Econometrics of Survey Data

Manuel Arellano

CEMFI

September 2014

Introduction

- Suppose that a population is stratified by groups to guarantee that there will be enough observations to permit sufficiently precise estimates for each of the groups. For example, these groups can be locations, age intervals, or wealth levels.
- Next, each subpopulation or stratum is divided into spatial clusters of households (apartment compounds, villages) and a random sample of clusters is selected from each stratum. Finally, a small number of households is selected at random within each cluster.
- In general, sampling clusters is more cost effective than sampling households directly, despite the fact that individual observations within a cluster will likely be correlated and therefore less informative than observations from individuals sampled at random.
- The above describes a prototypical multistage survey design in which both stratification and clustering are present.
- We wish to discuss the implications of using this type of data for an econometric analysis where the interest is in drawing conclusions about the original population. The econometric analysis could be of a descriptive nature (decomposition analysis or hedonic price indexes) or aim for non-structural or structural causal inference.

Introduction (continued)

- It is also of interest to think of complex survey designs in a figurative sense.
- Sometimes we wish to do statistical analysis of data which have not come to us as a result of actual sampling. In those situations we implicitly decide the abstract population of interest by imagining a sampling design that could have produced the data at hand from such population.
- This is the case when we work with the universe of firms of certain characteristics or when we work with cross-sectional or longitudinal data on countries.
- At a more general level, the question of inference is: how much confidence can I have that the pattern I see in the data is typical of some larger population, or that the apparent pattern is not only a random occurrence?
- From this perspective there is a role for inferential statistics as long as it can be anticipated that estimation error is not negligible relative to some conceptual population of interest.

Outline

- Introduction
    - Finite populations
- Part 1: Weighted estimation and stratification
    - Weighted means and estimating equations
    - Standard errors under stratification
    - Examples: OLS, IV, QR, ML, subpopulations
    - Endogenous and exogenous stratification
    - Population heterogeneity vs sample design
    - Choice-based sampling
- Part 2: Cluster standard errors
    - Introduction
    - Cluster fixed-effects vs cluster standard errors
    - General set up: estimating equations
    - Examples: panel data, quantile regression
- Part 3: Bootstrap methods
    - The idea of the bootstrap
    - Bootstrap standard errors
    - Asymptotic properties
    - Bootstrapping stratified clustered samples
    - Using replicate weights

Finite populations

- Inference in microeconometrics typically assumes that data are generated as a random sample with replacement or that is coming from an infinite population.
- These assumptions are not literally valid in the case of survey data that come from a finite population of units.
- There is an extensive literature in statistics dealing with the sampling schemes that arise in survey production (e.g. Levy and Lemeshow 1991).

Random sampling without replacement

- Consider a finite population $\overline{N}$ and the population mean given by

$$\theta_0 = \frac{1}{\overline{N}} \sum_{i=1}^{\overline{N}} Y_i.$$

- Also define the following population variance

$$S^2 = \frac{1}{\overline{N}-1} \sum_{i=1}^{\overline{N}} (Y_i - \theta_0)^2 = \frac{1}{\overline{N}-1} Y' Q_{\overline{N}} Y,$$

  where $Q_{\overline{N}} = I - \iota\iota'/\overline{N}$ is the deviation from means operator of order $\overline{N}$.

- The $Y = \left(Y_1, ..., Y_{\overline{N}}\right)'$ represent a full enumeration of the values of the variable and are therefore nonstochastic quantities.

Finite populations (continued)

- Let $y_1, ..., y_N$ be a random sample of size N *without replacement* and sample mean:

$$\overline{Y} = \frac{1}{N} \sum_{j=1}^{N} y_j$$

- The sample mean can be rewritten as a random selection of population values

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{\overline{N}} a_i Y_i$$

  where a_i is an indicator that takes the value 1 if i is in the sample, and 0 otherwise.

- Letting $f = N/\overline{N}$, under random sampling

$$E(a_i) = f$$

$$Var(a_i) = f(1-f)$$

$$Cov(a_i, a_j) = -\frac{1}{(\overline{N}-1)} f(1-f),$$

  or in compact form,

$$Var(a) = f(1-f) \frac{\overline{N}}{(\overline{N}-1)} Q_{\overline{N}}$$

- The result uses that $E(a_i a_j) = \frac{N}{\overline{N}} \frac{(N-1)}{(\overline{N}-1)}$ due to sampling without replacement.

Finite populations (continued)

- We immediately see that the sample mean is an unbiased estimator of the population mean:

$$E\left(\overline{Y}\right) = \frac{1}{N}\sum_{i=1}^{\overline{N}} E\left(a_i\right) Y_i = \frac{1}{N}\sum_{i=1}^{\overline{N}} \frac{N}{\overline{N}} Y_i = \theta_0$$

- Moreover, the variance is given by

$$Var\left(\overline{Y}\right) = \frac{1}{N^2} Y' Var\left(a\right) Y = \left(1 - f\right) \frac{S^2}{N}$$

- The term $1 - f$ is the finite population correction. When $N \longrightarrow \overline{N}$ the sampling variance goes to zero, $Var\left(\overline{Y}\right) \longrightarrow 0$.
- A multivariate generalization is straightforward.
- Typically, the population size is sufficiently large relative to the sample size that $1 - f \approx 1$ and the correction can be ignored. This is the approach we'll take in what follows.
- Standard asymptotic theory studies the properties of sample statistics from infinite populations as $N \longrightarrow \infty$.
- An alternative asymptotic framework to mimic situations where sample size is not negligible relative to population size, is one in which both N and $\overline{N}$ tend to infinity while f tends to some positive fraction.

Finite populations (continued)

- An unbiased estimate of the population variance S^2 is given by

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)^2 = \frac{1}{N-1}y'Q_N y$$

- Ullah and Breunig (1998) show that $E\left(s^2\right) = S^2$. They also consider results for linear regression and generalizations to stratified samples and clustering.

- It turns out that the OLS variance when sampling is without replacement is not necessarily smaller than the infinite population OLS variance. The direction of the inequality depends upon the matrix of regressors.

- Moreover, it is possible to consider a GLS estimator that takes into account that errors are not independent. Contrary to the mean model, GLS does not coincide with OLS.

Part 1: Weighted estimation and stratification

Stratified sampling

- We focus on the so called "standard stratified sampling": The population is divided into S non-overlapping groups or strata. Then for $s = 1, ..., S$, we draw a random sample of size N_s from stratum s. The total sample size is $N = N_1 + ... + N_S$.

- Related sampling schemes are "variable probability sampling" and "multinomial sampling" (e.g. Cosslett 1993, and Wooldridge 2010).

- Let q_s be the *known* population relative frequency for stratum s. Consider some variable Y whose probability distribution in the population is $f(y)$. This can be written as

$$f(y) = f(y \mid s = 1) q_1 + ... + f(y \mid s = S) q_S$$

  where $f(y \mid s)$ is the probability distribution of Y in stratum s.

