New Developments in Econometrics
Lecture 9: Stratified Sampling

Jeff Wooldridge

Cemmap Lectures, UCL, June 2009

1. Overview of Stratified Sampling
2. Regression Analysis
3. Clustering and Stratification

1. The Basic Methodology

- Typically, with stratified sampling, some segments of the population are over- or underrepresented by the sampling scheme. If we know enough information about the stratification scheme, we can modify standard econometric methods and consistently estimate population parameters.
- There are two common types of stratified sampling, standard stratified (SS) sampling and variable probability (VP) sampling. A third type of sampling, typically called multinomial sampling, is practically indistinguishable from SS sampling, but it generates a random sample from a modified population.

• **SS Sampling**: Partition the sample space, say W, into G non-overlapping, exhaustive groups, $\{W_g : g = 1, \ldots G\}$. Random sample is taken from each group g, say $\{w_{gi} : i = 1, \ldots, N_g\}$, where N_g is the number of observations drawn from stratum g and $N = N_1 + N_2 + \ldots + N_G$ is the total number of observations.

• Let w be a random vector representing the population. Each each random draw from stratum g has the same distribution as w conditional on w belonging to W_g:

$$D(w_{gi}) = D(w|w \in W_g), i = 1, \ldots, N_g. \quad (1)$$

We only know we have an SS sample if we are told.

• What if we want to estimate the mean of w from an SS sample? Let $\pi_g = P(w \in W_g)$ be the probability that w falls into stratum g; the π_g are often called the "aggregate shares." If we know the π_g (or can consistently estimate them), then $\mu_w = E(w)$ is identified by a weighted average of the expected values for the strata:

$$\mu_w = \pi_1 E(w|w \in W_1) + \ldots + \pi_G E(w|w \in W_G). \quad (2)$$

So an unbiased estimator is

$$\hat{\mu}_w = \pi_1 \bar{w}_1 + \pi_2 \bar{w}_2 \ldots + \pi_G \bar{w}_G, \quad (3)$$

where $\bar{w}_g$ is the sample average from stratum g.

• As the strata sample sizes grow, $\hat{\mu}_w$ is also a consistent estimator of μ_w. Also,

$$Var(\hat{\mu}_w) = \pi_1^2 Var(\bar{w}_1) + \ldots + \pi_G^2 Var(\bar{w}_G). \quad (4)$$

• Because $Var(\bar{w}_g) = \sigma_g^2/N_g$, each of the variances can be estimated in an unbiased fashion by using the usual unbiased variance estimator,

$$\hat{\sigma}_g^2 = (N_g - 1)^{-1} \sum_{i=1}^{N_g} (w_{gi} - \bar{w}_g)^2 \quad (5)$$

and

$$se(\hat{\mu}_w) = [\pi_1^2 \hat{\sigma}_1^2/N_1 + \ldots + \pi_G^2 \hat{\sigma}_G^2/N_G]^{1/2}. \quad (6)$$

• Useful to have a formula for $\hat{\mu}_w$ as a weighted average across all observations:

$$\hat{\mu}_w = (\pi_1/h_1)N^{-1}\sum_{i=1}^{N_1} w_{1i} + \ldots + (\pi_G/h_G)N^{-1}\sum_{i=1}^{N_G} w_{Gi}$$

$$= N^{-1}\sum_{i=1}^{N}(\pi_{g_i}/h_{g_i})w_i \qquad (7)$$

where $h_g = N_g/N$ is the fraction of observations in stratum g and in (7) we drop the strata index on the observations.

• **Variable Probability Sampling**: Often used where little, if anything, is known about respondents ahead of time. Still partition the sample space, but an observation is drawn at random. However, if the observation falls into stratum g, it is kept with (nonzero) sampling probability, p_g. That is, random draw w_i is kept with probability p_g if $w_i \in W_g$.

• The population is sampled N times (often N is not reported with VP samples). We always know how many data points were kept; call this M – a random variable. Let s_i be a selection indicator, equal to one if observation i is kept. So $M = \sum_{i=1}^{N} s_i$.

• Let $\mathbf{z}_i$ be a G-vector of stratum indicators for draw i, so

$$p(\mathbf{z}_i) = p_1 z_{i1} + \ldots + p_G z_{iG} \tag{8}$$

is the function that delivers the sampling probability for any random draw i.

