ECON 121, Applied Econometrics and Data Analysis **Prof. Tom Vogl**
Fall 2023 **University of California San Diego**

LECTURE NOTE 3: UNEQUAL PROBABILITY SAMPLING

1 Introduction

Most undergraduate statistics courses assume the data come from a *simple random sample*, in which all individuals in the population have equal probability of being selected for the sample. In practice, many datasets are not simple random samples. This lecture note is about how to analyze such samples.

2 Sampling Probabilities and Survey Weights

Many surveys choose participants randomly, but not with equal probability. One example is *stratified sampl*ing, in which survey designers first divide the population into groups (*strata*) and then draw a simple random sample within each group. The designers may oversample some groups, making the sample not directly representative of the population. Sometimes the oversampling is intended to generate a large subsample from a small subpopulation, enabling researchers to estimate more precise statistics for that group. Other times, the oversampling is intended to reduce survey costs, for instance by interviewing more intensively in areas with less violence or lower transportation costs. Another example is *survey nonresponse*. Even in a simple random sample, if individuals with different characteristics have different propensities to agree to respond to the survey, then the final sample overrepresents individuals with characteristics that increase response.

The idea can be a little confusing so deserves some extra explanation. Before a sample is chosen, every person in the population has a chance of being included in the sample. Call individual i's probability π_i. In a simple random sample, π_i is the same for everyone and equals the sample size divided by the number of people in the population. In a stratified sample that oversamples a minority group, a person from the minority group has a relatively high π_i. In a survey with differential nonresponse, a private or distrustful person has a relative low π_i. In the case of stratification, π_i is easy to determine because it is a direct consequence of the survey designer's sampling decisions, but in the case of differential nonresponse, it is more difficult because the drivers of non-response are usually not fully known.

Such survey data typically come with *survey weights* (or *sampling weights*). In the case of stratified sampling, the weights are called *design weights*; in the case of survey nonresponse, they are called *nonresponse weights* or sometimes *poststratification weights*. In both cases, survey designers set the weight based on their

assessment of π_i. In the next section, we will derive the most common survey weight in its original application, the estimation of a population total. The following sections will apply survey weights to mean estimation and regression estimation.

3 Horvitz-Thompson Estimator

The starting point for using survey weights in statistical analysis is the Horvitz-Thompson estimator for the estimating a total. Suppose we want to estimate total personal income in the United States, Y, the sum of individual income y_i across all N people in a finite population:

$$Y = \sum_{i=1}^{N} y_i$$

We have data from a survey that included each person i with probability π_i, resulting in a sample S with n people. We wish to estimate the population total using the data on the n people in the sample S in an unbiased way. We focus specifically on a linear estimator, such that it takes the form:

$$\hat{Y} = \sum_{i=1}^{n} w_i y_i$$

for some weights w_i. We take y_i and w_i as fixed, so that the only randomness in $\hat{Y}$ comes from sampling.

For $\hat{Y}$ to be unbiased, we need:

$$\begin{aligned}
E\left[\hat{Y}\right] &= Y \\
E\left[\sum_{i=1}^{n} w_i y_i\right] &= \sum_{i=1}^{N} y_i \\
E\left[\sum_{i=1}^{N} 1[i \in S] w_i y_i\right] &= \sum_{i=1}^{N} y_i \\
\sum_{i=1}^{N} E\left[1[i \in S] w_i y_i\right] &= \sum_{i=1}^{N} y_i \\
\sum_{i=1}^{N} \pi_i w_i y_i &= \sum_{i=1}^{N} y_i
\end{aligned}$$

The second line replaces $\hat{Y}$ and Y with their definitions. The third line changes from a summation over individuals $i = 1, \cdots, n$ in the *sample* to a summation individuals $i = 1, \cdots, N$ in the *population*, multiplying each $w_i y_i$ by $1[i \in S]$, a dummy variable for being in the sample. The fourth line replaces the expectation of a sum with the sum of expectations. The fifth line uses the fact that the expectation of a dummy variable is a probability, so that $E\left[1[i \in S] w_i y_i\right]$ can be written $\pi_i w_i y_i$.

For equality to hold generally in the last line, we need:

$$w_i = \frac{1}{\pi_i}$$

That is to say, to estimate the population total of y_i, we take a weighted sum of y_i in the sample, where the weights are equal to 1 over the probability of being included in the sample. The $\hat{Y}$ estimator using this definition of w_i is known as the *Horvitz-Thompson estimator*, after the statisticians who first divised it. We just proved the estimator if unbiased. It is consistent too, but we will not prove that property here.

This definition of w_i has intuitive properties. When π_i is small, w_i is large, so individuals from undersampled groups get larger weights. Furthermore, we can interpret w_i as the number of people from the population whom individual i represents. For example, in a 1% simple random sample, each sample member represents $w_i = \frac{1}{0.01} = 100$ people from the population.

