

## Lecture Note 4: GMM for 2-Step Estimators; Stratified Sampling

### GMM for 2-Step Estimators

Consider the following model for regression under sample selection, due to Heckman (1976):

$$y_i^* = w_i' \beta_0 + u_i$$

$$d_i^* = x_i' \gamma_0 + \epsilon_i,$$

$$d_i = 1(d_i^* > 0),$$

Here,  $w_i$  will typically be a subvector of  $x_i$ :  $x_i = (w_i, v_i)$ .

The joint disturbance is normally distributed:

$$\begin{pmatrix} u_i \\ \epsilon_i \end{pmatrix} | w_i, x_i \sim N(0, V),$$

$$V = \begin{bmatrix} \sigma_u^2 & \sigma_{u\epsilon} \\ \sigma_{u\epsilon} & 1 \end{bmatrix}.$$

We observe  $(w_i, x_i, d_i)$ , and  $y_i = y_i^* \cdot d_i$ . In other words, we don't observe the outcome of  $d_i = 0$ .

Note that

$$\begin{aligned} P(d_i = 1 | x_i) &= P(d_i^* > 0 | x_i) \\ &= P(x_i' \gamma_0 + \epsilon_i > 0 | x_i) \\ &= P(\epsilon_i > -x_i' \gamma_0 | x_i) \\ &= P(-\epsilon_i < x_i' \gamma_0 | x_i) \\ &= \Phi(x_i' \gamma_0). \end{aligned}$$

Simply regressing  $y_i$  on  $w_i$  for the  $d_i = 1$  subsample will not give consistent estimates of  $\beta_0$ , because

$$E[y_i | w_i, x_i, d_i = 1] = w_i' \beta_0 + E[u_i | \epsilon_i > -x_i' \gamma_0].$$

The last term is not equal to 0. It can be shown that the last term is

$$E[u_i | \epsilon_i > -x_i' \gamma_0] = \alpha_0 \lambda(x_i' \gamma_0),$$

where

$$\lambda(t) = \frac{\phi(t)}{\Phi(t)}.$$

Heckman suggested a two-step estimator for  $\beta_0$ :

1. First estimate  $\gamma_0$  by ML in the probit model of  $d_i$  on  $x_i$ . Let  $\hat{\gamma}$  denote the estimate.
2. For the  $d_i = 1$  subsample, regress  $y_i$  on  $w_i$  and  $\lambda(x_i'\hat{\gamma})$ .

The regression in the second step involves a generated regressor. This affects the sampling properties of the estimator  $\hat{\beta}$ , so the standard errors need to reflect the first-stage estimation of  $\hat{\gamma}$ .

One simple way to get valid standard errors is to fit this into the GMM framework. Then we can use the GMM formulas. Recall that the conditional log-likelihood for the probit model is

$$\log f(d_1, \dots, d_n | x_1, \dots, x_n) = \sum_{i=1}^n d_i \log \Phi(x_i' \gamma) + (1 - d_i) \log [1 - \Phi(x_i' \gamma)].$$

The score equation (first order condition) is

$$0 = \frac{\partial}{\partial \gamma} \log f(d_1, \dots, d_n | x_1, \dots, x_n).$$

This can be calculated as

$$0 = \sum_{i=1}^n x_i \frac{\phi(x_i' \gamma)}{\Phi(x_i' \gamma)(1 - \Phi(x_i' \gamma))} (d_i - \Phi(x_i' \gamma)).$$

Also, the normal equations for the second-stage least squares regression can be written as

$$0 = \sum_{i=1}^n d_i \begin{pmatrix} w_i \\ \lambda(x_i' \gamma) \end{pmatrix} (y_i - w_i' \beta - \alpha \lambda(x_i' \gamma)).$$

This suggest two moment functions: letting  $z_i = (w_i, x_i, d_i, y_i)$  and  $\theta = (\gamma, \alpha, \beta)$ :

$$m_1(z_i, \theta) = x_i \frac{\phi(x_i' \gamma)}{\Phi(x_i' \gamma)(1 - \Phi(x_i' \gamma))} (d_i - \Phi(x_i' \gamma)),$$

$$m_2(z_i, \theta) = d_i \begin{pmatrix} w_i \\ \lambda(x_i' \gamma) \end{pmatrix} (y_i - w_i' \beta - \alpha \lambda(x_i' \gamma)).$$

Let

$$g(z_i, \theta) = \begin{pmatrix} m_1(z_i, \theta) \\ m_2(z_i, \theta) \end{pmatrix}.$$

Then the 2-step estimator is a GMM estimator solving

$$0 = \frac{1}{n} \sum_{i=1}^n g(z_i, \theta).$$

## Random Sampling

Before talking about stratified sampling, it may be useful to first review what we mean by ordinary random sampling.