• Key assumption for VP sampling: Conditional on being in stratum g, the chance of keeping an observation is p_g. Statistically, conditional on $\mathbf{z}_i$ (knowing the stratum), s_i and w_i are independent. Then

$$E[(s_i/p(\mathbf{z}_i))w_i] = E(w_i). \tag{9}$$

• Equation (9) is the key result for VP sampling. It says that weighting a selected observation by the inverse of its sampling probability allows us to recover the population mean. Therefore,

$$N^{-1}\sum_{i=1}^{N}(s_i/p(\mathbf{z}_i))w_i \tag{10}$$

is a consistent estimator of $E(w_i)$. We can also write (10) as

$$(M/N)M^{-1}\sum_{i=1}^{N}(s_i/p(\mathbf{z}_i))w_i. \tag{11}$$

If we define weights as $\hat{v}_i = \hat{\rho}/p(\mathbf{z}_i)$ where $\hat{\rho} = M/N$ is the fraction of observations retained from the sampling scheme, then (11) is

$$M^{-1} \sum_{i=1}^{M} \hat{v}_i w_i, \tag{12}$$

where only the observed points are included in the sum.

- So, can write estimator as a weighted average of the observed data points. If $p_g < \hat{\rho}$, the observations for stratum g are underpresented in the eventual sample (asymptotically), and they receive weight greater than one.

2. **Regression Analysis**

- Almost any estimation method can be used with SS or VP sampled data: IV, MLE, quasi-MLE, nonlinear least squares.
- Linear population model:

$$y = \mathbf{x}\boldsymbol{\beta} + u. \tag{13}$$

Two assumptions on u are

$$E(u|\mathbf{x}) = 0 \tag{14}$$

$$E(\mathbf{x}'u) = \mathbf{0}. \tag{15}$$

(15) is enough for consistency, but (14) has important implications for whether or not to weight.

• SS Sampling: A consistent estimator $\hat{\boldsymbol{\beta}}$ is obtained from the "weighted" least squares problem

$$\min_{\mathbf{b}} \sum_{i=1}^{N} v_i (y_i - \mathbf{x}_i \mathbf{b})^2, \tag{16}$$

where $v_i = \pi_{g_i}/h_{g_i}$ is the weight for observation i. (Remember, the weighting used here is not to solve any heteroskedasticity problem; it is to reweight the sample in order to consistently estimate the population parameter $\boldsymbol{\beta}$.)

• Key Question: How can we conduct valid inference using $\hat{\boldsymbol{\beta}}$? One possibility: use the White (1980) "heteroskedasticity-robust" sandwich estimator. When is this estimator the correct one? If two conditions hold: (i) $E(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$, so that we are actually estimating a conditional mean; and (ii) the strata are determined by the explanatory variables, $\mathbf{x}$.

• When the White estimator is not consistent, it is conservative.

• Correct asymptotic variance requires more detailed formulation of the estimation problem:

$$\min_{\mathbf{b}} \left\{ \sum_{g=1}^{G} \pi_g \left[N_g^{-1} \sum_{i=1}^{N_g} (y_{gi} - \mathbf{x}_{gi}\mathbf{b})^2 \right] \right\}. \tag{17}$$

Asymptotic variance estimator:

$$\left[\sum_{i=1}^{N}(\pi_{g_i}/h_{g_i})\mathbf{x}_i'\mathbf{x}_i\right]^{-1}$$
$$\cdot\left\{\sum_{g=1}^{G}(\pi_g/h_g)^2\left[\sum_{i=1}^{N_g}(\mathbf{x}_{gi}'\hat{u}_{gi}-\overline{\mathbf{x}_g'\hat{u}_g})(\mathbf{x}_{gi}'\hat{u}_{gi}-\overline{\mathbf{x}_g'\hat{u}_g})'\right]\right\} \qquad (18)$$
$$\cdot\left[\sum_{i=1}^{N}(\pi_{g_i}/h_{g_i})\mathbf{x}_i'\mathbf{x}_i\right]^{-1}.$$

• Usual White estimator ignores the information on the strata of the observations, which is the same as dropping the within-stratum averages, $\overline{\mathbf{x}_g' \hat{u}_g}$. The estimate in (18) is always *smaller* than the usual White estimate.

• Econometrics packages such as Stata have survey sampling options that will compute (18) provided stratum membership is included along with the weights. If only the weights are provided, the larger asymptotic variance is computed.

- One case where there is no gain from subtracting within-strata means is when $E(u|\mathbf{x}) = 0$ and stratification is based on $\mathbf{x}$.
- If we add the homoskedasticity assumption $Var(u|\mathbf{x}) = \sigma^2$ with $E(u|\mathbf{x}) = 0$ and stratification is based on $\mathbf{x}$, the weighted estmator is less efficient than the unweighted estimator. (Both are consistent.)

• The debate about whether or not to weight centers on two facts: (i) The efficiency loss of weighting when the population model satisfies the classical linear model assumptions and stratification is exogenous. (ii) The failure of the unweighted estimator to consistently estimate $\boldsymbol{\beta}$ if we only assume a linear projection in the population,

$$y = \mathbf{x}\boldsymbol{\beta} + u, E(\mathbf{x}'u) = \mathbf{0}, \tag{19}$$

even when stratification is based on $\mathbf{x}$. The weighted estimator consistently estimates $\boldsymbol{\beta}$ under (19).

