Understanding and Teaching Unequal Probability of Selection

Humberto Barreto and Manu Raghav*

DePauw University

October 30, 2011

Corresponding author:

Manu Raghav
7 E. Larabee St
Harrison Hall 207
DePauw University
Greencastle, Indiana 46135
Email: manuraghav@depauw.edu
Phone: 540-460-7935
FAX: 765-658-1044

We thank the participants at the 2011 AEA Teaching Conference, especially our discussant Roisin O'Sullivan, and Frank M. Howland for helpful criticisms and suggestions.

*Author names are arranged in alphabetical order.

Abstract

This paper focuses on econometrics pedagogy. It demonstrates the importance of including probability weights in regression analysis using data from surveys that do not use simple random samples (SRS). We use concrete, numerical examples and simulation to show how to effectively teach this difficult material to a student audience. We relax the assumption of simple random sampling and show how unequal probability of selection can lead to biased, inconsistent OLS slope estimates. We then explain and apply probability weighted least squares, showing how weighting the observations by the reciprocal of the probability of inclusion in the sample improves performance. The exposition is non-mathematical and relies heavily on intuitive, visual displays to make the content accessible to students. This paper will enable professors to incorporate unequal probability of selection into their courses and allow students to use best practice techniques in analyzing data from complex surveys. The primary delivery vehicle is Microsoft Excel®. Two user-defined array functions, SAMPLE and LINESTW, are included in a prepared Excel workbook. We replicate all results in Stata® and offer a do-file for easy analysis in Stata. Documented code in Excel and Stata allows users to see each step in the sampling and probability weighted least squares algorithms. All files and code are available at *www.depauw.edu/learn/stata*.

Keywords: unequal probability, complex survey, simulation, weighted regression

1. Introduction

Given a population of size *N*, if each observation has the same chance of being selected, then we have a simple random sample (SRS). The properties of common statistics (e.g., the average or an ordinary least squares (OLS) regression slope estimate) based on data from such a sample (including sampling with replacement (WR) versus without replacement (WOR)) are well understood. Extensions to the basic model, such as heteroscedastic or autocorrelated errors, continue to assume SRS. Without question, simple random sampling is a core assumption of econometrics pedagogy.

Unfortunately, in sampling, simple does not mean easy to implement. It is exceedingly hard and often prohibitively expensive to obtain a truly simple random sample. In fact, most samples are actually produced by oversampling certain groups or areas, which means that others are less likely to be chosen. Since each observation does not have the same chance of being selected, the usual formulas do not apply. Correctly analyzing data generated by non-simple random sampling is rarely, if ever, explained to students—the complication is simply ignored.

This paper shows under which conditions and exactly where unequal sampling causes problems with OLS. This is important because all large, public data sets, such as the Current Population Survey (CPS), which are widely used in empirical research, employ complex survey designs that are not simple random samples.

To maintain focus on unequal probability of selection, this paper will not consider an allied issue: the effect on OLS estimated standard errors (SEs) from the use of cluster and stratified sampling applied in real-world surveys. This design effect is also important and properly using data produced by complex surveys requires that both unequal probability and design be incorporated in the analysis. We will focus only on the effects of unequal probability of selection to emphasize that these are two separate issues and to keep this paper reasonably short. Furthermore, since failure to incorporate unequal probability of selection can lead to biased and inconsistent OLS estimates while ignoring the survey design merely affects the estimated precision, correctly handling the unequal probability is more important than the design effect.

Our primary motivation is to provide a clear, intuitive presentation that can be used in an undergraduate econometrics course and we provide suggestions for bringing these ideas into the classroom. We use Microsoft Excel® to illustrate the data generation process and Monte Carlo simulation to demonstrate properties and claims. Given our target audience, we eschew mathematical formalism in favor of concrete examples that enable strong visual exposition. We include user-defined Excel functions in a prepared workbook to sample with unequal probabilities and estimate probability weighted regressions. In addition, we include Stata code to enable replication and presentation of the material in Stata. All files used in this paper are available at *www.depauw.edu/learn/stata*.

The next section explains the data generation process and its implementation in Excel. Section 3 discusses random sampling with equal and unequal probabilities of selection. Sections 4 and 5 explore the sampling distributions of OLS and probability weighted least squares, respectively, under several sampling schemes. Section 6 replicates the analysis in Stata. Having explained the material, we offer learning objectives and suggestions for classroom use in section 7 and conclude with section 8.

2. Generating a Finite Population

We are interested in estimating the slope coefficient of the relationship $Y = \beta_0 + \beta_1 X$ in a finite population of *N* (*X*, *Y)* pairs. We proceed by taking a random sample of size *n* from the population, then regressing *Y* on *X* to obtain $Y = b_0 + b_1 X$. Our focus is on the sampling distribution of b_1.

To see a concrete presentation of the way the finite population is generated, download and open the Excel file EqualUnequalProb.xls from *www.depauw.edu/learn/stata*. Be sure to enable macros when opening this workbook so that the functions and buttons in the file are operational. The workbook contains two crucial user-defined (not included in a standard Excel installation) functions, SAMPLE and LINESTW. They are both array functions. To enter a formula as an array function in Excel, you must simultaneously press the Ctrl, Shift, and Enter keys. Press the Esc key if you are in a cell that is part of an array and want to exit without making any changes. Excel uses curly brackets, {}, to convey when formulas are part of arrays.

The *DGP* sheet begins with parameters in cells B2 and B3 that were used to produce the finite population observations (in columns C and D). To add further concreteness to the problem, we implemented a rudimentary earnings function (without a log functional form or heteroscedasticity). For years of schooling, the formula in column C was =ROUND(RAND()*5+11.5,0); the wage (in column D, in units of dollars per hour) was =ROUND(slope*C2+NORMINV(RAND(),0,SD),2). The relative reference, C2, changes row value and always refers to the cell to the left, the number of years of schooling, as the formula is filled down. NORMINV(RAND(), 0, 2) produces random draws from a normal distribution with mean zero and a constant standard deviation of 2. Finally, the cell range C2:D1001 was copied and then pasted as values, yielding the finite population data. Samples will be taken from the finite population in columns C and D.

The chart fits the population regression line to the 1,000 observations in the population. The population regression equation is $Wage = -0.6232 + 2.02565 * Schooling$. The tables below the chart report summary information on the population.

We are not interested in estimating the slope parameter of the DGP, with a value of 2. Instead, our focus is on estimating the finite population regression slope, with a value of approximately 2.02565. We will sample from the 1,000 observations and regress Wage on Schooling in the sample data. We want to know how well conventional regression, OLS, performs under various conditions, but first we must examine how a sample is produced from the finite population.

