6

Regression With Survey Data From Complex Samples

Secondary analysis of data from large national surveys figures prominently in social science and public health research, and these surveys use complex sample designs in lieu of the simple random sample (SRS) that is assumed by most conventional statistical software. Examples are the National Health Interview Survey, the Medical Expenditure Panel Survey, the National Health and Nutrition Examination Survey, the Collaborative Psychiatric Epidemiology Surveys, the National Longitudinal Survey of Adolescents Health (Add Health), and the General Social Survey.

Researchers are drawn to these data sets because of their methodological strengths, such as the use of probability sample designs that can be generalized to the population, large sample sizes that permit sophisticated statistical analysis, and oversamples of relatively small groups in the population. The analysis of publicly available data sets is inexpensive for the analyst; and the agencies that funded the original data collection often are eager to have these data analyzed, which can result in funding for secondary data analysis. In addition to these public use data sets, survey researchers who conduct primary data collection rely on complex samples more often than not.

These complex samples differ from SRSs in such fundamental ways that it is not appropriate to analyze the data as if these data were obtained from an SRS. These design features include **stratification**, which is partitioning the entire population into nonoverlapping segments (e.g., census tracts); **clustering**, which are groupings of similar persons (e.g., blocks within census tracts); and **unequal selection probabilities**, which means that the probability of being selected into the sample differs across the persons who comprise the population. As a result of stratification and clustering, individuals are not sampled independently of one another. Two issues are of paramount concern: (1) the representativeness of the sample and its impact on parameter estimates and (2) the estimation of population variances and standard errors (*SE*s), which form the basis for tests of statistical significance and for constructing confidence intervals (CIs).

In sharp contrast, in an SRS, every individual in the population has an *equal* chance of being randomly selected for the sample, and the selection of individuals is *independent* of one another. However, this type of sampling is extremely rare in surveys of the general population for several reasons.

First and foremost are cost and logistical difficulty, especially for interviews with a sample that is widely dispersed geographically. For a sample defined geographically, for instance, an SRS would result in people being sampled from anywhere within these boundaries. To drive this point home, Los Angeles County encompasses 4,060.87 square miles (1 square mile = 2.59 square kilometers); given local traffic conditions, interviewers might well log more hours traveling to and from interviews than actually conducting them. In contrast, it is more efficient and less costly to collect data when sampling units are grouped together in some way. For example, the Los Angeles Depression Study, conducted in 1979 and recently designated a Social Science Classic by the Library of Science (Aneshensel, 2009), drew a representative sample of adults residing in Los Angeles County by first sampling census tracts from within subsets of all tracts in the county (known as strata, see below), then blocks within sampled tracts, households (HH) within sampled blocks, and finally one individual within sampled HHs (Frerichs, Aneshensel, & Clark, 1981).

An equally important consideration is that there may not be an information source for randomly selecting individuals directly, as is done in an SRS. In the absence of a listing of all residents of Los Angeles County, for example, sample members cannot be selected randomly as individuals. For this reason, it is necessary to sample identifiable clusters of individuals and then sample individuals within those clusters. These clusters may have existing lists of the sampling units within them, for instance, sampled schools have lists of all students in the schools. Other times, it is necessary to list individuals expressly for the study to provide the information necessary to select individuals.

For the Los Angeles Depression Study, two listings were necessary. First, the HHs on sampled blocks were enumerated by canvassing each block and listing the address of every HH on the block so that a sample of HHs could be selected. Then residents of the sampled HHs were listed so that one adult could be randomly selected. Until this last stage in the sample design, clusters were sampled; individuals were selected for the sample only in the sense that their clusters were sampled, and only some of the people in selected clusters were ultimately selected for the sample. Specific individuals were selected only in the final stage.

Finally, complex samples are also preferred when it is desirable to obtain larger numbers of sample members with particular characteristics than would be obtained with an SRS. This is accomplished by oversampling people with those characteristics. In social science research, it is now commonplace to oversample members of racial/ethnic minority groups to improve the precision of parameter estimates for these groups and to provide sufficient sample sizes for within group analysis. For the same reasons, people of advanced old age, the "oldest old," are sometimes oversampled.