4 Weighted Average

A typical survey dataset includes a survey weight like w_i as a variable, and we will now consider its use in a more familiar application, the estimation of the population mean. Often, the survey dataset includes w_i exactly as defined above. Sometimes, survey statisticians will rescale the weight, e.g. $\omega_i = \alpha w_i$, because they want the weights to sum to a number other than N, the population size. For example, they may set α to the sample size n divided by the population size N, so that the weights sum to the sample size. Or they may set α to 1 million, so they can remove the decimal point while avoiding major rounding errors. Or they may set α to $\frac{1}{\sum_{i=1}^{n} w_i}$, so that the weights sum to 1. These adjustments obviously affect the estimation of the population total. But we will see below that the weighted estimators for the mean and regression coefficient are invariant to α. Below, I will continue to use w_i in all formulas to avoid confusion, but the estimators work for any weight that is proportional to the inverse of π_i.

To estimate the population mean, we use the weighted average:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

This estimator simply divides $\hat{Y}$ by the total weight in the sample, $\sum_{i=1}^{n} w_i$. Since w_i represents the number of people represented by individual i, the sum of all w_i's equals the population size N, so that $\hat{\mu}$ equals $\hat{Y}$ divided by N. In a finite population, the population mean μ is equal to the total Y divided by N. So if $\hat{Y}$ is an unbiased and consistent estimator of Y, then $\hat{\mu}$ is an unbiased and consistent estimator of μ.

This weighted average is especially intuitive in the case of a stratified sample. Suppose surveyors sam-

ple urban and rural areas with different probabilities. The urban subsample is representative of the urban subpopulation, and the rural subsample is representative of the rural subpopulation, but the overall sample is not representative of the overall population. One approach to estimating the overall population mean would be to estimate a simple average within each stratum, $\bar{y}^s$ for stratum $s \in \{u, r\}$, and then take the population-weighted average of the $\bar{y}^s$'s:

$$\hat{\mu} = \frac{N^u}{N}\bar{y}^u + \frac{N^r}{N}\bar{y}^r$$

We weight each stratum-specific average by the population share of the stratum. This approach is equivalent to the individually-weighted average above. To see this point, first note that the sampling scheme leads to a sampling probability of $\pi^u = \frac{n^u}{N^u}$ for all urban residents and $\pi^r = \frac{n^r}{N^r}$ for all rural residents. As a result, the survey weights are constant within each sector: $w^u = \frac{N^u}{n^u}$ and $w^r = \frac{N^r}{n^r}$. Now consider the individually-weighted average:

$$\begin{aligned}
\hat{\mu} &= \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \\
&= \frac{1}{\sum_{i=1}^n w_i}\left(\sum_{i\in u} w_i y_i + \sum_{i\in r} w_i y_i\right) \\
&= \frac{1}{\sum_{i\in u} w^u + \sum_{i\in r} w^r}\left(w^u \sum_{i\in u} y_i + w^r \sum_{i\in r} y_i\right) \\
&= \frac{1}{n^u w^u + n^r w^r}\left(w^u n^u \bar{y}^u + w^r n^r \bar{y}^r\right) \\
&= \frac{1}{n^u \frac{N^u}{n^u} + n^r \frac{N^r}{n^r}}\left(\frac{N^u}{n^u} n^u \bar{y}^u + \frac{N^r}{n^r} n^r \bar{y}^r\right) \\
&= \frac{1}{N}\left(N^u \bar{y}^u + N^r \bar{y}^r\right)
\end{aligned}$$

We can see that taking an individually-weighted average is the same as taking a population-weighted average of stratum-specific simple averages. Both approaches lead to unbiased estimates of the population mean. We can also say that either weighted average is *representative* of the population.

5 Weighted Least Squares

Can we also use survey weights to obtain representative estimates of regression coefficients? We will see that the answer is yes and no, depending on how we think about representativeness in the context of regression.

To avoid confusion, we will expand on the notation from Sections 2-4 to accommodate two variables. We are interested in the relationship between an independent variable x_i and a dependent variable y_i in a population of size N. For the sake of example, we will suppose x_i is education and y_i is income. We have a sample of n individuals, with data on x_i, y_i, and a survey weight w_i.

The regression analogue to the weighted average is called *weighted least squares* (WLS) estimation. It solves the minimization problem:

$$\min_{\hat{b}_0^{WLS}, \hat{b}_1^{WLS}} \sum_{i=1}^{n} w_i \left(y_i - \hat{b}_0^{WLS} - \hat{b}_1^{WLS} x_i \right)^2$$

which is the same as the ordinary least squares (OLS) minimization problem, except that each squared deviation is multiplied by the survey weight. The solution for the slope coefficient is:

$$\hat{\beta}_1^{WLS} = \frac{\sum_{i=1}^{n} w_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} w_i (x_i - \bar{x})^2}$$

This estimator is similar to the OLS estimator. The OLS estimator equals the covariance of x_i and y_i divided by the variance of x_i. The WLS estimator equals the *weighted* covariance of x_i and y_i divided by the *weighted* variance of x_i.