3. Random Sampling with Equal and Unequal Probabilities

We will use three different probabilities of selection: (1) equal probability (column K) which is also known as simple random sampling, (2) unequal probability based on the independent variable, *Schooling* (column O), and (3) unequal probability based on the dependent variable, *Wage* (column P). We will show that OLS works well in the first two cases, but breaks down when unequal probabilities depend on the dependent variable. We will follow standard practice and sample with replacement. If we replace observations as they are sampled, then the probabilities of selection stay the same in each draw. This greatly simplifies computing each element's probability of inclusion in a sample.

Scroll right (if needed) to see three alternative probabilities of selection on the first draw. Column K uses a formula, =1/1000 (because there are 1,000 observations in the finite population), to indicate that each observation has an equal likelihood, 0.1%, of being chosen on the first draw. This equal probability of selection is the hallmark of a simple random sample.

Columns O and P are both based on unequal probabilities of selection on the first draw, but they differ in an important respect. In column O, an IF statement is used to assign the probability based on the value of *Schooling*, the independent variable. For example, each of the 196 observations in the population with 12 years of schooling have a 0.3% chance of being chosen on the first draw; while each of the 206 observations with 16 years of schooling have only a 0.02% chance. In column P, the unequal probabilities are based on the dependent variable, *Wage*. The 81 observations with low wages, less than 20, are the ones with high, 0.5%, probabilities of selection in the first draw. As wage increases, the probability of selection falls.

Next to column M, which has the probability of selection on the first draw based on *Schooling* and *Wage*, the probability of being included in the sample of 100 observations (with replacement) is computed (in column N). The calculation of the probability of inclusion given probability of selection on the first draw is explained in Section 5, when unequal probabilities of selection are used.

The next area of the sheet (scroll right, if needed) shows a single sample, two sets of regression results (OLS and probability weighted least squares), and a chart. The sample is in cell range R2:T101. Each of the 100 observations in the range was pulled from the C2:D1001 range. If the sample was chosen with replacement, then an individual observation from the population may appear more than once in the sample; sampling without replacement removes a chosen observation from the population so it can appear only once.

The sample is produced with a user-defined array function, SAMPLE, that is included in the workbook. This function inputs a population (range C2:D1001) and probabilities of selection in the first draw (column K, O, or P) and outputs a sample from the given population. To use the function, the cell range R2:T101 was selected. The SAMPLE function requires that the number of selected rows (100, in this example) must equal the sample size and the number of selected columns (3) must equal the number of independent (*X)* variables plus one (for the *Y* variable) plus one (for the probabilities column).

With the cell range selected, the formula, =SAMPLE(C2:D1001,K2:K1001,100,1), was entered in the formula bar. The SAMPLE function has four arguments: a cell range with the population data, a cell range with probabilities of selection on the first draw, the sample size, and an optional value, 0 or 1, for whether the sample is taken without or with replacement, respectively. For example, by using cell range K2:K1001 and 1 for the optional argument, we are specifying a simple random sample with replacement (SRSWR).

Because array functions may be unfamiliar and their behavior often frustrating, we repeat that, like all Excel array functions, SAMPLE requires the keystroke combination Ctrl-Shift-Enter (instead of simply the Enter or Tab key) to input the formula. Click the Esc key to exit an array function without making any changes.

The *SampleFn* sheet offers a behind-the-scenes look at the SAMPLE function. Knowledge of how the function is coded is not necessary to use the workbook, but the detailed explanation of how the function works and ready access to the source code provides information for those interested in understanding the details or extending the SAMPLE function.

From the *DGP* sheet, with columns R to AB visible, press the F9 key to recalculate the sheet and draw another sample via the SAMPLE function. The values displayed in columns S and T, along with the chart, will change every time you press the F9 key. Actually, the Prob column (R) is changing also, but since all observations have the same 1/1000 probability of being chosen on the first draw, the column values displayed remain the same.

Notice also that the regression results in cell ranges V3:W7 and Z3:AA5 change as you press F9. The former uses Excel's native LINEST array function to compute the OLS fit (in the first row), estimated standard errors (in the second row) and basic regression diagnostics (R^2 in cell V5, root mean squared error (RMSE) in W5, whole model F statistic in cell V6, model degrees of freedom in cell W6, regression sum of squares in cell V7, and SSR in cell W7). The results in the latter cell range are produced by LINESTW, a user-defined function built-in to the workbook that enables probability weighted least squares (probWLS) estimation. LINESTW reports the same information as LINEST, but omits the last two rows because these statistics are not meaningful in a probability weighted regression. The coefficient estimates (cells V3:W3 and Z3:AA3) are the same when equal probabilities of selection (in column K) are used.

The chart displays a scatter plot of the sample data, with the red line indicating the population regression function. The population regression line remains constant as new samples are drawn. The Refresh button recalculates the sheet and updates the PivotTable under the chart.

The cells of special interest, of course, are V3 and Z3. These are the slope (b_1) estimates from OLS and probWLS applied to the sample data. As usual, we are interested in the sampling distribution and properties of these estimators. We first study OLS, then turn our attention to probWLS.

4. The Sampling Distribution of the OLS Slope Estimator

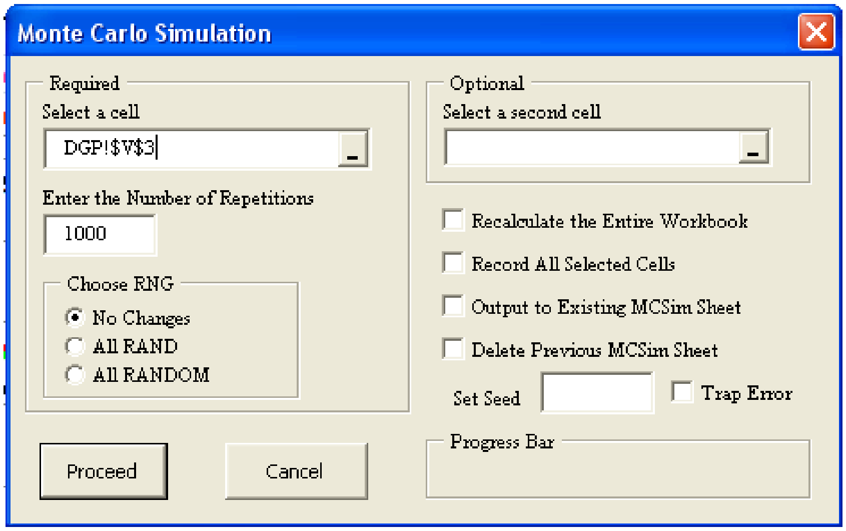
Begin analysis of the performance of the OLS slope estimator by clicking the Show OLS button (near cell AD9). The black line added to the chart is the OLS fitted line, with intercept and slope given by cells W3 and V3, respectively. Click the Refresh button a few times and notice how well OLS, using the 100 sampled observations, seems to be doing. The black OLS line stays quite close to the red (true) population regression line. Below the chart, cells Z24:Z28 show that the 100 observations are roughly 10% of the number of observations for a given level of *Schooling*.