Suppose we denote a population of interest by the set  $\Omega$ . For example,  $\Omega$  could be the entire population of the U.S. in 2000, and an element  $\omega \in \Omega$  denotes a single individual. Here, the population is a finite (but large) set. Let  $N$  be the size of the population.

We might be interested in some variable  $Y$  (say, earnings). Let  $Y(\omega)$  denote the earnings of individual  $\omega$ . The population average earnings is

$$PAE := \frac{1}{N} \sum_{\omega \in \Omega} Y(\omega).$$

Here, each  $Y(\omega)$  is a fixed number. If we could observe every individual's earnings, then we could calculate  $PAE$  directly.

In practice, we might take a sample from the population, calculate the average earnings in the sample, and use this as an estimate of the population average earnings.

Random sampling with replacement: suppose we randomly pick an individual from the population and record their earnings at  $Y_1$ . Then we randomly pick another individual (possibly picking the first individual again), and denote this  $Y_2$ , and so on, up to  $n$ . (Note the distinction between  $n$  and  $N$ .)

Each  $Y_i$  is a random draw from the distribution that puts probability  $\frac{1}{N}$  on each  $Y(\omega)$ . Moreover, due to replacement, the draws  $Y_i$  are independent and identically distributed. Then we could form

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i.$$

It is straightforward to show that  $\bar{Y}_n$  has expected value  $PAE$  and that the variance of the estimator is  $\sigma^2/n$ , where  $\sigma^2$  is the variance in the population:  $\sigma^2 = \frac{1}{N} \sum_{\omega} (Y(\omega) - PAE)^2$ .

Random sampling without replacement: most surveys are not actually done in the previous fashion. Instead, we usually take a random sample of size  $n$  without replacement. This can be thought of as placing equal probabilities on each of the  $\binom{N}{n}$  possible subsets of  $\Omega$  of size  $n$ .

Equivalently, we can think of drawing  $Y_1$  from the entire population with probabilities  $1/N$ ; then conditional on the first draw, we draw  $Y_2$  from the remaining population (of size  $N - 1$ ) with probabilities  $1/(N - 1)$ , and so on. Clearly, this is not an IID sample. The estimator  $\bar{Y}_n$  is still unbiased, but its variance can be shown to be

$$\frac{(N - n)(N - 1) \sigma^2}{N^2} \frac{1}{n},$$

which is slightly smaller.

## Large Population

In the previous variance formula, notice that if  $N$  is very large and large relative to  $n$ , then the difference will be slight. More generally, for very large populations, there is very little difference in the distribution of  $(Y_1, \dots, Y_n)$  under random sampling with replacement and random sampling without replacement. If the population  $\Omega$  is uncountably infinite, then there is no difference. So the usual assumption of an IID random sample is often a convenient fiction. In some cases, we are interested not in the population but some larger “superpopulation.” For example, we might test a medical therapy on a sample from the population of diseased individuals in 2006, but we have in mind using the results for making decisions about treatments on future individuals, who are not formally part of the 2006 disease population.

For the rest of this note, we’ll assume a large population. But it may be useful to consider how the analysis would change in the finite-population case.

## Stratified Sampling

Many surveys are not actually random samples. Instead, they use a stratified sampling scheme, in which the population is divided into “strata” based on some observable characteristics. Within each stratum, a random sample is drawn, but the survey may give more weight to some strata than others.

Why stratify? Two common reasons: operational convenience; deliberate oversampling of some subpopulations of particular interest.

Examples of stratified sampling:

1. Manski and Lerman (1977) examined a data set on transportation usage. They were interested in understanding the factors that caused people to adopt automobiles vs. buses as their means of transportation. The survey defined strata based on the individual’s choice of transportation (“choice-based sampling”), in order to get a reasonable number of observations for each form of transportation.
2. In some surveys such as the Panel Study of Income Dynamics, there is a random sample drawn from the US population, which is augmented by a second sample drawn from a group of relatively disadvantaged individuals. Here we have two strata, which partially overlap.