- Suppose we have drawn a stratified sample $y_1, ..., y_N$. This is as if we had drawn a random sample from the artificial population

$$f^*(y) = f(y \mid s = 1) \overline{h}_1 + ... + f(y \mid s = S) \overline{h}_S$$

  where $h_s = N_s / N$ and $\overline{h}_s = \lim h_s$.

Using weighted estimators

- If we want to consistently estimate the population mean of Y from the stratified data the stratification needs to be undone by re-weighting the observations.
- First, sample means for each strata are calculated

$$\widehat{E}\left(Y \mid s\right) = \frac{1}{N_s} \sum_{i \in s} y_i$$

- Next the overall mean is calculated by aggregating over strata using population frequencies instead of sample frequencies:

$$\widehat{E}\left(Y\right) = \sum_{s=1}^{S} \widehat{E}\left(Y \mid s\right) q_s = \frac{1}{N} \sum_{i=1}^{N} \omega_{s(i)} y_i$$

where $\omega_s = q_s / h_s$ and $s\left(i\right)$ is the stratum to which individual i belongs.

Weight factors

- The weight factor $\omega_{s(i)}$ is the ratio of the population share to the sample share.
- It informs us that household i is representative of $\omega_{s(i)}$–times the number of households it would represent in a random sample.
- Over-sampled categories are given a low weight, and vice-versa.
- In a random sample $\omega_{s(i)} = 1$ for all i, so that each unit in a sample of size N is representative of $\overline{N}/N$ units in a population of size $\overline{N}$ (e.g. $\overline{N}/N = 2000$ if $N = 500$ and $\overline{N} = 1,000,000$).
- For example, $\omega_{s(i)} = 3$ means that unit i represents $\phi_i = \omega_{s(i)} \times \overline{N}/N = 6000$ comparable units.
- Weight factors add up to N:

$$\sum_{i=1}^{N} \omega_{s(i)} = N_1\omega_1 + ... + N_S\omega_S = q_1 N + ... + q_S N = N$$

  Moreover, $\sum_{i=1}^{N} \phi_i = \overline{N}$.

- Using ϕ_i, $\widehat{E}(Y)$ can be seen as an approximate calculation of the population mean:

$$\widehat{E}(Y) = \frac{1}{\overline{N}} \sum_{i=1}^{N} \phi_i y_i.$$

- The ϕ_i are the standard definition of population weights in survey methodology.

Weighted estimating equations

- A generalization of the weighted mean in a GMM framework is as follows.
- Suppose we are interested in the estimation of some parameter θ_0, which is a characteristic of the possibly multivariate distribution $f(w)$ identified by the moment condition

$$E\psi(W, \theta_0) = 0.$$

- In a stratified sample we would therefore expect

$$\sum_{s=1}^{S} \widehat{E}[\psi(W, \theta_0) \mid s] q_s = \frac{1}{N} \sum_{i=1}^{N} \omega_{s(i)} \psi(w_i, \theta_0) \approx 0$$

- So it is natural to estimate θ_0 by $\widehat{\theta}$, which precisely solves

$$\frac{1}{N} \sum_{i=1}^{N} \omega_{s(i)} \psi\left(w_i, \widehat{\theta}\right) = 0. \tag{1}$$

- $\widehat{\theta}$ is a weighted generalized-method-of-moments (GMM) estimator.
- Here $\dim \psi = \dim \theta$. If $\dim \psi > \dim \theta$, $\widehat{\theta}$ will solve $\dim \theta$ linear combinations of (1).

Asymptotic Variance

- Under standard regularity conditions the asymptotic variance of $\widehat{\theta}$ can be estimated as

$$H^{-1}\sum_{s=1}^{S}\omega_s^2\left[\sum_{i\in s}^{N_s}\left(\psi\left(w_i,\widehat{\theta}\right)-\overline{\psi}_s\right)\left(\psi\left(w_i,\widehat{\theta}\right)-\overline{\psi}_s\right)'\right]H'^{-1}$$

  where $H=\sum_{i=1}^{N}\omega_{s(i)}\frac{\partial\psi\left(w_i,\widehat{\theta}\right)}{\partial\theta'}$ and $\overline{\psi}_s=\frac{1}{N_s}\sum_{i\in s}^{N_s}\psi\left(w_i,\widehat{\theta}\right)$.

- Means by strata need to be subtracted because in general $E\left[\psi\left(W,\theta_0\right)\mid s\right]\neq 0$.

Example: Asymptotic variance of the weighted sample mean

- If $\psi\left(w_i,\theta\right)=y_i-\theta$, we get $H=-\sum_{i=1}^{N}\omega_{s(i)}=-N$ and hence

$$\frac{1}{N^2}\sum_{s=1}^{S}\omega_s^2\left[\sum_{i\in s}^{N_s}\left(y_i-\overline{y}_s\right)^2\right]=\sum_{s=1}^{S}\frac{N_s^2}{N^2}\omega_s^2\left[\frac{1}{N_s^2}\sum_{i\in s}^{N_s}\left(y_i-\overline{y}_s\right)^2\right]=\sum_{s=1}^{S}q_s^2\frac{\widehat{\sigma}_s^2}{N_s}$$

  where

$$\widehat{\sigma}_s^2=\frac{1}{N_s}\sum_{i\in s}^{N_s}\left(y_i-\overline{y}_s\right)^2.$$

- A suitable choice of $h_1,...,h_S$ will be variance reducing relative to random sampling.

Examples

- The first example is *linear regression*:

$$\psi\left(w_i, \theta\right) = x_i\left(y_i - x_i'\theta\right)$$

- Solving (1) we get the weighted-least squares estimator

$$\widehat{\theta} = \left(\sum_{i=1}^{N} \omega_{s(i)} x_i x_i'\right)^{-1} \sum_{i=1}^{N} \omega_{s(i)} x_i y_i$$

- The second example is *linear instrumental-variables*:

$$\psi\left(w_i, \theta\right) = z_i\left(y_i - x_i'\theta\right)$$

leading to

$$\widehat{\theta} = \left(\sum_{i=1}^{N} \omega_{s(i)} z_i x_i'\right)^{-1} \sum_{i=1}^{N} \omega_{s(i)} z_i y_i.$$

- The third example is *quantile regression*:

$$\psi\left(w_i, \theta\right) = x_i\left[1\left(y_i \leq x_i'\theta\right) - \tau\right].$$

In this case $\widehat{\theta}$ solves for $\tau \in (0,1)$

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{N} \omega_{s(i)} \rho_\tau\left(y_i - x_i'\theta\right)$$

where $\rho_\tau(u)$ is the "check" function $\rho_\tau(u) = \left[\tau \mathbf{1}(u \geq 0) + (1-\tau)\mathbf{1}(u < 0)\right] \times |u|$.