We will evaluate the performance of the OLS slope estimator via simulation. By repeatedly resampling from the finite population (using the SAMPLE function), applying OLS to each sample (using the LINEST function), and keeping track of the results, we obtain an approximation to the exact sampling distribution, expected value, and standard error. The workbook has code based on Barreto and Howland (2010) that makes it easy to conduct Monte Carlo simulation analysis.

Click the MC Sim button (cell AD 15) to bring up a dialog box, shown in Figure 1. In the upper left corner, click inside the “Select a cell” input box and remove the contents, then click on cell V3 on the *DGP* sheet so that Excel enters the cell address (as shown in Figure 1). Click Proceed to run a simulation.

The simulation does 1,000 presses of the F9 key, taking 1,000 samples, computing the fitted line in each sample and keeping track of the 1,000 OLS slope estimates (in cell V3). The results, displayed in Figure 2, are contained in an *MCSim* sheet. Of course, your results will be similar, but not exactly the same as Figure 2 because your simulation is based on a different set of 1,000 samples.

In Figure 2, the average of the 1,000 slope estimates, 2.039, is an approximation to the expected value (EV) of the OLS estimator of the finite population slope. The fact that it is close to the population slope, approximately 2.02565, is evidence that the OLS slope estimator is performing well. A rough gauge of the variability in the average can be computed as $SD \;/\; \sqrt(\#reps)$. With 1,000 repetitions, 0.3456/sqrt(1000) $\approx$ 0.01. This means we can rely up to the first decimal place in the average reported in Figure 2. To improve precision, we can run simulations with more repetitions. In fact, OLS is a biased, but consistent estimator of the finite population conditional mean function because it is not exactly linear, in our example.



Monte Carlo Simulation

Required

Select a cell: DGP!$V$3

Enter the Number of Repetitions: 1000

Choose RNG

● No Changes

○ All RAND

○ All RANDOM

Optional

Select a second cell:

□ Recalculate the Entire Workbook

□ Record All Selected Cells

□ Output to Existing MCSim Sheet

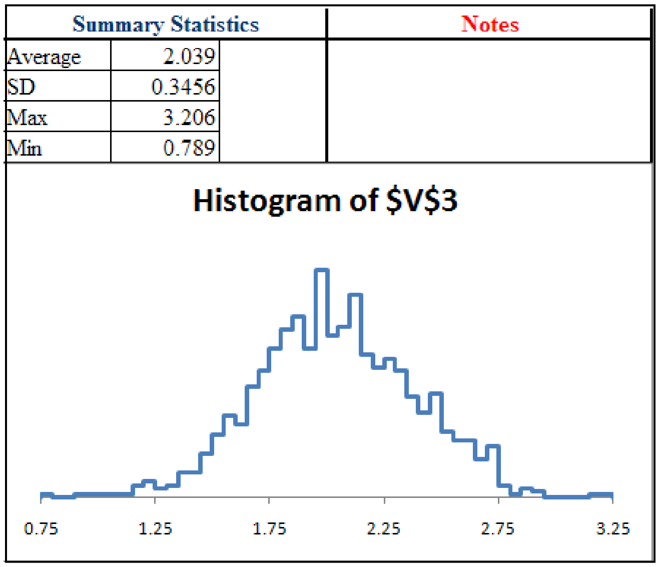
□ Delete Previous MCSim Sheet

Set Seed:

□ Trap Error

Progress Bar

Figure 1: Click the MC Sim button to run a simulation.



Summary Statistics		Notes
Average	2.039	
SD	0.3456	
Max	3.206	
Min	0.789	

Figure 2: Simulation of OLS slope with equal probabilities.

The SD of the 1,000 slope estimates, 0.3456, is an approximation to the exact standard error (SE) of the OLS slope estimator. The maximum and minimum OLS slope estimates are also included in the summary statistics for the simulation. The chart displays an empirical histogram of the 1,000 sample slopes, which is an approximation to the exact sampling distribution of the OLS slope estimator. To improve the approximation, increase the number of repetitions (or samples).

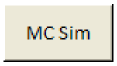
To see if the OLS estimated standard error is performing well, run a simulation that tracks cell V4. Your results should show that the average estimated SE is quite close to the SD of the 1,000 OLS sample slopes (0.3456 in Figure 2, which is an approximation to the exact SE). This means that our simulation results support the claim that the OLS estimated SE from a single sample (cell V4) is producing good estimates of the OLS exact SE. The fact that the OLS slope and estimated SE are working as advertised should not be surprising since we are working with a SRSWR design that meets the usual classical linear model requirements.

Confusion is to be expected when simulating cells V3, the estimated slope, and V4, the estimated SE. To be clear, we are working with three standard errors: (1) the exact SE, a constant that utilizes the population SD of the error distribution, (2) the approximate SE, produced by computing the SD of the estimated slopes generated by simulation, and (3) the estimated SE, a random variable based on the estimated SD of the error distribution. Each new sample produces a new estimated slope and a new estimated SE. There are two ways we are using the 1,000 repetitions in the simulation: (1) the average of the estimated slopes (cell V3) is compared to the population slope to determine if the estimated slope is biased and (2) the average of the estimated SEs (cell V4) is compared to the SD of the estimated slopes (which we call the approximate SE) to determine if the estimated SE is a good estimator of the exact SE (using the simulation's approximate SE as a proxy). We have just seen that with SRS, OLS performs well.

We can explore the effect of sampling with unequal probabilities by changing the arguments in the SAMPLE array function in cells R2:T101. This can be done manually (directly editing the formula in cell R2, changing the second argument from K2:K1001 to O2:O1001 and pressing Ctrl-Shift-Enter), but it is easier to simply click the radio button labeled Unequal (Schooling), near cell AA27.

With unequal probabilities of selection, the graph looks markedly different than before. In fact, the chart is not a simple Excel Scatter chart, but a Bubble chart, and the size of the bubbles (markers) is based on the Weight (in column U). OLS ignores the bubble size and considers each observation as having equal weight; we will see that probability weighted least squares incorporates the weight in computing the coefficient and SE estimates.

