

Lecture Note 5: Stratified Sampling continued

Recall notation: $z = (y, x)$, with population density $f(z) = f(y, x) = f(y|x)f(x)$.

Strata: $s = 1, \dots, S$, which partition \mathcal{Z} into $\mathcal{J}_1, \dots, \mathcal{J}_S$.

Let $Q_s = Pr(z \in \mathcal{J}_s)$, and let $Q = (Q_1, \dots, Q_S)$.

Q may or may not be known.

Recall the three main stratification schemes:

1. Standard Stratified Sampling: for each stratum s , we fixed number of observations n_s and draw a random sample of size n_s from \mathcal{J}_s .

The density of z in stratum s is given by the truncated density

$$\frac{f(z)1(z \in \mathcal{J}_s)}{Q_s}.$$

2. Variable Probability Sampling: π_s is the probability of retaining an observation which lies in stratum s .
3. Multinomial Sampling: for $i = 1, \dots, n$, draw a stratum s_i randomly according to the multinomial distribution with probabilities P_1, \dots, P_S , and then draw z_i from stratum s_i .

Parametric Model

Suppose the joint density of z has a parametric form

$$f(z|\theta, \alpha) = f(y|x, \theta)f(x|\alpha),$$

where θ and α are parameters. Importantly, we assume that θ and α are variation-independent: the set of possible values of θ and α is a product $\Theta \times \mathcal{A}$.

Also, suppose we are only interested in θ , so that α is a “nuisance” parameter.

Usually, the marginal density of x is “completely unknown,” so that α essentially indexes all possible density functions. To save some notation we’ll write $h(x) = f(x|\alpha)$, and think of h as the nuisance parameter. Then we can write

$$f(z|\theta, h) = f(y|x, \theta)h(x).$$

Keep in mind that h is unknown, whereas $f(y|x, \theta)$ has known form (a parametric family) but θ is unknown.

Now we can write the stratum probabilities as

$$\begin{aligned} Q_s(\theta, h) &= Pr_{\theta, h}(z \in \mathcal{J}_s) \\ &= \int_{\mathcal{J}_s} f(z|\theta, h) dz \\ &= \iint_{\mathcal{J}_s} f(y|x, \theta) h(x) dx dy. \end{aligned}$$

We use $Q_s(\theta, h)$ to indicate that the probabilities depend on both θ and h .

Ancillarity

Suppose $f(z|\gamma)$ is a parametric family of probability densities.

Ancillary statistic: a statistic $t(z)$ whose distribution does not depend on γ .

Fisher: should conduct inference conditional on ancillary statistics.

Let $f(z|t, \gamma)$ denote the conditional density of z given $t(z) = t$.

Consider multinomial sampling: the stratum sizes n_1, \dots, n_S have a multinomial distribution with parameters n and $P = (P_1, \dots, P_S)$ which are determined by the researcher.

This distribution does not depend on θ, h , so the stratum sizes are ancillary statistics.

Conditional on n_1, \dots, n_S , we have random samples of size n_s from each stratum s , which is equivalent to standard stratified sampling. So by the ancillarity principle, we can treat a multinomial stratified sample as if it were a standard stratified sample.

Similar argument works for variable probability sampling.

So we will focus on standard stratified sampling.

The log likelihood under standard stratified sampling is

$$\sum_{s=1}^S \sum_{i=1}^{n_s} \log \left[\frac{f(y_{si}|x_{si}, \theta) h(x_{si})}{Q_s(\theta, h)} \right].$$

Where we go from here depends on whether the stratification involves y or only x :

Exogenous Stratification

Suppose stratification only involves x .

Let $\tilde{\mathcal{J}}_1, \dots, \tilde{\mathcal{J}}_S$ partition \mathcal{X} , and write

$$Q_s(h) = Pr_{\theta, h}(x \in \tilde{\mathcal{J}}_s) = \int_{\tilde{\mathcal{J}}_s} h(x) dx.$$

Notice that Q_s only involves h .

The log likelihood can be written as

$$\sum_{s=1}^S \sum_{i=1}^{n_s} \log \left[\frac{f(y_{si}|x_{si}, \theta) h(x_{si})}{Q_s(h)} \right] = \sum_{s=1}^S \sum_{i=1}^{n_s} \left\{ \log f(y_{si}|x_{si}, \theta) + \log \left[\frac{h(x_{si})}{Q_s(h)} \right] \right\}.$$

So the log likelihood separates into two pieces, and we only need to use the part involving θ to conduct estimation and inference about θ .

Endogenous Stratification

Suppose stratification depends on y (but for simplicity, not on x).