Analyzing data obtained from a complex sample as if it had been collected from an SRS may yield biased parameter estimates and tends to underestimate *SE*s, which leads to inflated tests of statistical significance and increases the chances of making a Type I error, failing to

reject the null hypothesis when it is true, for example, finding an association when in fact the variables are independent of one another. For these reasons, the features of a complex sample design should be taken into consideration during data analysis by using specialized software to obtain unbiased parameter estimates that are representative of the population and robust *SE*s that accurately reflect the variability due to the sample design.[1, 2]

Although the statistical development of the techniques for analyzing data from complex samples is beyond the scope of this text, implementing these methods is usually straightforward (unless, of course, you are the statistician who is compiling the sample weights, for example, or the programmer who is developing the software). In most instances, all that is involved is (a) reading the study documentation pertaining to the derivation of the sample to identify the variables that define the sample, (b) inserting these variables into the specialized software as indicated in the software documentation, (c) selecting among a few analytic options based on both sources of documentation, (d) executing the analysis using specialized software, and (e) interpreting results in the same manner as results based on the analysis of an SRS using ordinary statistical techniques.

Although these steps are sufficient to get the job done, this "black box" approach leaves one in the dark about how the derivation of the sample and adjustments for its design have influenced analytic results and, thereby, study findings and conclusions. These adjustments to the data are just as important as the methods of data collection for arriving at accurate conclusions. The overview of these topics in this chapter is intended to shed some light on the internal mechanisms of this black box and the implications for drawing valid inferences from the data. For a complete treatment of these topics, the reader should consult Heeringa, West, and Berglund (2010), an excellent text on the statistical basis for the analysis of survey data from complex samples.

This chapter first describes the basic elements of complex samples: stratification, clustering, and unequal selection probabilities. The second part explains the impact of these sample design elements on parameter estimates and their *SE*s. Next, techniques to adjust for these effects to obtain unbiased parameter estimates and robust *SE*s are described, procedures that are available in major statistical software packages. Specifically, methods of estimation for multiple linear regression that take the sample design into account are presented, including sample weights and adjustments to *SE*s. Implications for inferences from sample estimates to true population values are then discussed. The Health and Retirement Study (HRS) provides an example of a complex sample and is used to illustrate the application of design-based methods of analysis, and to demonstrate inferential errors that may arise from the incorrect analysis of these types of data as if they were obtained from an SRS.

Complex Samples

Complex samples can be understood best by comparing their structure to an SRS, and the multistage area probability sample is a particularly instructive example because it incorporates features found in other less complicated designs that utilize only some of these features. Also,

it is the design most often used in the large-scale population-based surveys that are used for secondary data analysis by social scientists and public health researchers.

A multistage area probability sample has the following characteristics: (a) sampling occurs at more than one stage, (b) the boundaries of the sample are delineated by geographical and demographic characteristics, such as U.S. Metropolitan (core urban population of at least 50,000) Statistical Areas and nonurban counties, and (c) the elements of the sample are selected using a random mechanism so that every unit has a known chance of being selected for the sample (hence its name). This type of design is used to select not only individuals but other units as well, such as couples, families, or HHs, but for simplicity, the presentation that follows assumes that individuals are being sampled.

The multiple stages are arranged in a hierarchy in which the sampling units of the later stages are nested within the sampling units of the earlier stages (e.g., blocks are contained within census tracts, HH are situated within blocks). Clusters are used in place of the direct sampling of individuals from within strata because they can be identified, whereas it may not be possible to identity individuals at this stage, for example, blocks within census tracts are known, but the residents of a census tract are not listed (except in the decennial census, and these primary data are not available to researchers). Equally important, cluster sampling substantially reduces data collection costs and facilitates the administration of the data collection because the individuals within a cluster who are selected into the sample are less widely dispersed geographically than if they had been selected independently of one another.

In the initial stages, clusters of individuals with shared characteristics are sampled; individuals are selected only in the sense that they are part of a cluster that is selected. In this manner, individuals are carried along in the derivation of the sample as long as their cluster is among the clusters that are selected at that stage. When a cluster is not selected, it is dropped from the sampling frame along with all the individuals that constitute it. Specific individuals are not selected until the final stage, and even then, they are sampled from among the individuals contained within the cluster selected at the preceding stage (e.g., an individual is selected from within an HH).