Conditional likelihood model

- The next example is one where $w = (y, x)$ and $f(w) = f(y \mid x, \theta)\, g(x)$. A parametric model is specified for $f(y \mid x, \theta)$ while $g(x)$ is left unrestricted. Here

$$\psi(w_i, \theta) = \frac{\partial \ln f(y_i \mid x_i, \theta)}{\partial \theta}$$

  and we solve

$$\frac{1}{N} \sum_{i=1}^{N} \omega_{s(i)} \frac{\partial \ln f(y_i \mid x_i, \theta)}{\partial \theta} = 0.$$

- This estimator was first proposed by Manski and Lerman (1977) in the context of estimation of discrete-choice models from choice-based samples (sometimes confusingly called "weighted exogenous sample ML") –more to follow.
- It provides consistent estimates under endogenously stratified sampling (with known q_s) but it is not efficient.
- The reason why it is not efficient is that it does not attempt to reconcile the actual known values of q_s with those predicted by the model.
- Efficient estimation is discussed in Cosslett (1993), Imbens (1992), and Imbens and Lancaster (1996). These estimators are applicable in an ML context were the model specifies a conditional likelihood.

Estimating parameters from subpopulations

- In the last example interest is in estimating a characteristic θ_0 of some subpopulation.
- Let d_i indicate membership of the subpopulation and suppose that θ_0 satisfies

$$E\left[\psi\left(w_i, \theta_0\right) \mid d_i = 1\right] = 0$$

or equivalently

$$E\left[\psi\left(w_i, \theta_0\right) d_i\right] = 0.$$

- Estimation is based on the sample moment

$$\sum_{i=1}^{N} \omega_{s(i)} d_i \psi\left(w_i, \widehat{\theta}\right) = 0.$$

- For example, if interested in the mean of y_i for the units with $d_i = 1$ we get:

$$\sum_{i=1}^{N} \omega_{s(i)} d_i \left(y_i - \widehat{\theta}\right) = 0.$$

so that $\widehat{\theta}$ is the conditional weighted average:

$$\widehat{\theta} = \frac{1}{\sum_{j=1}^{N} \omega_{s(j)} d_j} \sum_{i=1}^{N} \omega_{s(i)} d_i y_i.$$

Endogenous and exogenous stratification

- Strata are based on the values of certain variables like geographic location or age.
- Often we are interested in estimating characteristics of the conditional distribution of some variable y given other variables x.
- For example, in regression analysis one is interested in a conditional mean $E(y \mid x)$ or a conditional quantile $Q_\tau(y \mid x)$. In a linear approach one assumes $E(y \mid x) = x'\theta_0$ so that θ_0 becomes the object of interest in estimation.
- If stratification is based exclusively on x (i.e. the stratification variables are functions of x) there is no need to use sample weights in estimation. The standard unweighted estimator remains consistent in the stratified sample.
- However, weighting is essential if interested in the estimation of unconditional quantities, such as $E(y)$ or $Q_\tau(y)$, or on conditional quantities that are not conditioned on the stratification variables.

Endogenous and exogenous stratification (continued)

- More generally, in a GMM framework the condition for *ignorable stratification* is that the moment equation $E\psi(W, \theta_0) = 0$ holds in each strata:

$$E[\psi(W, \theta_0) \mid s] = 0. \tag{2}$$

- The intuition is that if (2) holds, $\widehat{E}[\psi(W, \theta_0) \mid s] \approx 0$ for all s, so we not only expect

$$\sum_{s=1}^{S} \widehat{E}[\psi(W, \theta_0) \mid s] q_s = \frac{1}{N} \sum_{i=1}^{N} \omega_{s(i)} \psi(w_i, \theta_0) \approx 0$$

  but also

$$\sum_{s=1}^{S} \widehat{E}[\psi(W, \theta_0) \mid s] h_s = \frac{1}{N} \sum_{i=1}^{N} \psi(w_i, \theta_0) \approx 0$$

Endogenous and exogenous stratification (continued)

Linear projections and exogenous stratification

- If a linear regression parameter θ_0 only denotes a linear projection approximation to a nonlinear conditional expectation $E(y \mid x)$, weighting is required even if $s = s(x)$.
- The problem is that if $E[x(y - x'\theta_0)] = E(xu) = 0$ holds but $E(y \mid x) \neq x'\theta_0$, u and x will not be uncorrelated within strata:

$$E(xu \mid s) = E[xE(u \mid x, s) \mid s] = E[xE(u \mid x) \mid s] \neq 0$$

- Intuitively, a linear regression best approximates $E(y \mid x)$ at high density values of x. Since stratification changes the density of x it will also change the approximating linear projection.
- The same is true in an instrumental-variable model:

$$E[z(y - x'\theta_0)] = E(zu) = 0$$

  if the conditional mean restriction $E(u \mid z) = 0$ does not hold. In this situation weighting will be required even if the stratification variables depend exclusively on z (more later).

Endogenous and exogenous stratification (continued)

- The context of the endogenous/exogenous stratification terminology is an econometric model of y and x, which only specifies the conditional distribution of y (endogenous) given x (exogenous).
- Since the distribution of x is left unrestricted and plays no part in the estimation

$$f(w) = f(y \mid x, \theta)\, g(x),$$

  stratification can be ignored if only modifies the form of $g(x)$.
- In contrast, in endogenous or *non-ignorable stratification*, the strata are defined in terms of y (and possibly other variables).

Population heterogeneity vs sample design

- One interpretation of the weighted sample mean is that different means in each strata are first calculated and then combined into an average weighted by population frequencies. We wish to discuss the merits of this way of thinking in a regression case.
- Suppose that θ_s is a regression coefficient in stratum s (or some other characteristic of the distribution in stratum s), which varies across strata, and the object of interest is

$$\overline{\theta} = \sum_{s=1}^{S} \theta_s q_s$$

- In general, the weighted regression estimator will not be a consistent estimator of $\overline{\theta}$.
- This is sometimes held as an argument against weighted regression: when θ_s are homogeneous, unweighted regression (OLS) is more efficient and when they are not, both weighted and unweighted estimators are inconsistent for $\overline{\theta}$ (Deaton 1997).
- Where is the catch? The key is whether the substantive interest is in $\overline{\theta}$ (an average of effects across strata) or on some quantity defined for the overall population.
- In general, the parameter $\overline{\theta}$ will be of interest when variability in the θ_s captures meaningful population heterogeneity in individual responses.
- If we are using a regression as a device to summarize characteristics of the population, then $\overline{\theta}$ is not the parameter of interest but rather the population-wide regression coefficient, which is precisely the object that weighted regression tries to approximate.