• Analogous results hold for maximum likelihood, quasi-MLE, nonlinear least squares, instrumental variables. If one knows stratum identification along with the weights, the appropriate asymptotic variance matrix (which subtracts off within-stratum means of the score of the objective function) is smaller than the form derived by White (1982). For, say, MLE, if the density of y given $\mathbf{x}$ is correctly specified, and stratification is based on $\mathbf{x}$, it is better not to weight. (But there are cases – including certain treatment effect estimators – where it is important to estimate the solution to a misspecified population problem.)

• Findings for SS sampling have analogs for VP sampling, and some additional results. First, the Huber-White sandwich matrix applied to the weighted objective function (weighted by the $1/p_g$) *is* consistent when the *known* p_g are used. Second, an asymptotically more efficient estimator is available when the retention frequencies, $\hat{p}_g = M_g/N_g$, are observed, where M_g is the number of observed data points in stratum g and N_g is the number of times stratum g was sampled. (Is N_g known?)

- The estimated asymptotic variance in that case is

$$\left[\sum_{i=1}^{M}\mathbf{x}_i'\mathbf{x}_i/\hat{p}_{g_i}\right]^{-1} \cdot\left\{\sum_{g=1}^{G}\hat{p}_g^{-2}\left[\sum_{i=1}^{M_g}(\mathbf{x}_{gi}'\hat{u}_{gi}-\overline{\mathbf{x}_g'\hat{u}_g})(\mathbf{x}_{gi}'\hat{u}_{gi}-\overline{\mathbf{x}_g'\hat{u}_g})'\right]\right\} \cdot\left[\sum_{i=1}^{M}\mathbf{x}_i'\mathbf{x}_i/\hat{p}_{g_i}\right]^{-1}, \tag{20}$$

where M_g is the number of observed data points in stratum g.

Essentially the same as SS case in (18).

- If we use the known sampling weights, we drop $\overline{\mathbf{x}_g' \hat{u}_g}$ from (20). If $E(u|\mathbf{x}) = 0$ and the sampling is exogenous, we also drop this term because $E(\mathbf{x}'u|\mathbf{w} \in W_g) = \mathbf{0}$ for all g, and this is whether or not we estimate the p_g.
- If we are estimating a linear projection and use the estimated weights, $\overline{\mathbf{x}_g' \hat{u}_g}$ remains in the formula even when stratification is based on $\mathbf{x}$.
- Similar results carry over to nonlinear models.

3. Clustering and Stratification

- Survey data often characterized by clustering and VP sampling.

Suppose that g represents the primary sampling unit (say, city) and individuals or families (indexed by m) are sampled within each PSU with probability p_{gm}. If $\hat{\beta}$ is the pooled OLS estimator across PSUs and individuals, its variance is estimated as

$$\left(\sum_{g=1}^{G} \sum_{m=1}^{M_g} \mathbf{x}'_{gm} \mathbf{x}_{gm} / p_{gm} \right)^{-1}$$
$$\cdot \left[\sum_{g=1}^{G} \sum_{m=1}^{M_g} \sum_{r=1}^{M_g} \hat{u}_{gm} \hat{u}_{gr} \mathbf{x}'_{gm} \mathbf{x}_{gr} / (p_{gm} p_{gr}) \right] \tag{21}$$
$$\cdot \left(\sum_{g=1}^{G} \sum_{m=1}^{M_g} \mathbf{x}'_{gm} \mathbf{x}_{gm} / p_{gm} \right)^{-1} .$$

If the probabilities are estimated using retention frequencies, (21) is conservative, as before.

• Multi-stage sampling schemes introduce even more complications. Let there be S strata (e.g., states in the U.S.), exhaustive and mutually exclusive. Within stratum s, there are C_s clusters (e.g., neighborhoods).

• Large-sample approximations: the number of clusters sampled, N_s, gets large. This allows for arbitrary correlation (say, across households) within cluster.