In this manner, every member of the sample belongs to clusters that have been selected at every stage in the sampling frame. Each person's probability of being selected into the sample accumulates across the stages of the sample design. Because the probability of selection is known for the clusters, along with the probability of selection for the individual in the last stage, the probability of selection for each individual is known. However, selection probabilities for clusters may differ; for example, some clusters may be oversampled, which filters down to generate individual differences in the probability of being selected into the sample. Also, the selection probabilities of individuals in the last stage tends to differ as well, for instance, someone living in a two-person HH has a greater selection probability than someone living in a three-person HH when only one person per HH is selected. For these reasons, the members of a multistage area probability sample have a known probability of selection, but these selection probabilities are unequal.

Sampling techniques are not necessarily the same across stages. For example, probabilities proportionate to size might be used to select census tracts; a random draw could be used to select two blocks per census tract; a systematic sample could be used to select every *k*th HH

within a sampled block starting with a random HH; and, within an HH, simple random selection could be used by selecting the person with the most recent birthday.

With this general description, we turn now to the most critical features of complex samples: stratification, clustering, and unequal selection probabilities. Their impact on parameter estimates and *SE*s is described along with the consequences of ignoring the sample design by analyzing data as if the data had been generated by an SRS.

Stratification

Multistage area probability sample designs typically employ stratification, which entails subdividing the entire population into mutually exclusive and exhaustive subpopulations, which are known as strata (Kish, 1965). Stratification is used for several reasons, including decreasing the size of *SE*s relative to an SRS of the same size, enabling the oversampling of specific subpopulations, facilitating the use of different survey methods within strata, and permitting analysis within strata (Heeringa et al., 2010). Stratification affects *SE*s because the variance is estimated within stratum, which are internally homogeneous, and then averaged across stratum. It improves the precision of estimates by making the variance of at least some variables smaller within the strata than within the sample as a whole. This is achieved by stratifying on characteristics that are closely associated with selected variables so that there is less variability within stratum than between them. For example, socioeconomic status (SES) is strongly related to many health outcomes, which accounts for the use of stratification by SES in health research. Given that the characteristics used to define strata are more strongly associated with some variables than with others, stratification will affect the variance estimates of these variables more than the others.

Strata are distinct from one another and encompass the entire population. They are often delineated by geographic boundaries and demographic characteristics. However, strata may be defined by other characteristics, such as school districts. There are usually only a few strata, with each stratum containing a large number of people. To increase the precision of parameter estimates, there should be as much homogeneity within strata as possible on the variables under investigation, and the greatest amount of heterogeneity between strata (Heeringa et al., 2010).

In the first stage, the eventual sampling units (e.g., individuals) are grouped together into the largest clusters, which are known as **Primary Sampling Units** (PSUs; e.g., census tracts). PSUs are sampled from every stratum, and sampling is independent across strata. As a result, stratum can be analyzed separately. Although PSUs are sampled from each stratum, not all PSUs within a stratum are selected for inclusion in the sample. Only the sampled PSUs are carried forward to the next stage in the sample design.

It is important to emphasize that at this stage, PSUs are sampled, not individuals. Individuals are selected only insofar as they are members of a PSU that has been selected; most individuals within a sampled PSU eventually will be dropped as a result of the sampling that occurs at subsequent stages.

In contrast, with SRSs, sampling occurs in one stage and individuals are selected directly. Stratified samples usually result in somewhat smaller variance estimates than an SRS of the

same size, that is, smaller estimates of *SE*s. Therefore, *SE*s are overestimated when stratification is not taken into consideration during analysis by using software that assumes an SRS; tests of statistical significance are deflated and CIs are artificially wide, which may lead to a Type II error.

Clustering

In the next stage of a multistage area probability sample design, the clusters are known as **Secondary Sampling Units** (SSUs), and SSUs are independently sampled from within each of the PSUs that were selected at the previous stage. Thus, clusters are linked between stages: The PSUs that are selected at the first stage contain within them the SSUs for the second stage. If, for example, the PSUs are census tracts, then the subset of selected census tracts provides the clusters of blocks that are sampled at the second stage. As before, it is the SSUs that are sampled at this stage (not individuals), and only some SSUs are selected from within each PSU. All individuals within a given cluster are selected at this stage and remain in the sampling frame at this stage, although only some of these individuals will be selected into the final sample at subsequent stages. Sampling of SSUs from within PSUs is independent of sampling of SSUs within other PSUs.