Let $\bar{\mathcal{J}}_1, \dots, \bar{\mathcal{J}}_S$ partition \mathcal{Y} .

$$Q_s(\theta, h) = Pr_{\theta, h}(y \in \bar{\mathcal{J}}_s) = \int_{\bar{\mathcal{J}}_s} \int_{\mathcal{X}} f(y|x, \theta) h(x) dx dy.$$

Suppose that the stratum probabilities $Q = (Q_1, \dots, Q_S)$ are known to be (q_1, \dots, q_S) . We could replace the $Q_s(\theta, h)$ by the known values in the log likelihood:

$$\sum_{s=1}^S \sum_{i=1}^{n_s} \log \left[\frac{f(y_{si}|x_{si}, \theta) h(x_{si})}{q_s} \right].$$

However, there is additional information about the parameters θ, h contained in the moment equations

$$q_s = Q_s(\theta, h) = \int_{\bar{\mathcal{J}}_s} \int_{\mathcal{X}} f(y|x, \theta) h(x) dx dy.$$

Identification

In some cases, stratification can induce lack of identification. A simple example: suppose that y is a choice variable, taking on values $1, \dots, K$, and there are no covariates. Then we could model

$$Pr(y = k) = \delta_k, \quad k = 1, \dots, K.$$

Assume $\delta_k > 0$ for all k , and that $\sum_{k=1}^K \delta_k = 1$.

Suppose we do stratified sampling, where each possible choice k corresponds to a stratum. This is called choice-based sampling.

Within stratum k , $Pr(y = k) = 1$. So there is no information about the δ_k in the stratified sample.

Of course, if we knew the stratum probabilities $Q_k = Pr(y = k) = \delta_k$, then we would know all the parameters of the model.

Here is a slightly more elaborate example of identification failure.

Suppose we have K choices as before, and choice-based sampling, but now we are interested in relating choices to some explanatory variables.

Let $x = (x_1, \dots, x_K)$ be explanatory variables, with $x_K = 0$ as a normalization, and suppose

$$Pr(y = k|x) = \frac{\exp(\delta_k + x'_k \beta)}{\sum_{l=1}^K \exp(\delta_l + x'_l \beta)}.$$

x has unknown density $h(x)$.

Write $\theta = (\delta_1, \dots, \delta_K, \beta)$.

Suppose θ and h are the “true” parameter values, and consider alternative values θ^* and h^* constructed in the following way.

$$c_k e^{\delta_k^*} = e^{\delta_k},$$

for some c_1, \dots, c_K that are not all equal, and

$$\beta^* = \beta.$$

Note that

$$\begin{aligned} Pr(y = k|x, \theta) &= \frac{e^{\delta_k} \exp(x'_k \beta)}{\sum_{l=1}^K e^{\delta_l} \exp(x'_l \beta)} \\ &= \frac{c_k e^{\delta_k^*} \exp(x'_k \beta)}{\sum_{l=1}^K c_l e^{\delta_l^*} \exp(x'_l \beta)} \\ &= \frac{c_k Pr(y = k|x, \theta^*)}{\sum_{l=1}^K Pr(y = l|x, \theta^*)}. \end{aligned}$$

Also note that the density of observations in stratum k is

$$\frac{Pr(y = k|x, \theta)h(x)}{Q_k(\theta, h)},$$

where Q_k is a constant (with respect to x and y) that normalizes this expression to be a probability density.

By our previous expression for $Pr(y = k|x, \theta)$,

$$\frac{Pr(y = k|x, \theta)h(x)}{Q_k(\theta, h)} = \frac{c_k Pr(y = k|x, \theta^*)}{\sum_{l=1}^K c_l Pr(y = l|x, \theta^*)} \cdot \frac{h(x)}{Q_k(\theta, h)}.$$

Let

$$h^*(x) = \frac{c_0 h(x)}{\sum_{l=1}^K c_l Pr(y = l|x, \theta^*)},$$

where c_0 is a normalizing constant. Then we can write

$$\frac{Pr(y = k|x, \theta)h(x)}{Q_k(\theta, h)} = \frac{Pr(y = k|x, \theta^*)h^*(x)}{Q_k(\theta, h)c_0/c_k} = \frac{Pr(y = k|x, \theta^*)h^*(x)}{Q_k(\theta^*, h^*)}.$$

So (θ, h) and (θ^*, h^*) yield the same likelihood functions.

ML Based Estimation

Return to the general stratified sampling setup with endogenous stratification.

A “naive” log conditional likelihood function (ignoring the stratification) is

$$L(\theta) = \frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} \log f(y_{si}|x_{si}, \theta).$$

The naive MLE satisfies the score equation

$$\frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} \nabla_{\theta} \log f(y_{si}|x_{si}, \theta) = 0.$$

To see why this will not lead to a good estimate, recall that in the overall population,

$$E[\nabla_{\theta} \log f(y|x, \theta)] = 0,$$

which we can write in terms of strata-specific expectations as

$$\sum_{s=1}^S Q_s E[\nabla_{\theta} \log f(y|x, \theta)|\bar{\mathcal{J}}_s] = 0.$$

But if we take the expectation of the naive MLE score equation, we get

$$E\left[\frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} \nabla_{\theta} \log f(y_{si}|x_{si}, \theta)\right] = \sum_{s=1}^S \frac{n_s}{n} E[\nabla_{\theta} \log f(y|x, \theta)|\bar{\mathcal{J}}_s],$$

which does not equal 0 in general since n_s/n does not equal Q_s .