With columns R to AB visible on the *DGP* sheet and with the OLS fitted line displayed in the chart and "Unequal (Schooling) Prob, WR" title in the first row, press F9 repeatedly. OLS seems to be bouncing around the finite population regression line (in red) that we are trying to estimate. You may notice that the markers are small on the left and large on the right of the chart, but remember that OLS ignores this information and considers each weight equally.

Click the  button to track cell V3 and confirm that unequal probabilities based on *Schooling* (the *X* variable) do not bias the OLS estimator of the sample slope. Notice, however, that the SD of the sample slopes (an approximate SE) is around 0.44, or about 25% larger than the approximate SE in our first simulation (in Figure 2). The effect is large enough to see by pressing F9 repeatedly and observing the increased bounce in the OLS fitted line relative to the equal probability case. To return the DGP to the initial SRSWR case, click on the Equal radio button and then press F9 repeatedly.

The OLS exact SE in a bivariate regression can be expressed as $\frac{SD_{\varepsilon}}{\sqrt{n}SD_{X}}$. The OLS estimated SE simply replaces the numerator with an estimate of the SD of the error distribution, the sample RMSE. This formula offers a clear explanation of the increase in the variability of the OLS slope estimator (from roughly 0.35 to 0.44) as probabilities were changed, ceteris paribus, from Equal (SRS) to Unequal (Schooling). The SD of *Schooling* is the key. Under Equal probabilities of selection, the 100 observations in the sample are roughly evenly spread across the five values of schooling. Unequal (Schooling) probabilities create samples where lower values of schooling are more common and the number of observations tends to fall as schooling rises. The PivotTable under the chart makes this clear. Switch back and forth from Equal to Unequal (Schooling) (by clicking the appropriate radio buttons) and notice how the distribution of *Schooling* values changes. A consequence of the Unequal (Schooling) sampling scheme is that the SD of *Schooling* is lower compared to the SRS case.

Cell S104 computes the SD of *Schooling*. Simulation of cell S104 shows that with the probability set to Equal (SRS) the SD of *Schooling* is about 1.41 and it falls to roughly 1.15 with Unequal (Schooling). This fall in the SD of *Schooling* matches the rise in the approximate SE from 0.35 to 0.44. Unequal (Schooling) leads to higher variability in the OLS slope estimator compared to SRS because the SD of *Schooling* falls.

This analysis makes clear, however, that the result is not guaranteed. Were we to concoct an unequal probability of selection that favored smaller and larger *X*s (12 and 16 in our example), the SD of *X* would rise and the OLS SE of the slope would fall compared to SRS.

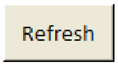
To this point, we have offered evidence from simulation that OLS does a good job of estimating the slope with unequal probabilities that depend on years of schooling and that the SE of the OLS slope estimator depends on the SD of *Schooling*. We now turn our attention to the sample estimate of the variability of the estimated slope and ask, can we rely on the OLS estimated SE? With the probability of selection set to Unequal (Schooling), run a simulation of the OLS estimated SE (cell V4) to find out. As before, we compare the average of the OLS estimated SEs (tracking cell V4) to the approximate SE obtained from the previous simulation (tracking cell V3 and using the SD of the sample slopes). While agreement is not perfect, the average estimated SE (cell V4) is about 0.42 and the approximate SE is a little higher at 0.43, the OLS estimated SE is measuring the variability of the slope reasonably well.

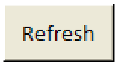
So, in this example, OLS behaves as expected under both equal and unequal probabilities of selection based on *Schooling*, but what about unequal probabilities of selection based on the dependent variable? To answer this question, click the radio button labeled Unequal (Wage), near cell AA27, or edit the array formula in cell R2, making the formula =SAMPLE(C2:D1001,P2:P1001,100,1) and press Ctrl-Shift-Enter. The chart immediately shows that OLS is broken. It is no longer bouncing around the population regression line. Keeping an eye on cell V3 while repeatedly pressing F9 shows that it is too often under 2, the population parameter that is being estimated. (The estimated intercept is also way off, on average, but we will continue to focus on the slope.) A simulation (tracking cell V3) reveals that the OLS slope estimator is now centered near 1.6, not the population slope of 2 that we are trying to estimate.

This finding, i.e., that the OLS slope estimator is biased when unequal probabilities are a function of the dependent variable, is an important point. If we use OLS on a sample that is produced by Unequal (Wage), sampling observations with low wages more often than those with high wages, we will systematically under shoot our target. The inaccuracy of OLS has led to the development of an alternative to OLS, probability weighted least squares.

The workbook enables exploration of the consistency properties of OLS. Scroll right to an empty area of the sheet and use the SAMPLE function with 50 observations (=SAMPLE(C2:D1001,P2:P1001,50,1)) and 150 observations (=SAMPLE(C2:D1001,P2:P1001,150,1)). Use LINEST on each sample to compute the estimated slope and then run simulations that track the slopes from the two sample sizes. To compare your results to ours, download the Excel file EqualUnequalProbConsistency.xls from *www.depauw.edu/learn/stata*.

Before leaving OLS and turning our attention to probWLS, we take a moment to consider why OLS works with Unequal (Schooling) and not Unequal (Wage). The chart can be used to provide an intuitive explanation. Remember that OLS ignores marker size, treating all observations equally. We will use the marker size, which is based on the reciprocal of the probability of inclusion, to understand how the samples produced are different under different sampling schemes.

Select the Unequal (Schooling) radio button to sample based on schooling. Now watch the markers in the chart carefully as you click the  button repeatedly. The markers start small (on the left) and gradually increase in size as schooling rises, reaching their largest values in the rightmost vertical strip. We can conclude that with Unequal (Schooling), a sampling scheme based on the value of *Schooling*, we will get more observations in certain vertical strips and fewer observations in other vertical strips. With homoscedastic errors (which we have), the distribution of values in each strip will have the same spread. Thus, in essence, what unequal sampling based on the independent variables does, in our example, is to more or less densely populate a given vertical strip, but this does not bias the estimated slope (however, as we know, the SE is affected) because the spread within a strip remains unaffected. An extreme case illustrates the point: suppose that the probability of sampling an observation with 16 years of schooling was zero. We would then have 100 observations with values of schooling from 12 to 15. OLS would remain undamaged in the sense that it would accurately estimate the population slope, but its precision would fall. The sampling distribution of the OLS slope estimator would remain centered on the finite population slope, but the histogram would be more spread out.