Local average treatment effects

- As an illustration, consider an instrumental-variable treatment effect setting where there is an outcome variable Y_i, a binary treatment D_i and a binary instrument Z_i.
- The parameter of interest is

$$\theta_0 = \frac{E\left(Y_i \mid Z_i = 1\right) - E\left(Y_i \mid Z_i = 0\right)}{E\left(D_i \mid Z_i = 1\right) - E\left(D_i \mid Z_i = 0\right)},$$

  which can be interpreted as a local average treatment effect (LATE) defined over the group of compliers, i.e. those induced to treat by a change in Z_i (Imbens and Angrist 1994).
- For example, Y_i=earnings, D_i=college education, and Z_i=college nearby. Compliers are individuals who would go to college if college is nearby but not if it is far away.
- Suppose there are two sampling strata (rural & urban). For each strata, we could get

$$\theta_s = \frac{E\left(Y_i \mid Z_i = 1, s\right) - E\left(Y_i \mid Z_i = 0, s\right)}{E\left(D_i \mid Z_i = 1, s\right) - E\left(D_i \mid Z_i = 0, s\right)}$$

- However, in this case the relevant aggregate effect is not $\overline{\theta}$ but θ_0. In this situation, weighted IV is essential to get a consistent estimation of θ_0 under stratified sampling.
- The θ_s are LATEs for different subpopulations of compliers.

Quantile treatment effects

- A further illustration involves quantile treatment effects (QTE), as used in the matching literature in program evaluation (Firpo, 2007).
- Let (Y_1, Y_0) be potential outcomes with marginal cdfs $F_1(r), F_0(r)$ and quantile functions $Q_{1\tau} = F_1^{-1}(\tau), Q_{0\tau} = F_0^{-1}(\tau)$. The QTE is defined to be

$$\theta_0 = Q_{1\tau} - Q_{0\tau}$$

- Under conditional exogeneity $F_j(r) = \int \Pr(Y \leq r \mid D = j, X)\, dG(X)$, $(j = 0, 1)$. Moreover, $Q_{1\tau}, Q_{0\tau}$ satisfy the moment conditions:

$$\begin{aligned} E\left[\frac{D}{p(X)} 1(Y \leq Q_{1\tau}) - \tau\right] &= 0 \\ E\left[\frac{1-D}{1-p(X)} 1(Y \leq Q_{0\tau}) - \tau\right] &= 0 \end{aligned}$$

  and

$$Q_{1\tau} = \arg\min_q E\left[\frac{D}{p(X)}\rho_\tau(Y-q)\right], Q_{0\tau} = \arg\min_q E\left[\frac{1-D}{1-p(X)}\rho_\tau(Y-q)\right].$$

- Firpo's method is a two-step weighting procedure in which the propensity score $p(X) = \Pr(D = 1 \mid X)$ is estimated first.
- The point here is that in the presence of stratified sampling we would like to introduce an additional layer of weights in both steps to undo stratification and obtain population-level quantities.

Choice-based sampling

- Choice-based sampling is a simple case of endogenous stratification in which y is discrete and its values define the strata.
- This is the context of the first econometric work on endogenous stratification.
- Manski and Lerman (1977) proposed an estimator for data on individual choices of transportation mode (travel to work by bus or by car), using choice based sampling.
- Their weighted estimator is consistent and computationally straightforward, but inefficient in an ML context such as this one.
- The problem is that while endogenous stratification is used to oversample the rare alternative, the Manski-Lerman estimator dilutes the information by assigning low weights to the extra observations (Coslett 1993).
- A GMM efficient estimator is proposed in Imbens (1992).

- In an exogenously stratified sample, the informative part of the likelihood is the density of y conditioned on x. In contrast, in a choice-based sample, it is the density of x conditioned on y, which is connected to the original model by Bayes' rule.

Binary model example

- Suppose that y and x are binary and we are interested in the logit model

$$\begin{aligned} \pi_1 &= \Pr(y = 1 \mid x = 1) = \Lambda(\alpha + \beta) \\ \pi_0 &= \Pr(y = 1 \mid x = 0) = \Lambda(\alpha) \end{aligned}$$

- We have a random sample of x of size N_1 for units with $y = 1$ and one of size N_0 for units with $y = 0$. The population value $q = \Pr(y = 1)$ is known.
- Thus, there are two strata with population frequencies q and $1 - q$, $N = N_1 + N_0$, and sample frequencies $h = N_1/N$ and $1 - h = N_0/N$.
- Using Bayes' rule

$$\pi_1 = \frac{\Pr(x = 1 \mid y = 1)\, q}{\Pr(x = 1 \mid y = 1)\, q + \Pr(x = 1 \mid y = 0)\,(1 - q)}$$

  with a similar expression for π_0.

- Intuitively, $\Pr(x = 1 \mid y = 1)$ can be consistently estimated in the first stratum and $\Pr(x = 1 \mid y = 0)$ in the second:

$$\begin{aligned} \widehat{\Pr}(x = 1 \mid y = 1) &= N_{11}/N_1 = \widehat{\phi}_1 \\ \widehat{\Pr}(x = 1 \mid y = 0) &= N_{01}/N_0 = \widehat{\phi}_0 \end{aligned}$$

- Using them to replace population frequencies, we obtain estimates of π_1 and π_0.
- To estimate (α, β), just solve $\alpha = \text{logit}(\pi_0)$ and $\beta = \text{logit}(\pi_1) - \text{logit}(\pi_0)$.
- For logit, knowledge of q is needed to determine α but not β.

Binary model example (continued)

- In this example weighted logit produces the same result. The first-order conditions are

$$\sum_{i=1}^{N} \frac{q_i}{h_i} \begin{pmatrix} x_i \\ 1 - x_i \end{pmatrix} [y_i - \Lambda(\alpha + \beta x_i)] = 0$$

  or

$$\sum_{y=1}^{N_1} \frac{q}{N_1/N} \begin{pmatrix} x_i \\ 1 - x_i \end{pmatrix} [1 - \Lambda(\alpha + \beta x_i)] - \sum_{y=0}^{N_0} \frac{1-q}{N_0/N} \begin{pmatrix} x_i \\ 1 - x_i \end{pmatrix} \Lambda(\alpha + \beta x_i) = 0.$$

- Collecting terms we get:

$$\begin{aligned} q\widehat{\phi}_1 [1 - \Lambda(\alpha + \beta)] - (1-q)\widehat{\phi}_0 \Lambda(\alpha + \beta) &= 0 \\ q\left(1 - \widehat{\phi}_1\right)[1 - \Lambda(\alpha)] - (1-q)\left(1 - \widehat{\phi}_0\right)\Lambda(\alpha) &= 0 \end{aligned}$$

- Solving the first equation we get an estimate of π_1

$$\widehat{\Lambda(\alpha + \beta)} = \widehat{\pi}_1 = \frac{\widehat{\phi}_1 q}{\widehat{\phi}_1 q + \widehat{\phi}_0 (1-q)}$$

  and solving the second we get $\widehat{\Lambda(\alpha)} = \widehat{\pi}_0$.
- In more general cases, like multinomial models or models with continuous x's, the direct approach may not be feasible while weighted logit remains a simple method.
- The example is intended to illustrate the nature of identification under choice-based sampling and the connection with weighted estimation.