The sampled SSUs contain within them the clusters to be used at the next stage and so on until the final stage when individuals are selected. For example, each block that has been sampled from a census tract at Stage 2 encompasses numerous HHs, from which a sample of HHs is then selected at Stage 3.

People within clusters tend to have similar characteristics precisely because they are in the same cluster. For example, they reside on the same block. As a result, they are more like one another than like people in other clusters and more alike than they would be if they had been selected individually from an SRS. For example, sampled residents from HHs on a given block are likely to be somewhat the same on many characteristics, such as annual family income. This violates the assumption of an SRS that observations are independent of one another. The extent to which observations within a cluster resemble one another is known as the **intraclass correlation coefficient**. Although a cluster needs to be large enough to make stable parameter estimates, large clusters are not an optimal use of resources because the observations are providing somewhat redundant information as a result of this similarity.

The homogeneity within clusters results in an effective sample size for a clustered sample that is smaller than an SRS of the same size. Therefore, the actual sample size may need to be increased to achieve the desired level of precision, offsetting some of the cost-effectiveness of clustering.

Although stratification tends to decrease the variance of parameter estimates relative to an SRS of the same size, as just discussed, clustering has the opposite effect. These countervailing influences may be perplexing because strata and clusters are similar in that they are internally homogeneous with regard to the variables being studied. However, while all stratums are sampled, only some clusters are sampled. As a result, the clusters that are sampled may contain cases that are more dissimilar from one another than would be the case if all clusters were sampled, leading to greater variance (Menard, 2010).

The final stage of a multistage area probability sample entails the random selection of individuals into the sample. Unlike an SRS, however, the selection of individuals into the sample is not independent because the individuals who are selected in the final stage have been selected because they were part of clusters that were sampled in earlier stages.

Clustering is an exceedingly consequential feature of complex samples for the estimation of *SE*s, which then affects significance tests for hypotheses and inferences to the population made on the basis of these tests. Cluster sampling typically increases sampling variation compared with the direct sampling of individuals in an SRS, resulting in less efficient *SE*s. In other words, parameter estimates are less precise and *SE*s are larger in a clustered sample than in an SRS of the same size.

The impact of stratification on reducing variance estimates typically is substantially smaller than the impact of clustering on increasing variance estimates. For this reason, the net effect is to increase the variance of parameter estimates when both stratification and clustering are used.

As a result, when SRS statistical techniques are mistakenly applied to a clustered sample, *SE*s are underestimated and tests of statistical significance are inflated, leading to an increased chance of making a Type I error. In this manner, one may obtain a statistically significant result, when in actuality, there is none. These considerations do not apply to data from an SRS because individuals are selected directly, not as elements in clusters and strata.

Unequal Selection Probabilities

A representative sample of the population is essential to making accurate estimates of population parameters, but sample characteristics may not align well with population characteristics for a number of reasons. From a sample design perspective, a major consideration is variation in the probability of being selected into the sample. In an SRS design, each person has an equal probability of being selected. In a complex sample, however, members of the population usually (although not always) do not have equal selection probabilities by design. Differences typically occur at every stage in a multistage sample and accrue over the stages. As a result, the raw sample may not be an accurate representation of the population, necessitating the use of sample weights (see the following discussion).

Selection probabilities are likely to vary from person to person as a result of the design because strata differ in size from one another as do clusters. For example, if the sampling of census tracts from within a stratum is based on probabilities proportional to the size of the tract, then a large tract will have a greater probability of selection than a small tract. In contrast, if two blocks are sampled per census tract, then the selection probability will be higher for blocks within a census tract containing only a few blocks than blocks within a tract with many blocks.

The probability of selection into the sample accumulates over the stages of a sample design in that the sampling fractions across stages are multiplied to yield a total sampling fraction. Suppose that for a given individual, the probability of being selected into a sample like the one used in the Los Angeles Depression Study is 1/18 for the tract (sampling 100 of the approximately

1,800 tracts in the county), 1/25 for the block (sampling 2 of the 50 blocks in the tract), 1/10 for the HH (sampling 3 of the 30 HH on the block), and 1/2 for the individual (sampling 1 of 2 persons in the HH). This person would have a sampling fraction of $1/18 \times 1/25 \times 1/10 \times 1/2 = 0.00011$. Someone residing on the same block and who lives alone, in comparison, would have a sampling fraction of 0.00022 ($1/18 \times 1/25 \times 1/10 \times 1$).