• Within stratum s and cluster c, let there be M_{sc} total units (household or individuals). Therefore, the total number of units in the population is

$$M = \sum_{s=1}^{S} \sum_{c=1}^{C_s} M_{sc}. \tag{22}$$

• Let z be a variable whose mean we want to estimate. List all population values as $\{z^o_{scm} : m = 1, \ldots, M_{sc}, c = 1, \ldots, C_s, s = 1, \ldots, S\}$, so the population mean is

$$\mu = M^{-1} \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{m=1}^{M_{sc}} z^o_{scm}. \tag{23}$$

Define the total in the population as

$$\tau = \sum_{s=1}^{S} \sum_{c=1}^{C_s} \sum_{m=1}^{M_{sc}} z^o_{scm} = M\mu. \tag{24}$$

Totals within each cluster and then stratum are, respectively,

$$\tau_{sc} = \sum_{m=1}^{M_{sc}} z_{scm}^{o} \tag{25}$$

$$\tau_{s} = \sum_{c=1}^{C_s} \tau_{sc} \tag{26}$$

- Sampling scheme:

(i) For each stratum s, randomly draw N_s clusters, with replacement.

(Fine for "large" N_s.)

(ii) For each cluster c drawn in step (i), randomly sample K_{sc} households with replacement.

• For each pair (s, c), define

$$\hat{\mu}_{sc} = K_{sc}^{-1} \sum_{m=1}^{K_{sc}} z_{scm}. \tag{27}$$

Because this is a random sample within (s, c),

$$E(\hat{\mu}_{sc}) = \mu_{sc} = M_{sc}^{-1} \sum_{m=1}^{M_{sc}} z_{scm}^{o}. \tag{28}$$

• To continue up to the cluster level we need the total, $\tau_{sc} = M_{sc}\mu_{sc}$.

So, $\hat{\tau}_{sc} = M_{sc}\hat{\mu}_{sc}$ is an unbiased estimator of τ_{sc} for all $\{(s, c) : c = 1, \ldots, C_s, s = 1, \ldots, S\}$ (even if we eventually do not use some clusters).

• Next, consider randomly drawing N_s clusters from stratum s. Can show that an unbiased estimator of the total τ_s for stratum s is

$$C_s \cdot N_s^{-1} \sum_{c=1}^{N_s} \hat{\tau}_{sc}. \tag{29}$$

Finally, the total in the population is estimated as

$$\sum_{s=1}^{S} \left(C_s \cdot N_s^{-1} \sum_{c=1}^{N_s} \hat{\tau}_{sc} \right) \equiv \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} z_{scm} \tag{30}$$

where the weight for stratum-cluster pair (s, c) is

$$\omega_{sc} \equiv \frac{C_s}{N_s} \cdot \frac{M_{sc}}{K_{sc}}. \tag{31}$$

- Note how $\omega_{sc} = (C_s/N_s)(M_{sc}/K_{sc})$ accounts for under- or over-sampled clusters within strata and under- or over-sampled units within clusters.
- (30) appears in the literature on complex survey sampling, sometimes without M_{sc}/K_{sc} when each cluster is sampled as a complete unit, and so $M_{sc}/K_{sc} = 1$.
- To estimate the mean μ, just divide by M, the total number of units sampled.

$$\hat{\mu} = M^{-1}\left(\sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}}\omega_{sc}z_{scm}\right). \tag{32}$$

• To study regression (and many other estimation methods), specify the problem as

$$\min_{\beta} \sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} (y_{scm} - \mathbf{x}_{scm}\boldsymbol{\beta})^2. \tag{33}$$

The asymptotic variance combines clustering with weighting to account for the multi-stage sampling. Following Bhattacharya (2005), an appropriate asymptotic variance estimate has a sandwich form,

$$\left(\sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} \mathbf{x}'_{scm} \mathbf{x}_{scm} \right)^{-1} \hat{\mathbf{B}} \left(\sum_{s=1}^{S} \sum_{c=1}^{N_s} \sum_{m=1}^{K_{sc}} \omega_{sc} \mathbf{x}'_{scm} \mathbf{x}_{scm} \right)^{-1} \tag{34}$$

where $\hat{\mathbf{B}}$ is somewhat complicated:

$$\hat{\mathbf{B}} = \sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} \omega_{sc}^2 \hat{u}_{scm}^2 \mathbf{x}_{scm}' \mathbf{x}_{scm}$$

$$+ \sum_{s=1}^{S}\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}}\sum_{r\neq m}^{K_{sc}} \omega_{sc}^2 \hat{u}_{scm} \hat{u}_{scr} \mathbf{x}_{scm}' \mathbf{x}_{scr} \qquad (35)$$

$$- \sum_{s=1}^{S} N_s^{-1} \left(\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} \omega_{sc} \mathbf{x}_{scm}' \hat{u}_{scm} \right) \left(\sum_{c=1}^{N_s}\sum_{m=1}^{K_{sc}} \omega_{sc} \mathbf{x}_{scm}' \hat{u}_{scm} \right)'$$

- The first part of $\hat{\mathbf{B}}$ is obtained using the White "heteroskedasticity"-robust form. The second piece accounts for the clustering. The third piece reduces the variance by accounting for the nonzero means of the "score" within strata.