This approach works well if we think of regression as tool to describe an empirical relationship in a finite population. In this sense, define the *population slope coefficient* as OLS estimator applied to the entire population:

$$\beta_1^{POP} = \frac{\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Note that the summations here go to N rather than n. Using a proof similar to Section 2's derivation of the Horvitz-Thompson estimator, one can show that $\hat{\beta}_1^{WLS}$ is an unbiased estimator of β_1^{POP}. That is to say, when we use WLS in an unequal probability sample, we estimate the same quantity as when we use OLS in a complete census of the population. In this sense, we can say that a survey-weighted regression is representative.

However, the approach works less well if we think of the regression equation:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

as representing a *structural* relationship. In our example, a relevant structural relationship is the causal effect of education on income. If this effect β_1 is the same for everyone, as the regression equation suggests, then the sample OLS and WLS estimators are both unbiased and consistent. But the Gauss-Markov Theorem tells us of a downside to using WLS: the OLS estimator is efficient, whereas the WLS estimator is not. If both estimators are unbiased and consistent but only one is efficient, we should prefer the efficient one. OLS beats WLS here.

Now suppose the structural relationship is heterogeneous. In our example, the causal effect of education

on income might differ between urban and rural areas. We can represent this situation by adding i subscripts to the coefficients:

$$y_i = \beta_{0i} + \beta_{1i} x_i + \varepsilon_i$$

Now each individual has their own β_{1i}. In the urban-rural example, we might have $\beta_{1i} = \beta^u$ if individual i lives in an urban area and $\beta_{1i} = \beta^r$ if individual i lives in a rural area. But more generally, the effect of education on income might differ by cognitive ability, family income, and so on.

When slopes are heterogeneous, a natural quantity of interest is the average slope in the population $\overline{\beta}_1$. Many analysts of survey data apply survey weights to regression estimation because they believe that doing so yields unbiased or consistent estimates of $\overline{\beta}_1$. In general, it does not. This point is not really about survey weights, but instead about how the population regression coefficient β_1^{POP} is not an unbiased or consistent estimator of $\bar{\beta}_1$. To see this point, let's express β_1^{POP} in terms of its underlying covariances and variances:

$$\beta_1^{POP} = \frac{cov(x_i, y_i)}{V[x_i]} = \frac{cov(x_i, \beta_{0i} + \beta_{1i} x_i)}{V[x_i]} = \frac{cov(x_i, \beta_{0i})}{V[x_i]} + \frac{cov(x_i, \beta_{1i} x_i)}{V[x_i]}$$

When β_0 and β_1 are constant, it is straightforward to show that the expression on the right simplifies to β_1. However, now that they vary, β_{0i} and β_{1i} may be correlated with x_i, and the expression on the right does not simplify in general. In the special case of a randomized controlled trial, however, the math does work out. In particular, if x_i is randomly assigned, then it is independent of β_{0i} and β_{1i}, so that $cov(x_i, \beta_{0i}) = 0$ and $cov(x_i, \beta_{1i} x_i) = \overline{\beta} V[x_i]$. In other circumstances, we are forced to ask: are we interested in the regression coefficient that would obtain in a census of the population, or are we interested in the average structural parameter?

When the form of heterogeneity in β_{1i} is known, we can estimate $\overline{\beta}_{1i}$ by fitting a model with interaction terms, or by fitting separate models for groups with difference β_{1i}'s. In the urban-rural example, we could run separate regressions in urban and rural areas and then take the weighted average of the sector-specific slopes:

$$\hat{\overline{\beta}}_1 = \frac{N^u}{N} \hat{\beta}_1^u + \frac{N^r}{N} \hat{\beta}_1^r$$

If β_{1i} varies only between urban and rural areas, then this weighted average is an unbiased and consistent estimator of $\overline{\beta}_1$. In general, the weighted average gives a different answer from a pooled regression of income on education.

This discussion should highlight that there is no single correct way to weight in regression, but instead a correct way to think about weighting decisions and their tradeoffs. In an unequal probability sample from a population with *heterogeneous* slopes, $\hat{\beta}^{WLS}$ is a good estimator for β_1^{POP} but not $\overline{\beta}_1$. In an unequal prob-

ability sample from a population with *homogeneous* slopes, both $\hat{\beta}^{OLS}$ and $\hat{\beta}^{WLS}$ are acceptable estimators for β_1^{POP} and $\overline{\beta}_1$, but $\hat{\beta}^{OLS}$ has a smaller variance under the Gauss-Markov assumptions.

This section introduced the idea of *parameter heterogeneity*, perhaps for the first time in your statistics and econometrics training. The idea that slopes or causal effects may differ across individuals may seem abstract or strange. We will return to it later in the quarter when we discuss the Rubin Causal Model.