To keep notation simple, we’ll focus on the case where the strata are nonoverlapping. See Cosslett (1993) for the extension to overlapping strata.

## Notation

Let  $\Omega$  denote the population, which we assume is large and consider as a probability space.

$z(\omega)$ : observable variables for individual  $\omega \in \Omega$ , taking values in  $\mathcal{Z}$ .

(We could have  $z = (y, x)$ , with  $y$  taking values in  $\mathcal{Y}$  and  $x$  taking values in  $\mathcal{X}$ , so  $\mathcal{Z} \subset \mathcal{Y} \times \mathcal{X}$ .)

Strata:  $s = 1, \dots, S$ . We assume a finite number of strata.

For each  $s$ , let  $\mathcal{J}_s \subset \mathcal{Z}$  be the subpopulation. We assume  $\mathcal{J}_1, \dots, \mathcal{J}_S$  form a partition of  $\mathcal{Z}$ .

Let  $f(z)$  denote the density (with respect of some measure) of the random variable  $z = z(\omega)$ .

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

In some cases,  $Q$  is known; in other cases, it is unknown. Sometimes there is additional information that can be used to estimate  $Q$ .

### Stratification Schemes

There are different ways to implement stratified sampling:

#### 1. Standard Stratified Sampling:

We fix the number of observations from each stratum:  $n_1, n_2, \dots, n_S$ .

For each stratum  $s$ , we draw  $n_s$  observations using random sampling from  $\mathcal{J}_s$ .

Note:  $n_s$  is not a random variable.

We end with  $S$  random samples, each from a different subpopulation.

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

We start with a pure random sample of size  $\bar{n}$ .

For each  $i = 1, \dots, \bar{n}$ , we determine what stratum that individual is in. For an individual in stratum  $s$ , they are kept in the sample with probability  $\pi_s \geq 0$ .

The probabilities  $\pi = (\pi_1, \dots, \pi_S)$  are determined by the researcher and assumed known.

Let  $n$  be the number of retained samples, and relabel the observations  $i = 1, \dots, n$ . So  $n$  is a random variable.

This type of sampling is sometimes used when it is easy to determine which stratum an individual belongs to, but costly to gather more information about the individual.

#### 3. Multinomial Sampling

This works in two stages. First, we select a stratum  $s_i$  at random according to probabilities  $P_s$ . Then, we draw  $z_i$  from that stratum randomly.

In this scheme, the observations are IID.

## Sample Averages and Weighting

For now, let's focus on standard stratified sampling. Suppose we are interested in estimating the population mean of  $y$ . Our target  $E[y]$  can be written as

$$E[y] = \sum_{s=1}^S Q_s E[y|\mathcal{J}_s].$$

The notation  $E[y|\mathcal{J}_s]$  is shorthand for  $E[y|z \in \mathcal{J}_s]$ , and this formula follows by the law of iterated expectations.

Consider the sample average

$$\bar{y}_n = \frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} y_{si},$$

where  $n = \sum_{s=1}^S n_s$ . The notation  $y_{si}$  refers to the  $i$ th random draw in the  $s$ th stratum.

Then

$$\begin{aligned} E[\bar{y}_n] &= \frac{1}{n} \sum_{s=1}^S n_s E[y|\mathcal{J}_s] \\ &= \sum_{s=1}^S \frac{n_s}{n} E[y|\mathcal{J}_s]. \end{aligned}$$

So unless  $n_s/n$  equals  $Q_s$ , this estimator is biased.

Consider a weighted average, where we weight each observation by  $Q_s \cdot \frac{n}{n_s}$ :

$$\bar{y}_w = \frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} y_{si} \left( Q_s \cdot \frac{n}{n_s} \right).$$

It's easy to verify that this will be unbiased.

Intuition: individuals in stratum  $s$  represent  $n_s/n$  of the sample, but represent  $Q_s$  of the population. For example, if  $n_s/n = 1/10$ , but  $Q_s = 1/2$ , we need to use each observation in stratum  $s$  to represent 5 individuals "in the population."

## Weighting and Linear Regression

Suppose we're interested in estimating the relationship between  $y$  and  $x$ , and stratification only depends on  $x$ .