Now select Unequal (Wage) to force use of the probabilities in column P and watch the markers in the chart carefully as you click the  button repeatedly. The visual effect is dramatic. Instead of bubble size increasing from left to right, it increases from bottom to top. We are now getting many observations with low wages, which are concentrated in the bottom left, and relatively fewer observations as wages rise. This explains why the OLS line breaks off and does such a poor job of estimating the population fitted line. With too few high wage observations, which are concentrated in the top right (more schooling leads to higher wages), the OLS fitted line ends up too flat. Unlike selection based on *X*, which maintains the given distribution of *Y* values in a vertical strip (altering only the number of observations in a vertical strip), probability of selection based on *Y* is a horizontal effect (as wage rises, the chances of being selected fall) which alters the distribution of *Y* values in a vertical strip, effectively destroying the ability of OLS to accurately estimate the population regression line.

This visual analysis offers a clear hint to the upcoming success of probWLS—by putting more weight on the rarely sampled high wage observations that are more likely to be found in the upper right area of the chart (and depicted with large markers), the fitted line will be steeper and, once again, we will have access to an accurate estimator.

5. Probability Weighted Least Squares (probWLS)

Begin by clicking the Equal radio button (near cell AA27). Next, click the Show probWLS button. An additional line, probWLS, has been added to the chart, but it is exactly the same as OLS with SRSWR (compare V3 to Z3 and W3 to AA3). Click the Unequal (Schooling) radio button (near cell AA27) to change the sampling scheme from SRSWR to unequal probabilities based on *Schooling*, the independent variable. Click the Refresh button repeatedly to see that OLS and probWLS are now slightly different. To understand probWLS, we first have to understand the concept of weights. After explaining how probability weights are calculated and interpreted, we explore the sampling distribution of the probWLS slope estimator under varying scenarios, and then explain how probWLS is computed.

The probability weight is the reciprocal of a single observation's probability of inclusion in the sample. If sampling is done with replacement, this weight can be computed exactly because the probability of inclusion equals $1 - (ProbNotSelectedinaSingleDraw)^n$. When sampling without replacement and with unequal probabilities of selection, there is no analytical formula for computing the probability of inclusion from the probability of selection in the first draw. The computation required for ever-changing conditional probabilities, as observations are removed from the population and change an observation's probability of selection, quickly becomes extremely cumbersome as the sample size increases. Beginning in cell M14, the *SampleFn* sheet offers a simple example of the exact and approximate (via simulation) probabilities of inclusion with and without replacement. In practice, the effect of WOR is ignored and weights are calculated based on WR. The workbook ExploringWRvWOR.xls (available at *www.depauw.edu/learn/stata*) shows that the probability of inclusion based on WR is close to WOR even in our example. In a real-world application (such as the Current Population Survey), small probabilities of selection, large finite population sizes, and relatively small sample sizes guarantee little difference between WR and WOR probabilities of inclusion.

Each weight indicates how many observations in the population are represented by that observation in the sample. Observations with small probabilities of inclusion are unlikely to be chosen so when one is selected, it stands to reason that it be assigned greater importance. For example, consider the ninth observation in the population, with 12 years of schooling and a wage of 20.11 (in cells C10 and D10, respectively). This observation's probability of selection on the first draw with unequal probabilities based on *Schooling* (column O) is 0.3%. Cell N4 shows that the probability of this observation being included in a sample (chosen with replacement) of 100 observations is roughly 25%. Click the Track 9th Obs button (to the left of cell O10) to easily see if this observation is included in the sample. Press F9 or click the Refresh button until you see a strip of green colored cells in the R2:U101 range. (With a 25% probability of inclusion, you should sample the 9th observation reasonably soon.) When the 9th (or any *Schooling* = 12) observation in the population is chosen, it represents only about four observations in the population because it is common to sample such observations. The 9th observation will be plotted in the chart with a very small marker, like the other sample observations with 12 years of schooling, because they are all relatively likely to be chosen and, therefore, have small weights.

Notice the markers associated with 16 years of schooling. They are large because these observations are relatively unlikely to be chosen. When an observation with 16 years of schooling is chosen, it has a relatively large weight of about 50 (which is the reciprocal of its probability of inclusion of nearly 2%), meaning it represents 50 individuals in the population. You can find just a few observations with 16 years of schooling in any given sample, but many with 12 years of schooling.

The performance of probWLS can be evaluated via simulation. With Unequal (Schooling) selected (so that range O2:O1001 is used in the SAMPLE function for cells R2:T101), track cells V3 (OLS) and Z3 (probWLS) by including Z3 in the "Select a second cell" input box in the Monte Carlo simulation dialog box. Your results will be similar to Figure 3.

Simulation suggests that both OLS and probWLS are working well (both averages in Figure 3 are close to 2.02565, the finite population slope), but the smaller approximate SE produced by OLS (0.4351 versus 0.5367 in Figure 3) is evidence that OLS is more precise and, therefore, preferable to probWLS.

DGP!$V$3		DGP!$Z$3		Notes
Average	1.998	Average	2.030	
SD	0.4351	SD	0.5367	
Max	3.564	Max	3.967	
Min	0.579	Min	0.143	

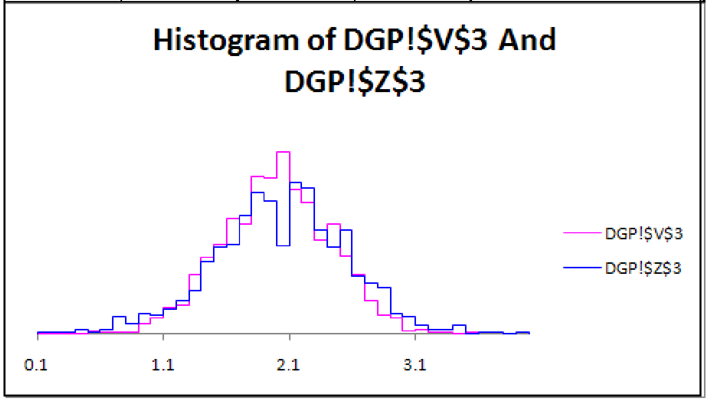


Figure 3: Comparing OLS and probWLS with Unequal (Schooling) probabilities.
Note: $V$3 = OLS slope; $Z$3 = probWLS slope.

DGP!$V$3		DGP!$Z$3		Notes
Average	1.581	Average	1.971	
SD	0.3829	SD	0.7043	
Max	2.865	Max	4.315	
Min	0.279	Min	-0.602	

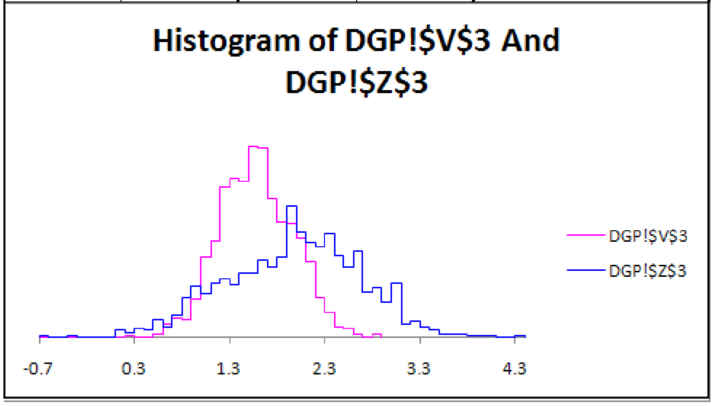


Figure 4: Comparing OLS and probWLS with Unequal (Wage) probabilities.
Note: $V$3 = OLS slope; $Z$3 = probWLS slope.