In addition, it is common for some subgroups of the population to be oversampled by having a disproportionately large selection probability compared with members of groups that are not oversampled. For example, it is often desirable to oversample members of minority groups such as African Americans to ensure sufficient sample sizes for analysis and to improve the precision of parameter estimates for these groups. One method of accomplishing this end is by selecting a disproportionately large number of census tracts with relatively high concentrations of African Americans; this is done by assigning these tracts relatively high selection probabilities. Given that African Americans, on average, have lower SES than do non-Hispanic Whites, especially African Americans living in impoverished, hypersegregated inner-city neighborhoods, the resultant sample may not accurately reflect the full spectrum of variation in SES among African Americans. To compensate, PSUs could be stratified simultaneously on both racial composition and SES, for example, increasing the selection probabilities of high-SES predominantly African American tracts even more. As a case in point, the Add Health Survey mentioned in Chapter 2 oversampled African American students with college-educated parents.

Oversampling most often is based on sociodemographic characteristics, such as membership in a racial/ethnic minority group, although other criteria are sometimes used. For example, in the Add Health Study cited in Chapter 2, twins and adoptees were selected with certainty.

It should be apparent that if some groups are oversampled in this manner, other groups necessarily are undersampled (given a fixed sample size), although this consequence is rarely made explicit in sample descriptions. These groups compose a smaller proportion of the total sample than would be the case if not for the oversampling of other groups.

As a result, characteristics that are associated with the criteria used to oversample have sample distributions that typically do not correspond very well with the distributions of those characteristics in the population. Indeed, these sample characteristics may be extremely distorted. For instance, if persons over the age of 85 are oversampled, then sample estimates of population parameters that are associated with age will be inaccurate. Suppose that subjective life expectancy is estimated—the age to which you expect to live minus your current age. The disproportionately large number of people nearing the end of their lives will yield an overall estimate that is shorter than the true value for the population where this group is a relatively small (albeit growing) portion of the population.

More generally, unequal selection probabilities generate a sample in which the proportional representation of at least some segments of the population is not the same as their proportional representation in the population. In other words, the sample is not representative of the target population because some groups are overrepresented, while others are underrepresented. Consequently, estimates of population parameters based on the raw sample tend to be biased, and findings do not generalize accurately to the population.

In contrast, in the SRS everyone has an equal probability of selection so that the distribution of characteristics of the sample mirrors those of the population. Therefore, parameter estimates can be generalized to the population.

HRS ANALYSIS JOURNAL 6.1

The Health and Retirement Study Sample Design

The Health and Retirement Study is an ongoing, biennial longitudinal study of a large, nationally representative multicohort sample of persons aged 50 and older begun in 1992 (Health and Retirement Study, n.d.). In this chapter, it is used to illustrate the design of complex samples and the analysis of survey data collected from these types of samples. The HRS data also provide a running example of the implementation of the elaboration model with multiple linear regression that spans the beginning of the exclusionary strategy of analysis through the end of the inclusive strategy (Chapters 7 through 11).

The HRS sample design is a good illustration of the type of complex sampling design just discussed. The target population for each cohort was all adults in the contiguous United States born during the birth cohort years who resided in HHs. Samples were selected using a multistage area probability sample design with four selection stages. At the first stage, 2 PSUs were selected from among each of 56 strata, where the PSUs were U.S. Metropolitan Statistical Areas (MSA) and non-MSA counties selected based on probabilities proportionate to size. The SSU were area segments within PSUs. Third, a complete enumeration was made of all housing units (HU) physically located within the selected SSU, followed by the random selection of HUs that contained at least one person from the birth cohort. The final stage selected one or more persons within a sampled HU: (a) a single unmarried age-eligible person, (b) a married couple who were both age-eligible, or (c) a married couple in which only one spouse was age-eligible. If there was more than one unrelated age-eligible person in the HU, one person was randomly selected. In addition, there were three oversamples: Blacks, Hispanics, and residents of Florida.