Formally, let  $z = (y, x)$ , and suppose that the strata  $\mathcal{J}_1, \dots, \mathcal{J}_S$  are such that we can define a

corresponding partition  $\tilde{\mathcal{J}}_1, \dots, \tilde{\mathcal{J}}_S$  of  $\mathcal{X}$  such that

$$z = (y, x) \in \mathcal{J}_s \quad \leftrightarrow \quad x \in \tilde{\mathcal{J}}_s.$$

So

$$Q_s = Pr(z \in \mathcal{J}_s) = Pr(x \in \tilde{\mathcal{J}}_s).$$

Note that the density of  $z$  in subpopulation  $s$  can be written

$$\frac{f(z)1(z \in \mathcal{J}_s)}{Q_s} = \frac{f(y|x)f(x)1(x \in \tilde{\mathcal{J}}_s)}{Q_s}.$$

Suppose we assume a classical linear regression model:

$$E[y|x] = x'\beta, \quad \forall x \in \mathcal{X}.$$

Consider the LS estimator

$$\hat{\beta} = (X'X)^{-1}X'Y = \left( \sum_{s=1}^S \sum_{i=1}^{n_s} x_{si}x'_{si} \right)^{-1} \sum_{s=1}^S \sum_{i=1}^{n_s} x_{si}y_{si}.$$

Then

$$E[\hat{\beta}|X] = \left( \sum_{s=1}^S \sum_{i=1}^{n_s} x_{si}x'_{si} \right)^{-1} \sum_{s=1}^S \sum_{i=1}^{n_s} x_{si}E[y_{si}|x_{si}] = \beta.$$

Therefore

$$E[\hat{\beta}] = \beta.$$

When stratification depends only on exogenous variables, we call this “exogenous sampling.”

But be careful: suppose we do not assume that the conditional mean of  $y$  given  $x$  is linear, but instead want to estimate the best linear predictor coefficient:

$$\gamma = (E[xx'])^{-1}E[xy].$$

This can be written as

$$\gamma = \left( \sum_{s=1}^S Q_s E[xx'|\tilde{\mathcal{J}}_s] \right)^{-1} \left( \sum_{s=1}^S Q_s E[xy|\tilde{\mathcal{J}}_s] \right).$$

As we showed in LN1, under random sampling, the LS estimator converges in probability to  $\gamma$ .

Assume that for each  $s$ ,  $n_s \rightarrow \infty$  and  $n_s/n \rightarrow \xi_s$ . Write

$$\hat{\beta} = \left( \frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} x_{si} x'_{si} \right)^{-1} \frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} x_{si} y_{si}.$$

Consider the term

$$\frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} x_{si} x'_{si} = \sum_{s=1}^S \frac{n_s}{n} \frac{1}{n_s} \sum_{i=1}^{n_s} x_{si} x'_{si}.$$

The inner term

$$\frac{n_s}{n} \frac{1}{n_s} \sum_{i=1}^{n_s} x_{si} x'_{si} \xrightarrow{p} \xi_s E[xx' | \tilde{\mathcal{J}}_s].$$

So

$$\sum_{s=1}^S \frac{n_s}{n} \frac{1}{n_s} \sum_{i=1}^{n_s} x_{si} x'_{si} \xrightarrow{p} \sum_{s=1}^S \xi_s E[xx' | \tilde{\mathcal{J}}_s].$$

Likewise,

$$\frac{1}{n} \sum_{s=1}^S \sum_{i=1}^{n_s} x_{si} y_{si} \xrightarrow{p} \sum_{s=1}^S \xi_s E[xy | \tilde{\mathcal{J}}_s].$$

So

$$\hat{\beta} \xrightarrow{p} \left( \sum_{s=1}^S \xi_s E[xx' | \tilde{\mathcal{J}}_s] \right)^{-1} \sum_{s=1}^S \xi_s E[xy | \tilde{\mathcal{J}}_s] \neq \gamma.$$

It would make sense to instead use a weighted LS estimator. Let the weight for observation  $i$  in stratum  $s$  be

$$w_{si} = Q_s \cdot \frac{n}{n_s},$$

and let  $W$  be the diagonal matrix of weights  $w_{si}$ . Then

$$\hat{\beta}_w = (X'WX)^{-1} X'W y \xrightarrow{p} \gamma.$$