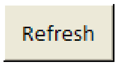
Click the Unequal (Wage) radio button (near cell AA27) to change the sampling scheme from having the unequal probabilities be a function of *Schooling*, the independent variable, to *Wage*, the dependent variable. Click the  button repeatedly to see that probWLS is now clearly superior to OLS.

Figure 4 shows the results of a simulation tracking both V3 and Z3, the OLS and probWLS slope estimators, under Unequal (Wage) sampling. The conclusion from the previous section, that OLS is broken, is confirmed and there is evidence that probWLS performs well because the average of 1,000 probWLS slopes is near 2.02565, the population parameter.

Unfortunately, a simulation of the probWLS estimated SE (cell Z4) does not yield such positive results. On average, the probWLS estimated SE is roughly 0.55, which is too low compared to the approximate SE of 0.7 from Figure 4. The LINESTW function is using a Taylor linearization (which is the default approach used by commercial software such as STATA and SAS), but the complex survey literature offers many alternatives for estimating the probWLS SE, including a variety of bootstrapping approaches. The poor performance of the probWLS estimator of the SE helps explain the presence of so many competing estimators.

The secret to the success of the probWLS slope estimator and failure of OLS lies in the way they utilize the data. While OLS ignores the bubble size in the chart, probWLS weights the observations according to the inverse of their probability of inclusion in the sample. With Unequal (Wage) probabilities, relatively few observations will appear in the top right corner of the scatter plot. Probability weighted least squares compensates for this by assigning a large weight to the few observations that do appear at high schooling, high wage points and this pulls the fitted line up, correcting the tendency to get too shallow a slope.

Figure 5 illustrates the logic. The probWLS fitted line does a better job estimating the population regression function because the three large markers above a wage of 32 have only a 1% chance of being selected and, thus, have a weight of 100. The small markers below 20 have a roughly 40% probability of inclusion and, thus, a weight of about 2.5. In fitting the line, the probWLS algorithm treats the three observations that are unlikely to be chosen as 40 observations each compared to the observations with wages less than 20. Weighting the observations according to the inverse of the probability of inclusion compensates for the fact that high wage, high schooling observations are much less likely to appear in the sample than the low wage, low schooling observations.

We conclude this section by reviewing the computation underlying probWLS. While OLS fits the regression line according to the familiar $(X'X)^{-1}X'y$ matrix multiplication, probWLS inserts *W*, an *n* x *n* diagonal matrix of weights between the *X* matrices: $(X'WX)^{-1}X'Wy$. Each weight is composed of the reciprocal of the probability of inclusion divided by the number of observations in the sample. OLS assumes SRS and implicitly treats *W* as an identity matrix because each probability of inclusion is 1/*n*, so that each weight is *n*/*n*. Unequal probabilities of inclusion guarantee that *W* will not be an identity matrix. The effect is obvious—instead of treating each observation equally, observations with greater weight exert greater influence in computing the slope estimate and this explains the superiority of probWLS over OLS.

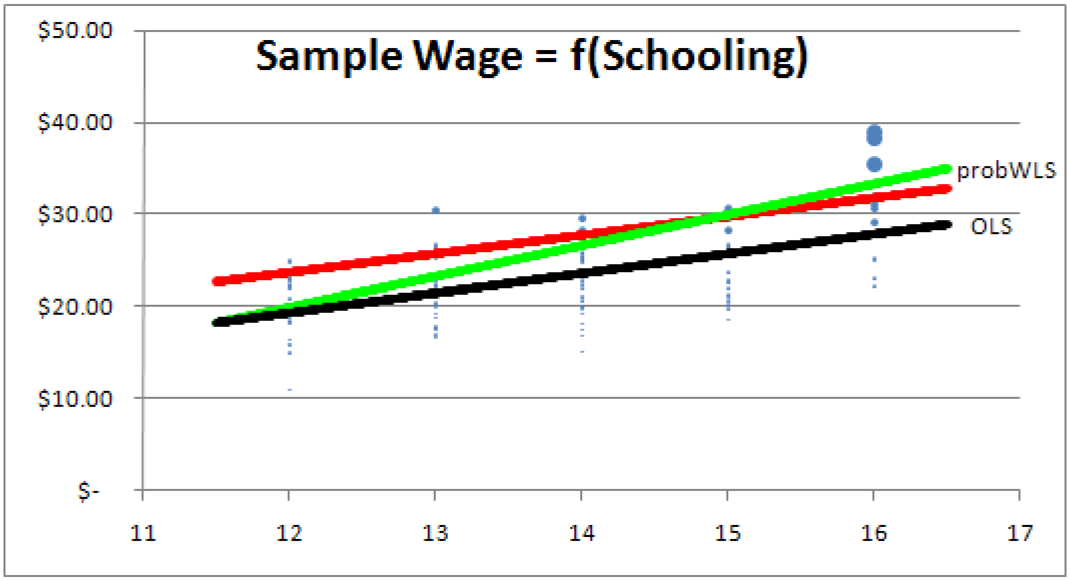


Figure 5: Giving greater weight to unlikely observations is how probWLS beats OLS.

The computation for the probWLS estimated SE is not as straightforward. The probWLS estimated SE is a non-linear function that is linearized using a Taylor series expansion. Instead of the usual variance-covariance matrix, $\sigma^2(X'X)^{-1}$, probWLS uses a sandwich estimator, $(X'WX)^{-1}\ V(X'WX)^{-1}$. With SRS, the *V* matrix collapses to the variance of the errors times *(X'X)* and *W* is an identity matrix. Substituting into the sandwich estimator, we get $(X'X)^{-1}\sigma^2(X'X)(X'X)^{-1}$, producing the usual result. Unequal probabilities, however, produce a *V* matrix that is a function of each observation's variance of errors (the same if homoscedastic), sampling weight, and *X* values. Our Excel LINESTW function (along with software that supports complex survey design, such as R, SAS, and Stata) produce estimated SEs based on the $(X'WX)^{-1} V(X'WX)^{-1}$ sandwich estimator be default.