The original HRS study that commenced in 1992 was designed to follow a cohort of adults then in their fifties as they made the transition from active work into retirement (HRS1, born 1931–1941, $n = 12{,}654$). It was joined in 1993 by the companion Assets and Health Dynamics of the Oldest Old Study (AHEAD, born before 1924, $n = 8{,}222$), consisting of persons aged 70 and over and designed to examine the postretirement and end of life period. The two studies were merged in 1998, when two new cohorts were added: Children of the Depression Era (CODA, born 1924–1930, $n = 2{,}320$) and War Babies (WB, born 1942–1947, $n = 2{,}529$). In 2004, an additional cohort was added, Early Baby Boomers (EBB, born 1948–1953, $n = 3{,}340$).[3] Thus, the sample encompasses people in late middle age through the oldest old.

The substantial differences in the number of participants enrolled for each cohort signify differences in selection probabilities across cohorts that are consequential to sample estimates of characteristics of the population. For instance, the 10-year HRS1

cohort enrolled nearly four times as many participants as the 5-year EBB cohort. This difference suggests that persons in the birth years for HRS1 were in effect oversampled, having a higher probability of being selected into the sample than persons in the EBB birth years.

The analyses reported in this text utilize the sample as it was constituted in 2006 and 2008 in order to use data from an enhanced face-to-face interview that included a leave-behind "Psychosocial Questionnaire" (PQ) that measures key constructs for the theories used as examples in this text. Half of the sample was randomly selected to complete the PQ in 2006; the other half of the sample was assessed in the same way at the next data collection in 2008. The 2006 sample was divided to select respondents who were sampled for the PQ that year: the same procedure was used to select respondents sampled for the PQ in 2008. Then the entire sample was reconstituted by combining these two halves to increase the analytic sample size and statistical power relative to using only the 2006 or the 2008 data.

Procedurally, the PQ was left with sampled respondents at the end of the interview, and respondents were asked to complete the questionnaire and mail it back to the field office. Telephone follow-ups were conducted with respondents who had not returned the questionnaire after a second reminder notice. Measures from the PQ include life satisfaction, loneliness, discrimination, and social support. Additional information, including sociodemographic characteristics and health status, comes from the concurrent HRS core interview and from the respondent's baseline interview.

Of the approximately 30,000 persons ever enrolled in the HRS sample through 2008, a total of 15,176 were eligible for the PQ, selected for it, and additionally were age-eligible for this analysis because they were in one of the five birth cohorts (excluded, e.g., are younger spouses of participants). Of these persons, 12,983 (85.5%) returned the survey. Respondents were dropped from the analytic sample if they did not have a valid sample weight ($n = 127$) and/or had excessive missing data (following limited imputation of missing data with the mean or the mode, $n = 620$). The final analytic sample for all analyses is 12,236.

Accounting for the Sample Design

The design features of paramount importance to the analysis of data from a complex sample concern the structure of the sample in terms of stratification, clustering, and unequal selection probabilities. Although our primary interest is the impact of these design features in multiple linear regression and logistic regression (see Chapter 12), the procedures described in this chapter apply to other procedures as well, including univariate parameter estimates and their *SE*s (see HRS Analysis Journal 6.2 below). In brief, it is necessary to use sample weights to obtain unbiased parameter estimates and to take the sample design into account in the estimation of *SE*s for tests of statistical significance and CIs. For the statistical basis for these estimates, see Heeringa and colleagues (2010) and the sources cited therein. Major statistical

software packages now contain specialized procedures for analysis of survey data that accommodate sample weights and adjust estimates of *SE*s for complex sample designs. The ease with which these procedures can now be implemented eliminates a number of obstacles that in the past deterred survey analysts from taking the sample design into consideration. Thus, current accepted practice is to analyze complex survey data with procedures that are tailor made for such designs.

Sample Weights

As discussed above, the complex sample design typically results in unequal selection probabilities. This feature distinguishes these designs from SRSs in which each individual in the population has the same probability of being selected into the sample. This variation in selection probabilities needs to be taken into consideration during analysis in order to make unbiased parameter estimates.

This adjustment is accomplished with the use of **sample weights**, also referred to as **probability weights** or **pweights**, which are weights assigned to each observation that equal the inverse of the probability of being selected into the sample based on the sample design. This weight is interpreted as the number of persons in the population who are represented by a particular person in the sample. In an SRS, each person in the population has the same probability of being selected and, therefore, each observation has the same implicit weight of 1.00. In most complex samples, however, selection probabilities vary from person to person, which necessitates the use of sample weights.