One solution is to weight the observations. Units in strata s would get weight $Q_s/(n_s/n)$, so that the weighted log likelihood would be

$$L_w(\theta) = \frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} \frac{Q_s}{n_s/n} \log f(y_{si}|x_{si}, \theta).$$

The FOC for the MLE are

$$\frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} \frac{Q_s}{n_s/n} \nabla_{\theta} \log f(y_{si}|x_{si}, \theta) = 0.$$

At the true value of θ , the score equation has expectation

$$E \left[\frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} \frac{Q_s}{n_s/n} \nabla_{\theta} \log f(y_{si}|x_{si}, \theta) \right] = 0.$$

Thus it is a valid moment equation, and can be viewed as a special case of GMM in order to get standard errors.

The weighted MLE is commonly used, as it is relatively straightforward to modify standard routines for MLE to handle the weighting. However, it is not efficient in general.

Efficient Estimation

Efficient estimators can be constructed by starting with the “correct” likelihood function

$$\sum_{s=1}^S \sum_{i=1}^{n_s} \log \frac{f(y_{si}|x_{si}, \theta) h(x_{si})}{Q_s(\theta, h)}.$$

To start, let’s go back to our earlier notation for the marginal density of x :

$$h(x) = f(x|\alpha),$$

and suppose that α is a finite-dimensional parameter. In other words, we have parametrically specified the marginal density of x . In this case, the log likelihood would be

$$\sum_{s=1}^S \sum_{i=1}^{n_s} \log \frac{f(y_{si}|x_{si}, \theta) f(x_{si}|\alpha)}{Q_s(\theta, \alpha)}.$$

We could then define the maximum likelihood estimator of θ, α as the solution to

$$\max_{\theta, \alpha} \sum_{s=1}^S \sum_{i=1}^{n_s} [\log f(y_{si}|x_{si}, \theta) + \log f(x_{si}|\alpha) - \log Q_s(\theta, \alpha)].$$

It can be useful to concentrate the log likelihood: for any θ , let $\hat{\alpha}(\theta)$ maximize the log likelihood with respect to α . Then choose $\hat{\theta}$ to solve

$$\max_{\theta} \sum_{s=1}^S \sum_{i=1}^{n_s} [\log f(y_{si}|x_{si}, \theta) + \log f(x_{si}|\hat{\alpha}(\theta)) - \log Q_s(\theta, \hat{\alpha}(\theta))].$$

By standard results on MLE, this estimator should be consistent, asymptotically normal, and efficient (in the sense of achieving the Cramer-Rao variance bound asymptotically).

What if we don't have a parametric specification for the marginal density of x ?

Aside on Nonparametric MLE

Suppose we have x_1, \dots, x_n IID from an unknown distribution H .

Consider discrete distributions with support equal to the observed values of x . Any such distribution can be described by n probabilities h_1, \dots, h_n , where $h_1 = Pr(x = x_1)$, etc. So we can think of this as a parametric model with n parameters.

The empirical likelihood function is defined as

$$\mathcal{L}(h_1, \dots, h_n) = \prod_{i=1}^n h_i,$$

and the log empirical likelihood is

$$\sum_{i=1}^n \log h_i.$$

The NPMLE chooses $h = (h_1, \dots, h_n)$ to solve

$$\max_{h_1, \dots, h_n} \sum_{i=1}^n \log h_i, \quad \text{s.t. } h_i \geq 0, \sum_{i=1}^n h_i = 1.$$

The solution turns out to be $\hat{h}_i = \frac{1}{n}$ for all i .

In other words, the estimated distribution \hat{H} is the empirical distribution which places mass $\frac{1}{n}$ on each of the observed values of x .

Notice also that expectations with respect to the empirical distribution are just sample averages.

Now, returning to our stratified sampling problem with unknown $h(x)$:

Cosslett suggests to estimate θ and h by maximizing the joint likelihood, using the empirical likelihood representation of h . We write the likelihood as

$$\sum_{s=1}^S \sum_{i=1}^{n_s} \log \frac{f(y_{si}|x_{si}, \theta) h_{si}}{Q_s(\theta, h)},$$

where the h_{si} are parameters giving the probability that $x = x_{si}$ and

$$Q_s(\theta, h) = \int_{\bar{\mathcal{J}}_s} \sum_{s=1}^S \sum_{i=1}^{n_s} h_{si} \cdot f(y|x_{si}, \theta) dy.$$

This is a difficult maximization problem, and the solution is somewhat complicated (see the Cosslett survey for the formulas). The estimates of h_{si} are no longer the simple form of the basic NPMLE, because of the $Q_s(\theta, h)$ terms in the log likelihood.

But the benefit is that the estimator for θ is fully efficient, in the sense of achieving a semiparametric efficiency bound.