In an inferential setting, the *V* matrix must be estimated, using the residuals. The standard approach is to compute, for each observation, the probability weight times its residual times its *X* (row) vector. The transpose of this vector (*k* x 1) is then multiplied by itself (1 x *k*), yielding a *k* x *k* matrix for each observation. The sum of these matrices is multiplied by a degrees of freedom adjustment, *n*/(*n*-1) and used as an estimate of the *V* matrix in the sandwich estimator of the variance-covariance matrix. In addition, if the DGP is based on a complex survey design, using strata and clusters, then additional steps are needed to properly compute the estimate of the *V* matrix.

6. Replicating Results in Stata

We will keep the explanation for implementing the above exercise in Stata short because we have already discussed in-depth the same issues when we presented its implementation in Excel. Replicating the results offers a check and allows users who are more comfortable with Stata to incorporate this material into their courses. We have included a do-file that replicates results of the Excel file and can be executed in an interactive and a non-interactive mode.

Begin by downloading the EqualUnequalProbPopData.dta data file and EqualUnequalProbProgram.do do-file from *www.depauw.edu/learn/stata*. The data file contains values from the finite population (columns C and D) in the Excel workbook. Open the Do-file Editor (from the Window menu, select "Do-file Editor," then "New do-file Ctrl-8"). From the Do-file Editor window, open the do-file by File -> Open and navigating to your download folder and selecting the do-file. Edit the fourth line, making sure that the correct pathname is given for where the data file is saved on the computer. Run the do-file by clicking on the "do" button in the toolbar of the do-file editor.

Return to Stata to see that the program is running. The user can select 1 for the interactive mode or 2 for the non-interactive mode. Typing any other number in the command window and pressing the Enter key will result in exiting the do-file. If the user selects the non-interactive mode by typing 2 from the keyboard and pressing the Enter key, then Stata will draw a sample of size 100, followed by running a regression using that sample, drawing a graph showing population, OLS, and WLS lines, and running Monte Carlo simulations with sample size of 100 and repetitions of 100 for each of the three probabilities of selection. These numbers can be easily changed by modifying the do-file in the do-file editor before executing the do-file. Stata will generate three different graphs while running the non-interactive mode, one for each of the three probabilities of selection. All graphs will appear in separately generated windows. The title of each window will indicate the corresponding probability of selection for that graph.

The interactive mode of the do-file can be initiated by typing 1 from the keyboard in the command window of Stata and pressing the Enter key. Type 1 and press the Enter key for running regressions, 2 for drawing graphs, and 3 for running Monte Carlo simulations.

If the user decides to choose either to run regressions or to draw graphs, the next question that Stata will ask is if the user wants to draw a new sample with replacement of size 100. If the user types “yes” from the keyboard and presses the Enter key, then Stata will ask which of the three probabilities of selections the user wants to choose by typing either 1, 2, or 3 and pressing the Enter key. Stata will draw a new sample of the specified size with replacement using the specified probability of selection.

Once a new sample is chosen, it will be used to run regressions and draw graphs. But the user can also type any other key in response to the question about drawing a new graph. This will make Stata use the previously drawn sample. This is useful if the user first wants to run regressions and then subsequently wishes to use the same sample to draw graphs. However, when the user starts an interactive mode initially without cycling through the interactive mode before, the user must draw a new sample as there is no previously drawn sample. If the user does not answer “yes” to the question, Stata will gently remind the user that in the first iteration, the user must answer “yes” to the question and draw a new sample.

Stata will run a regression based on the sample that was either freshly drawn or drawn beforehand. Based on the probability of selection that was used to draw samples, Stata will run regressions using the appropriate probability weights that correspond to that probability of selection. Initially Stata uses the command `svyset`, whose generic syntax is svyset _n [pweight=*variable_name*], where _n (Stata’s symbol for individual observations) is used here because, in this case, there are no clusters and each observation in the population itself is sampled (observations are the primary sampling units or PSUs). Once Stata has defined the design of the survey using the `svyset` command, one can use `svy` prefix in front of any command to incorporate the survey design. For example, to run regression we can use `svy: regress wage schooling` and Stata will automatically use the appropriate details of survey design.

If the user had chosen to run a regression, then a regression output will be chosen using the above command. After showing the regression output, Stata will ask if the user wants to continue the interactive mode. A “yes” answer to this question leads the interactive mode to repeat itself. Otherwise, Stata will exit the do-file.

If the user had decided to have Stata draw graphs, then another question will appear that will ask the user to select one of the following choices: a graph with only OLS and population regression lines along with the scatter plot between wage and schooling, a graph with only WLS and population regression lines along with the scatter plot between wage and schooling, or lastly a graph, which has OLS, WLS, and population regression lines along with scatter plot between wage and schooling. We are using the `lfit` (which is an abbreviation of linear fit) command of Stata to draw various regression lines and the `scatter` command to draw the scatter plot. These different types of plots can be combined together in one graph using the `graph twoway` command of Stata. Please look at Stata’s help files or Stata Corporation (2009) for `lfit`, `scatter`, and `graph twoway` commands for further details.

After Stata has drawn one of the graphs based on user specified preferences, Stata will ask if the user wants to draw other graphs (using the same sample). If the user types “yes” and presses the Enter key, the previously shown question that asked about which graph to draw will reappear. The user can then choose again to draw another graph. If the user types anything else in response to the question about drawing more graphs, then Stata will ask the user if the user wants to continue in the interactive mode. If the user types “yes,” the interactive mode will repeat itself once again. If the user types something else, the do-file will exit after reminding the user that the user can execute the do-file again to try other modes.

Finally, at the very beginning of the interactive mode, the user could have also chosen to run Monte Carlo simulations by selecting 3 and pressing the Enter key. The next question that Stata will ask is to specify the sample size that will be used the simulation. Stata will then ask how many repetitions to use for the simulation. The user will then select one of the three probabilities of selection by typing 1, 2, or 3 and pressing the Enter key. Stata will then draw the specified number of samples of specified sample size, run a regression for each sample, store the values of slope coefficients for each sample as scalars and then calculate the average and standard deviations of the slope coefficients. The average and standard deviation of the Monte Carlo simulation will then be displayed.

After Stata has finished a particular task in the interactive mode, at the very end it asks the user if the user wants to continue in the interactive model or exit the do-file. If the user types “yes” to this question, then the interactive mode starts again.

We used Stata to run 1,000 repetitions of sample size 100 with equal, unequal (schooling) and unequal (wage) probabilities. Our results replicated the results in Excel.