Discrete choice from data on participants

- Bover and Arellano (2002) studied the determinants of migration using choice-based Census data on residential variations (RV) that only had characteristics of migrants.

- The difficulty in this case is that $\Pr(x = 1 \mid y = 0)$ is not identified, so that the model parameters are not identified without further information.

- The extra information is a complementary sample from labor force surveys (LFS), which contains characteristics of migrants and non-migrants but no migration variables. From these data the marginal distribution of x can be estimated.

- The goal is to estimate multinomial logit probabilities of migrating to small, medium and large towns, relying on Bayes' rule:

$$\Pr(y = j \mid x) = \frac{f(x \mid y = j)\Pr(y = j)}{f(x)} \qquad (j = 1, 2, 3)$$

- The RV data is informative about $f(x \mid y = j)$ and the LFS data about $f(x)$.

- Given RV data, LFS data are indirectly informative on nonmigrant characteristics $f(x \mid y = 0)$. Information on $p_j = \Pr(y = j)$ comes from aggregate census.

Discrete choice from data on participants (continued)

- Two sources of micro data and one source of macro data are combined to secure identification of the multinomial choice model.

- Since the x's used in Bover and Arellano were all discrete, they treated $f\left(x \mid y = j\right)$ and $f\left(x\right)$ as multinomial distributions.

- They formed cells for each value of x. Then they calculated cell by cell frequencies of migration $\Pr\left(y = j \mid x\right)$ combining the various data sources. Finally, the discrete choice models were fitted to cell-specific frequencies by minimum distance.

Continuous regressors

- If x's are continuous the previous method does not work. Here we consider a GMM alternative that works with continuous regressors. The multinomial logit model is

$$\Pr(y = j \mid x) = G_j(x, \theta) = \frac{e^{x'\theta_j}}{1 + e^{x'\theta_1} + e^{x'\theta_2} + e^{x'\theta_3}}$$

- Since we have

$$G_j(x, \theta) f(x) = p_j f(x \mid y = j),$$

  the following holds for any function $h(x)$:

$$\int h(x) G_j(x, \theta) f(x) dx = p_j \int h(x) f(x \mid y = j) dx$$

- Setting $h(x) = x$, we have

$$E\left[xG_j(x, \theta)\right] = p_j E[x \mid y = j].$$

- Let the LFS data be $\{x_k\}_{k=1}^M$ and the RV data $\{y_i, x_i\}_{i=1}^N$. Estimation of θ is based on the two-sample moment conditions

$$\frac{1}{M} \sum_{k=1}^{M} x_k G_j(x_k, \theta) = p_j \frac{1}{N_j} \sum_{i=1}^{N} x_i 1(y_i = j) \qquad (j = 1, 2, 3)$$

  where $N_j = \sum_{i=1}^N 1(y_i = j)$.

Continuous regressors (continued)

- To see what is going on, consider the simplified situation where the model is binary ($y = 1$ if mover, $y = 0$ if stayer) and we are using a linear probability model. The previous moment equation boils down to

$$\frac{1}{M}\sum_{k=1}^{M} x_k x_k'\theta = p\frac{1}{N}\sum_{i=1}^{N} x_i$$

- The estimator is just the following two-sample least squares method

$$\widehat{\theta} = p\left(\frac{1}{M}\sum_{k=1}^{M} x_k x_k'\right)^{-1}\frac{1}{N}\sum_{i=1}^{N} x_i$$

or

$$\widehat{\theta} = p\left[\widehat{E}_{LFS}\left(xx'\right)\right]^{-1}\widehat{E}_{RV}\left[x \mid y = 1\right].$$

Part 2: Cluster standard errors

Cluster standard errors

Introduction

- We have already discussed the consequences of clustered designs for the estimation of means and their standard errors. We now deal with the implications for estimation error in econometric work more broadly.

- The basic conclusion remains. Namely, that if the cluster design is ignored, conventional formulas for standard errors of estimated regression parameters and their generalizations are too small.

- Goal is to find ways of obtaining valid standard errors when using clustered samples.

- Moulton (1990) alerted economists about large underestimation of standard errors due to ignoring cluster effects.

- In an individual wage equation for the US with only state level x's, ignoring cluster effects would understate standard errors by a factor of more than three (Deaton 1997)

Remarks

- There is a close parallel between the econometrics of clustered samples and "large N, small T" panel data (Arellano, 2003). However, in the context of clusters, unbalanced designs are the norm (varying number of units per cluster).

- For example, in a household-level panel data set, the household plays the role of a cluster and time plays the role of the individual units.

- There may be also cluster samples of panel data. In those situations there are two potential sources of correlation across observations: across time within the same individual and across individuals within the same cluster (see Wooldridge 2010).

- In the econometrics of clustered samples there is also an efficiency issue, although less prominent in applied work. The point is that because the error terms in the regressions are correlated across observations, OLS is not efficient.

Cluster fixed-effects vs cluster-robust standard errors

- Sometimes within-cluster dependence is specified as an additive cluster effect in the error of a model. In this case there is constant within-cluster correlation among errors.
- If cluster effects are correlated with regressors, within-cluster LS (cluster "fixed effects") is consistent but OLS in levels is not.
- If cluster effects are uncorrelated with regressors, OLS in levels is consistent but standard errors need to be adjusted.
- Assuming constant within-cluster dependence, one can use formulas for standard errors that are more restrictive than the fully robust formulas discussed below, but may have better sampling properties when the number of clusters is small.
- Sometimes cluster fixed-effects are used for consistency together with cluster-robust standard errors for correcting any remaining non-constant within-cluster correlation.

General set up

- Let θ be a parameter identified from the moment condition

$$E\psi\left(w,\theta\right)=0$$

such that $\dim\left(\theta\right)=\dim\left(\psi\right)$. Consider a sample $\mathcal{W}=\left\{w_1,...,w_N\right\}$ and a consistent estimator $\widehat{\theta}$ that solves the sample moment conditions.

- $\psi\left(w,\theta\right)$ are GMM orthogonality conditions or first-order conditions in M-estimation.
- Assume that conditions for the large-sample linear representation of the scaled estimation error and asymptotic normality hold:

$$\sqrt{N}\left(\widehat{\theta}-\theta\right)\approx-D_0^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi\left(w_i,\theta\right)\xrightarrow{d}\mathcal{N}\left(0,D_0^{-1}V_0D_0^{-1\prime}\right)$$

where $D_0=\partial E\psi\left(w,\theta\right)/\partial c'$ are partial derivatives and V_0 is the limiting variance:

$$V_0=\lim_{N\to\infty}Var\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi\left(w_i,\theta\right)\right)$$

- To get standard errors of $\widehat{\theta}$ we need estimates of D_0 and V_0.
- A natural estimate of D_0 is to use numerical derivatives $\widehat{D}$ evaluated at $\widehat{\theta}$.
- As for V_0 the situation is different under independent or cluster sampling.