Since sample weights are the inverse of selection probabilities, they too accumulate over stages in a multiplicative manner. Continuing the example from above, the sample weight for the hypothetical person in the Los Angeles Depression study is $18 \times 25 \times 10 \times 2 = 9{,}000$; this person's neighbor who lives alone has a weight of 4,500 ($18 \times 25 \times 10 \times 1$). This weight is interpreted as the number of people in the population that this person represents. For this reason, the sum of these weights equals the size of the target population. Statistical packages sometimes apply the weight in a particular data set in such a way that the population size is mistaken for the sample size. Should this difficulty occur, a relative weight or normalized weight can be calculated and used instead by dividing the raw weight by its mean (Thomas & Heck, 2001). The same adjustment can be made by multiplying the sample weight by the unweighted n divided by the weighted n.

Weights for complex samples often take into consideration factors in addition to unequal sampling probabilities. One crucial factor is nonresponse among persons who are selected for the sample, especially differential nonresponse rates across subgroups of the population. For example, some racial/ethnic minority groups may have disproportionately low response rates because they are distrustful of research because of incidents of scientific misconduct, such as the infamous Tuskegee syphilis experiment, or because cultural norms discourage the disclosure of personal information. Sample weights may be adjusted to compensate for these differences, assigning disproportionately high weights to participants who are members of groups with low participation rates.

Adjustment for nonresponse requires information about the characteristics of the persons who did not respond, but this information may be very limited precisely because these persons did not respond. In some instances, information is available from an external source, such as information about all patients in a sample drawn from a medical care provider. In other instances, it is obtained during data collection, for example, through interviewer observations of the person or his or her surroundings. However, this information more often than not is limited to the information contained in the sample design (Heeringa et al., 2010), for example, the racial composition of the census tract.

These adjustments to the sample weights rest on the assumption that people who did participate in the survey are similar to those who did not participate on some characteristics—such as race/ethnicity or median HH income of their census tract—also are similar on other characteristics, most important, the variables being studied. This assumption is problematic given that they differ on one essential characteristic: whether they participated or not. Consequently, adjustments for nonresponse can be thought of as reducing bias in parameter estimates as distinct from providing completely unbiased estimates.

Beyond adjusting for unequal selection probabilities and differential nonresponse, weights are sometimes used to make the sample distribution of key sociodemographic characteristics conform to the population distribution of those characteristics.[4] These weights are called **poststratification weights** because they are computed *after* the sample is collected and because these characteristics define the various *strata* that constitute society, for example, race/ethnicity, gender, and SES. To calculate poststratification weights, information about the stratification of the population by these sociodemographic characteristics is needed from an external source, such as the U.S. Census or the Current Population Survey. Sampling weights are then adjusted to align the proportional distribution of subgroups in the sample to match those in the population, considering multiple characteristics simultaneously. For example, adjustments might be made for strata simultaneously defined by race/ethnicity, gender, and SES such that a low-income African American woman receives a different adjustment from a low-income African American man and from a low-income non-Hispanic White female.

As a result of these factors, sample weights can vary tremendously from person to person.

It is imperative to examine closely the documentation for sample weights for a given data set. Although some of the statistical details may be impenetrable, it is necessary to understand how these weights correspond to the sample design in order to be aware of how the application of these weights affects your analysis. Often, it is productive to examine the mean sample weights by select characteristics; for example, examining how these weights vary across the sociodemographic subgroups of the sample (see Table 6.2).

In addition, sample weights almost always increase estimates of *SE*s and, therefore, are consequential to inferences from sample estimates to population parameters, as discussed in the following.

These issues do not apply to data from an SRS because individuals have equal selection probabilities; each person in effect has an implicit weight of one. This feature eliminates the need for sample weights (although other weights might be called for, such as adjustments for differential nonresponse).

HRS ANALYSIS JOURNAL 6.2

The HRS Sample and Sample Weights

The unequal selection probabilities generated by the HRS sample design require sampling weights to yield unbiased parameter estimates. The HRS weights also adjust for initial nonresponse and for attrition over time. In addition, poststratification adjustments to the weights are made based on the corresponding Current Population Survey on the basis of the birth cohort, gender, and race/ethnicity.