7. Learning Objectives and Classroom Use

The work above makes clear that unequal probability of selection is an important aspect of sampling and regression that is commonly encountered in practice. This section shows how these concepts can be incorporated into a course. Importantly, students will learn that the desirable properties of the OLS estimator requires SRS and that applying OLS to an unequal probability DGP produces biased and inconsistent estimates. They will also learn that using probability weighted least squares produces better estimates. This alone is sufficient reason for incorporation in an undergraduate econometrics course, but below we offer other rationales and suggestions for course enrichment.

A fundamental pedagogical idea is that of extending a model by relaxing previously fixed assumptions. Students see what they have learned thus far as a special case of a more general model. This provides the opportunity to reinforce previous lessons and enables deeper understanding of the material. In most undergraduate econometrics courses, for example, after presenting the Classical Linear Model, heteroscedasticity is introduced as a more complicated extension of the error term. Students learn that homoscedastic errors are a special case of a more general model. In similar fashion, simple random sampling is a special case of the more general case of sampling with unequal probabilities. Incorporating unequal sampling allows for review of sampling and regression, guaranteeing increased learning. It also widens the student’s view of the data generation process and how it impacts estimation and inference.

There are other ways in which unequal probability is analogous to heteroscedasticity. The idea that errors were not identically or independently distributed was understood early in the development of inferential regression, but it was computationally overwhelming and, thus ignored in practice. First weighted least squares and, more recently, the robust SE was incorporated in statistical software packages, and econometrics textbooks soon added this material. Similarly, with complex survey algorithms now ubiquitous, it is time to catch up to the software and include this content in our courses.

Given the similarity in using weights to improve performance in the case of errors with varying spreads and unequal probability of selection, it would seem natural to have unequal probability follow heteroscedasticity. Students would again be reminded that regression is part of the family of weighted averages and basic ideas of unbiasedness, consistency, and minimum variance in the sampling distribution could be reinforced.

The exposition in this paper, with its emphasis on simulation and comparing estimators, offers a clear, intuitive delivery. The Excel workbook or Stata do-file could be projected and various simulations could be carried out in class and given as homework. An exploration of the performance of OLS and probWLS under varying sample sizes (as suggested near the end of Section 4) would be an appropriate assignment. Another approach would be to incorporate the material in student papers. In addition to OLS, students could estimate probWLS and compare the results. Even more ambitious would be incorporating the survey design to obtain better estimates of the SE.

Finally, there is a data-driven reason to bring unequal probability into the undergraduate curriculum: with easily-accessible, real-world surveys, such as the Current Population Survey, using unequal probability of sampling and user-friendly software enabling correct estimation, the return to incorporating unequal probability can be quite high. Point estimates are sometimes relatively unchanged, but other times the effect can be substantial. For example, using the CPS, Carrington, et al. (2000, p. 4) see little difference between OLS and probWLS coefficients on *Education*, *Age*, and *Female* variables, but slopes on *Black* and *Hispanic* dummies are substantially different.

8. Conclusion

This paper argues that unequal probability of selection and complex survey design should be incorporated in the standard econometrics curriculum. This material is well understood by statisticians and is part of modern software, such as Stata, SAS, and R. Our focus was on intuitive, visual explanations of the consequences of using OLS with unequal probabilities of selection and the way probability weighted least squares can be used to provide better estimates. Further reading, along with standard matrix exposition, is available from Deaton (1997), Lohr (2010) and Wolter (2003); Lumley (2010) offers a complex survey package in R.

We chose Excel as our primary platform for explaining this material to enable easy use for classroom presentation. Two user-defined array functions, SAMPLE and LINESTW, can be used to create additional examples or apply probWLS to data in Excel. We also showed how Stata can be used to demonstrate the effect of unequal probability sampling. Although the Excel workbook and Stata do-file come with our hypothetical finite population, users can use the Excel SAMPLE and LINESTW and Stata do-file on their own datasets and to construct their own examples.

We constructed a simple example of an earnings function and used it to illustrate how a finite population is produced. Then we applied three different sampling schemes. To recap the key points, in our example, conditioning unequal probabilities of selection on the independent variable, *Schooling*, did not affect the expected value of the OLS estimator, though precision was diminished. When unequal probability of inclusion depended on the *Y* variable, *Wage*, OLS broke down, producing obviously biased and inconsistent slope estimates. Probability weighted least squares performed much better because it gives greater weight to observations that are less likely to be chosen. In our example, the estimated SE used by probWLS, based on a Taylor expansion approximation, was, on average, about 80% of the approximate SE (via simulation), which is our proxy for the exact SE. This helps explain why there are several alternatives to the standard sandwich estimator, including a variety of bootstraps.

The results in this paper are dependent on our particular DGP and data. It is not true, for example, that OLS performs reasonably well if unequal probability sampling is a function of an independent variable. We constructed our example to produce strong visuals and enable intuitive explanations.

It is worth remembering that this paper focused exclusively on the effect of unequal probability of selection. While important, there is a second fundamental aspect that must be considered: the stratification and clustering design effect of a complex survey. Point estimates are unaffected by design considerations, but SEs depend critically on correlations within clusters. If these correlations are not taken into account, estimated SEs are biased low. On the other hand, stratification is an honest way of reducing SEs because we are using the additional information that different observations within the same stratum are more similar to each other than they are across different strata. As a result, when we use a different average for each stratum to calculate the standard error for that stratum, we obtain higher precision in our estimates. For example, if we know beforehand that apples and oranges are different types of fruits and are likely to be different in their weight, on average, then we can calculate the standard deviation of their weights separately using a different average for each fruit. This will increase precision. We strongly believe that, just like unequal probability of selection, explaining the implications of complex survey design with a clear, intuitive exposition also needs to be included in the econometrics curriculum and this will be the subject of future work.

References

Barreto, H. and Howland, F. M. (2010, 2nd printing). *Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel* (Cambridge University Press). *www.wabash.edu/econometrics*

Carrington, W. J., J. L. Eltinge, K. McCue (2000), "An Economist's Primer on Survey Samples," Working Papers from U.S. Census Bureau, Center for Economic Studies. *EconPapers.repec.org/RePEc:cen:wpaper:00-15*

Deaton, A. (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy* (World Bank). *books.google.com/books?&id=Mfe8ukMh_v4C*

Lohr, S. L. (2010, 2nd ed.), *Sampling: Design and Analysis* (Cengage Learning).

Lumley, T. S. (2010). *Complex Surveys: A Guide to Analysis Using R* (John Wiley & Sons, Inc.). *faculty.washington.edu/tlumley/survey*

Stata Corporation (2009). *Survey Data Reference Manual*, (Stata Press). *www.stata.com*

Wolter, K. M. (2003). *Introduction to Variance Estimation* (Springer Verlag).