Independent sampling

- Let $\psi_i = \psi(w_i, \theta)$, $\widehat{\psi}_i = \psi\left(w_i, \widehat{\theta}\right)$. With iid observations, $E\left(\psi_i \psi_j'\right) = 0$ $(i \neq j)$ and

$$V_0 = E\left(\psi_i \psi_i'\right)$$

- A consistent estimate of V_0 under independence is therefore

$$\widehat{V} = \frac{1}{N} \sum_{i=1}^{N} \widehat{\psi}_i \widehat{\psi}_i'$$

- and a consistent estimate of the asymptotic variance of $\widehat{\theta}$ under independent sampling:

$$\widehat{Var}\left(\widehat{\theta}\right) = \frac{1}{N} \widehat{D}^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{\psi}_i \widehat{\psi}_i' \right) \widehat{D}^{-1}. \tag{3}$$

- As an illustration, if $\psi_i = y_i - \theta$, $D_0 = -1$, $V_0 = E\left[(y_i - \theta)^2\right] \equiv \omega_v$, and

$$\widehat{Var}\left(\widehat{\theta}\right) = \frac{\widehat{\omega}_v}{N}, \qquad \widehat{\omega}_v = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \widehat{\theta}\right)^2$$

Cluster sampling

- The sample consists of H groups (or clusters) of M_h observations each ($N = M_1 + ... + M_H$), such that observations are independent across groups but dependent within groups, $H \to \infty$ and M_h is fixed for all h.
- For convenience we order observations by groups and use the double-index notation w_{hm} so that $\mathcal{W} = \{w_{11}, ..., w_{1M_1} \mid ... \mid w_{H1}, ..., w_{HM_H}\}$.
- Under cluster sampling, letting $\overline{\psi}_h = \sum_{m=1}^{M_h} \psi_{hm}$ and $\widetilde{\psi}_h = \sum_{m=1}^{M_h} \widehat{\psi}_{hm}$ we have

$$V_0 = \lim_{N\to\infty} Var\left(\frac{1}{\sqrt{N}} \sum_{h=1}^{H} \overline{\psi}_h\right) = \lim_{N\to\infty} \frac{1}{N} \sum_{h=1}^{H} E\left(\overline{\psi}_h \overline{\psi}_h'\right),$$

  so that $E\left(\psi_i \psi_j'\right) = 0$ only if i and j belong to different clusters, or $E\left(\psi_{hm} \psi_{h'm'}'\right) = 0$ for $h \neq h'$. Thus, a consistent estimate of V_0 is

$$\widetilde{V} = \frac{1}{N} \sum_{h=1}^{H} \widetilde{\psi}_h \widetilde{\psi}_h'.$$

- The estimated asymptotic variance of $\widehat{\theta}$ allowing within-cluster correlation is

$$\widetilde{Var}\left(\widehat{\theta}\right) = \frac{1}{N} \widehat{D}^{-1} \left(\frac{1}{N} \sum_{h=1}^{H} \widetilde{\psi}_h \widetilde{\psi}_h'\right) \widehat{D}^{-1}. \tag{4}$$

- Note that (4) is of order $1/H$ whereas (3) is of order $1/N$.

Illustration

- Consider the simple case of $\psi_{hm} = y_{hm} - \theta$ and a balanced clustered design $M_h = M$ for all h (so that $N = HM$).
- Letting $E\left[(y_{hm} - \theta)(y_{hs} - \theta)\right] = \omega_\eta$ and $E\left[(y_{hm} - \theta)^2\right] = \omega_\eta + \omega_v$, we have

$$V_0 = \frac{H}{HM} E\left(\overline{\psi}_h^2\right) = \frac{1}{M}\left(M\omega_v + M^2\omega_\eta\right) = \omega_v + M\omega_\eta$$

$$\widetilde{V} = \frac{1}{HM}\sum_{h=1}^{H} \widetilde{\psi}_h^2 = \widehat{\omega}_v + M\widehat{\omega}_\eta$$

where $\widehat{\omega}_v$ is the within-cluster sample variance

$$\widehat{\omega}_v = \frac{1}{H(M-1)}\sum_{h=1}^{H}\sum_{m=1}^{M}(y_{hm} - \overline{y}_h)^2,$$

$\overline{y}_h = \widetilde{\psi}_h / M$, and $\widehat{\omega}_\eta$ is the cluster-effect variance

$$\widehat{\omega}_\eta = \frac{1}{H}\sum_{h=1}^{H}\overline{y}_h^2 - \frac{\widehat{\omega}_v}{M}$$

- Therefore,

$$\widetilde{Var}\left(\widehat{\theta}\right) = \frac{1}{N}\left(\widehat{\omega}_v + M\widehat{\omega}_\eta\right) = \frac{\widehat{\omega}_v}{N} + \frac{\widehat{\omega}_\eta}{H}$$

- The first term is of order $1/N$ but the second is $1/H$.

Examples

Panel data model

- A panel data example is

$$\psi\left(w_{hm},\theta\right)=x_{hm}\left(y_{hm}-x_{hm}'\theta\right)$$

  where h denotes units and m time periods. If the variables are in deviations from means, $\widehat{\theta}$ is the within-group estimator.
- In this case $\widehat{D}_{hm}=-x_{hm}x_{hm}'$ and $\widetilde{\psi}_h=\sum_{m=1}^{M_h}x_{hm}\widehat{u}_{hm}$ where $\widehat{u}_{hm}$ are within-group residuals.
- The result is the (large H, fixed M_h) formula for within-group standard errors that are robust to heteroskedasticity and serial correlation of arbitrary form in Arellano (1987):

$$\widetilde{Var}\left(\widehat{\theta}\right)=\left(\sum_{h=1}^{H}\sum_{m=1}^{M_h}x_{hm}x_{hm}'\right)^{-1}\sum_{h=1}^{H}\sum_{m=1}^{M_h}\sum_{s=1}^{M_h}\widehat{u}_{hm}\widehat{u}_{hs}x_{hm}x_{hs}'\left(\sum_{h=1}^{H}\sum_{m=1}^{M_h}x_{hm}x_{hm}'\right)^{-1}$$

Examples (continued)

Quantile regression

- Another example is a τ-quantile regression with moments

$$\psi\left(w_{hm},\theta\right)=x_{hm}\left[1\left(y_{hm}\leq x_{hm}'\theta\right)-\tau\right]$$

where $\widehat{D}_{hm}=\widehat{\omega}_{hm}x_{hm}x_{hm}'$, $\widehat{\omega}_{hm}=\mathbf{1}\left(\left|y_{hm}-x_{hm}'\widehat{\theta}\right|\leq\xi_N\right)/\left(2\xi_N\right)$, $\widetilde{\psi}_h=\sum_{m=1}^{M_h}x_{hm}\widehat{u}_{hm}$ and $\widehat{u}_{hm}=1\left(y_{hm}\leq x_{hm}'\widehat{\theta}\right)-\tau$.

