
Two-Step GMM						
median bias	-0.00	0.00	0.00	0.00	0.00	0.00
relative median bias	-0.02	0.08	0.08	0.03	0.08	0.11
median absolute error	0.04	0.03	0.02	0.02	0.01	0.01
coverage rate 90% ci	0.88	0.85	0.80	0.80	0.78	0.76
covarage rate 95% ci	0.92	0.91	0.87	0.85	0.86	0.84
Exponential Tilting						
median bias	0.00	0.00	-0.00	0.00	-0.00	0.00
relative median bias	0.04	0.09	-0.00	0.01	-0.02	0.13
median absolute error	0.05	0.03	0.02	0.02	0.01	0.01
coverage rate 90% ci	0.87	0.86	0.86	0.88	0.87	0.87
covarage rate 95% ci	0.91	0.90	0.91	0.93	0.91	0.93