- Cluster-robust standard errors for QR coefficients are obtained from:

$$\widehat{Var}\left(\widehat{\theta}\right)=\widehat{B}^{-1}\left(\sum_{h=1}^{H}\sum_{m=1}^{M_h}\sum_{s=1}^{M_h}\widehat{u}_{hm}\widehat{u}_{hs}x_{hm}x_{hs}'\right)\widehat{B}^{-1}$$

where

$$\widehat{B}=\sum_{h=1}^{H}\sum_{m=1}^{M_h}\widehat{\omega}_{hm}x_{hm}x_{hm}'$$

Unconditional quantile

- Estimation of an unconditional τ-quantile is the same as QR on an intercept and therefore a special case of the previous result:

$$\psi\left(w_{hm}, \theta\right) = \left[1\left(y_{hm} \leq \theta\right) - \tau\right]$$

where $\widehat{\omega}_{hm} = \mathbf{1}\left(\left|y_{hm} - \widehat{\theta}\right| \leq \xi_N\right) / \left(2\xi_N\right)$, $\widetilde{\psi}_h = \sum_{m=1}^{M_h} \widehat{u}_{hm}$ and $\widehat{u}_{hm} = 1\left(y_{hm} \leq \widehat{\theta}\right) - \tau$.

- Cluster-robust standard errors for sample τ-quantiles are obtained from:

$$\widetilde{Var}\left(\widehat{\theta}\right) = \frac{\sum_{h=1}^{H} \sum_{m=1}^{M_h} \sum_{s=1}^{M_h} \left[1\left(y_{hm} \leq \widehat{\theta}\right) - \tau\right]\left[1\left(y_{hs} \leq \widehat{\theta}\right) - \tau\right]}{\left(\sum_{h=1}^{H} \sum_{m=1}^{M_h} \frac{\mathbf{1}\left(\left|y_{hm} - \widehat{\theta}\right| \leq \xi_N\right)}{2\xi_N}\right)^2}$$

Part 3: Bootstrap methods

Bootstrap methods

Introduction

- The bootstrap is an alternative method of assessing sampling variability. It is a mechanical procedure that can be applied in a wide variety of situations.

- It works in the same way regardless of whether something straightforward is being estimated or something more complex.

- The bootstrap was invented and given its name by Brad Efron in a paper published in 1979 in the *Annals of Statistics*.

- The bootstrap is probably the most widely used methodological innovation in statistics since Ronald Fisher's development of the analysis of variance in the 1920s (Erich Lehmann 2008).

The idea of the bootstrap

- Let $Y_1, ..., Y_N$ be a random sample according to some distribution F and let $\widehat{\theta}_N = h(Y_1, ..., Y_N)$ be some statistic of interest. We want to estimate the distribution of $\widehat{\theta}_N$

$$\Pr_F\left(\widehat{\theta}_N \leq r\right) = \Pr_F\left[h(Y_1, ..., Y_N) \leq r\right]$$

  where the subscript F indicates the distribution of the Y's.
- A simple estimator of $\Pr_F\left(\widehat{\theta}_N \leq r\right)$ is the plug-in estimator. It replaces F by the empirical cdf $\widehat{F}_N$:

$$\widehat{F}_N(s) = \frac{1}{N}\sum_{i=1}^{N} 1(Y_i \leq s),$$

  which assigns probability $1/N$ to each of the observed values $y_1, ..., y_N$ of $Y_1, ..., Y_N$.
- The resulting estimator is then

$$\Pr_{\widehat{F}_N}\left[h(Y_1^*, ..., Y_N^*) \leq r\right] \tag{5}$$

  where $Y_1^*, ..., Y_N^*$ denotes a random sample from $\widehat{F}_N$.
- The formula (5) for the estimator of the cdf of $\widehat{\theta}_N$ is easy to to write down, but it is prohibitive to calculate except for small N.
- To see this note that each of the Y_i^* is capable of taking on the N values $y_1, ..., y_N$, so that the total number of values of $h\left(Y_1^*, ..., Y_N^*\right)$ that has to be considered is N^N. To calculate (5), one would have to count how many of these N^N values are $\leq r$.

The idea of the bootstrap (continued)

- For example, suppose that Y_i is binary, F is given by $\Pr(Y = 1) = p$, $\widehat{\theta}_N$ is the sample mean and $N = 3$. There are eight possible samples:

y_1, y_2, y_3	$\Pr(y_1, y_2, y_3)$	$\widehat{\theta}_N$
$(0,0,0)$	$(1-p)^3$	0
$(1,0,0)$	$p(1-p)^2$	$1/3$
$(0,1,0)$	$p(1-p)^2$	$1/3$
$(0,0,1)$	$p(1-p)^2$	$1/3$
$(1,1,0)$	$p^2(1-p)$	$2/3$
$(1,0,1)$	$p^2(1-p)$	$2/3$
$(0,1,1)$	$p^2(1-p)$	$2/3$
$(1,1,1)$	p^3	1

- So that $\Pr_F\left(\widehat{\theta}_N \leq r\right)$ is determined by

r	$\Pr_F\left(\widehat{\theta}_N = r\right)$
0	$(1-p)^3$
$1/3$	$3p(1-p)^2$
$2/3$	$3p^2(1-p)$
1	p^3

The idea of the bootstrap (continued)

- Suppose that the observed values y_1, y_2, y_3 are $(0, 1, 1)$, so that the observed value of $\widehat{\theta}_N$ is $2/3$. Therefore, our estimate of $\Pr\left(\widehat{\theta}_N = r\right)$ is given by

r	$\Pr_{\widehat{F}_N}\left(\widehat{\theta}_N = r\right)$
0	$\left(1-\frac{2}{3}\right)^3 = \frac{1}{27} = .037$
$1/3$	$3\frac{2}{3}\left(1-\frac{2}{3}\right)^2 = \frac{6}{27} = .222$
$2/3$	$3\left(\frac{2}{3}\right)^2\left(1-\frac{2}{3}\right) = \frac{12}{27} = .444$
1	$\left(\frac{2}{3}\right)^3 = \frac{8}{27} = .297$

- The previous example is so simple that the calculation of $\Pr_{\widehat{F}_N}\left(\widehat{\theta}_N \leq r\right)$ can be done analytically, but in general this type of calculation is beyond reach.

The idea of the bootstrap (continued)

Estimation by simulation

- A standard device for (approximately) evaluating probabilities that are too difficult to calculate exactly is simulation.
- To calculate the probability of an event, one generates a sample from the underlying distribution and notes the frequency with which the event occurs in the generated sample.
- If the sample is sufficiently large, this frequency will provide an approximation to the original probability with a negligible error.
- Such approximation to the probability (5) constitutes the second step of the bootstrap method.
- A number M of samples $Y_1^*, ..., Y_N^*$ (the "bootstrap" samples) are drawn from $\widehat{F}_N$, and the frequency with which

$$h\left(Y_1^*, ..., Y_N^*\right) \leq r$$

provides the desired approximation to the estimator (5) (Lehmann 2008).

Numerical illustration

- To illustrate the method I have generated $M = 1000$ bootstrap samples of size $N = 3$ with $p = 2/3$, which is the value of p that corresponds to the sample distribution $(0, 1, 1)$. The result is

	r	#samples	bootstrap pdf	$\Pr_{\widehat{F}_N}\left(\widehat{\theta}_N = r\right)$
$(0,0,0)$	0	37	.037	.037
$(1,0,0), (0,1,0), (0,0,1)$	1/3	222	.222	.222
$(1,1,0), (1,0,1), (0,1,1)$	2/3	453	.453	.444
$(1,1,1)$	1	288	.288	.297

- The discrepancy between the last two columns can be made arbitrarily small by increasing M.

- The method we have described consisting in drawing random samples with replacement treating the observed sample as the population is called "non-parametric bootstrap".

Bootstrap standard errors

- The bootstrap procedure is very flexible and applicable to many different situations such as the bias and variance of an estimator, to the calculation of confidence intervals, etc.

- As a result of resampling we have available M estimates from the artificial samples: $\widehat{\theta}_N^{(1)}, ..., \widehat{\theta}_N^{(M)}$. A *bootstrap standard error* is then obtained as

$$\left[\frac{1}{M-1} \sum_{m=1}^{M} \left(\widehat{\theta}_N^{(m)} - \overline{\widehat{\theta}_N} \right)^2 \right]^{1/2} \tag{6}$$

  where $\overline{\widehat{\theta}_N} = \sum_{m=1}^{M} \widehat{\theta}_N^{(m)} / M$.

- In the previous example, the bootstrap mean is $\overline{\widehat{\theta}_N} = 0.664$, the bootstrap standard error is 0.271 calculated as in (6) with $M = 1000$, and the analytical standard error is

$$\left[\frac{\widehat{\theta}_N \left(1 - \widehat{\theta}_N \right)}{n} \right]^{1/2} = 0.272$$

  where $\widehat{\theta}_N = 2/3$ and $n = 3$.

Asymptotic properties of bootstrap methods

- Using the bootstrap standard error to construct test statistics cannot be shown to improve on the approximation provided by the usual asymptotic theory, but the good news is that under general regularity conditions it does have the same asymptotic justification as conventional asymptotic procedures.
- This is good news because bootstrap standard errors are often much easier to obtain than analytical standard errors.

Refinements for large-sample pivotal statistics

- Even better news is the fact that in many cases the bootstrap does improve the approximation of the distribution of test statistics, in the sense that the bootstrap can provide an asymptotic refinement compared with the usual asymptotic theory.
- The key aspect for achieving such refinements (consisting in having an asymptotic approximation with errors of a smaller order of magnitude in powers of the sample size) is that the statistic being bootstraped is *asymptotically pivotal*.
- An asymptotically pivotal statistic is one whose limiting distribution does not depend on unknown parameters (like standard normal or chi-square distributions).
- This is the case with t-ratios and Wald test statistics, for example.
- Note that for a t-ratio to be asymptotically pivotal in a regression with heteroskedasticity, the robust White form of the t-ratio needs to be used.

Asymptotic properties of bootstrap methods (continued)

- The upshot of the previous discussion is that the replication of the quantity of interest (mean, median, etc.) is not always the best way to use the bootstrap if improvements on asymptotic approximations are sought.
- In particular, when we wish to calculate a confidence interval, it is better not to bootstrap the estimate itself but rather to bootstrap the distribution of the t-value.
- This is feasible when we have a large sample estimate of the standard error, but one is skeptical about the accuracy of the normal probability approximation.
- Such procedure will provide more accurate estimates of confidence intervals than either the simple bootstrap or the asymptotic standard errors.

- However, often we are interested in bootstrap methods because an analytic standard error is not available or is hard to calculate.
- In those cases the motivation for the bootstrap is not necessarily improving on the asymptotic approximation but rather obtaining a simpler approximation with the same justification as the conventional asymptotic approximation.

Bootstrapping stratified clustered samples

- The bootstrap can be applied to a stratified clustered sample.
- All we have to do is to treat the strata separately, and resample, not the basic underlying units (the households) but rather the primary sample units (the clusters).

Using replicate weights

- Taking stratification and clustering sampling features into account, either analytically or by bootstrap, requires the availability of stratum and cluster indicators.

- Generally, Statistical Offices or survey providers do not make them available for confidentiality reasons.

- To enable the estimation of the sampling distribution of estimators and test statistics without disclosing stratum or cluster information, an alternative is to provide replicate weights.

- For example, the EFF provides 999 replicate weights. Specifically the EFF provides replicate cross-section weights, replicate longitudinal weights, and multiplicity factors (Bover, 2004).
- Multiplicity factors indicate the number of times a given household appears in a particular bootstrap sample.

- The provision of replicate weights is an important development because it facilitates the use of replication methods, which are simple and of general applicability, together with allowing for confidentiality safe ward.

References

- Arellano, M. (1987): "Computing Robust Standard Errors for Within-Group Estimators", *Oxford Bulletin of Economics and Statistics*, 49, 431-434.
- Arellano, M. (2003): *Panel Data Econometrics*, Oxford University Press.
- Bover, O. and M. Arellano (2002): "Learning About Migration Decisions From the Migrants," *Journal of Population Economics*, 15, 357-380.
- Bover, O. (2004), "The Spanish Survey of Household Finances (EFF): Description and Methods of the 2002 Wave", Banco de España Occasional Paper 0409.
- Deaton, A. (1997): *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Johns Hopkins Press.
- Cameron, C. and P. Trivedi (2005): *Microeconometrics*, Cambridge University Press.
- Coslett, S. R. (1993): "Estimation from Endogenously Stratified Samples," in G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.): *Handbook of Statistics*, Vol. 11, North Holland.
- Efron, B. (1979): "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1–26.
- Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75, 259–276.

References (continued)

- Imbens, G. (1992): "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling," *Econometrica*, 60, 1187–1214.
- Imbens, G. W. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, 467-475.
- Imbens, G. and T. Lancaster (1996): "Efficient Estimation and Stratified Sampling," *Journal of Econometrics*, 74, 289–318.
- Levy, P. S. and S. Lemeshow (1991): *Sampling of Populations: Methods and Applications*, 2nd edition, Wiley.
- Manski, C. F. and S. R. Lerman (1977): "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica*, 45, 1977-1988.
- Moulton, B. R. (1990): "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics*, 72, 334-338
- Ullah, A. and R. V. Breunig (1998): "Econometric Analysis in Complex Surveys," in D. Giles and A. Ullah (eds.): *Handbook of Applied Economic Statistics*, Marcel Dekker.
- Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, MIT Press.