The weights specially for the PQ data also adjust for nonresponse to it. The half samples that received the PQ in either 2006 or 2008 are each weighted to the population for that year. Consequently, an adjusted weight is calculated by dividing these weights in half. The application of the adjusted weight means that we can think of the sample as representing the U.S. population of persons born before 1954 midway between these two times in 2007.[5]

The characteristics of the analytic sample are summarized in Table 6.1. Unweighted and weighted data are presented to illustrate the effects of using sample weights on

Table 6.1 Characteristics of the HRS PQ Sample, U.S. Adults Aged 52 and Older

	Unweighted		*Weighted*	
Characteristic	n	*Proportion/ Mean*	*Proportion/ Mean*	*Robust* SE
Cohort				
AHEAD[a]	1,100	0.090	0.072	0.003
CODA[b]	1,844	0.151	0.114	0.005
HRS1[c]	5,087	0.416	0.266	0.005
WB[d]	2,004	0.164	0.231	0.006
EBB[e]	2,201	0.180	0.317	0.009
Gender				
Male	5,110	0.418	0.457	0.005
Female	7,126	0.582	0.543	0.005
Race/ethnicity				
non-Hispanic White	9,685	0.792	0.833	0.010
African American	1,454	0.119	0.082	0.005
Latino	859	0.072	0.063	0.009
"Other"	238	0.019	0.022	0.003

(Continued)

Table 6.1 (Continued)

	Unweighted		*Weighted*	
Characteristic	n	*Proportion/ Mean*	*Proportion/ Mean*	*Robust* SE
Marital status				
Married	8,215	0.671	0.674	0.006
Divorced/separated	1,310	0.107	0.130	0.004
Widowed	2,350	0.192	0.158	0.003
Never married	361	0.030	0.038	0.002
Employment status				
Employed	2,988	0.244	0.358	0.007
Retired	8,055	0.658	0.543	0.008
"Other"	1,193	0.097	0.099	0.004
Age (years)		69.102	65.952	0.182
Education (years)		12.712	13.003	0.071
Income (/$1,000)		62.619	71.920	1.777
Wealth (/$1,000)		427.692	431.440	14.903

Note: N = 12,236; *SE* = Standard error. Robust *SE*s are calculated using Balanced Repeated Replication (BRR) to adjust for the complex sample design. *SE*s are not shown for unweighted data because they are inaccurate; HRS = Health and Retirement Study; PQ = Psychosocial Questionnaire; AHEAD = Assets and Health Dynamics of the Oldest Old; CODA = Children of the Depression Era; HRS1 = Health and Retirement Study 1; WB = War Babies; EBB = Early Baby Boomers; Some proportions to not sum to 1.00 due to rounding error.

a. Born before 1924.

b. Born between 1924 and 1930.

c. Born between 1931 and 1941.

d. Born between 1942 and 1947.

e. Born between 1948 and 1953.

univariate parameter estimates. The most instructive comparison concerns the distribution of birth cohorts. In terms of sample composition, HRS1 retains the distinction of being the largest cohort, as shown by the unweighted *n* and proportion. However, when the data are weighted, the most recent EBB cohort is the largest one, even though fewer EBB participants entered the study and despite the fact that it encompasses only a 5-year birth interval (compared with the 10-year interval of HRS1).

The differences between the unweighted and weighted distributions are the result of corresponding differences in the sample weights by cohort, as shown in Table 6.2. PQ participants in the EBB cohort have an average weight that is almost three times the average sample weight of participants from the HRS1 cohort. Although these weights

Table 6.2 HRS PQ Sample Weights by Cohort

Cohort	*Mean*	SD
AHEAD	4562.55	2020.06
CODA	4320.57	1738.71
HRS1	3650.08	1581.80
WB	8073.38	4173.52
EBB	10069.77	4010.53
Total	*5712.37*	*3756.80*

Note: *N* = 12,236; HRS = Health and Retirement Study; PQ = Psychosocial Questionnaire; *SD* = standard deviation; AHEAD = Assets and Health Dynamics of the Oldest Old; CODA = Children of the Depression Era; HRS1 = Health and Retirement Study 1; WB = War Babies; EBB = Early Baby Boomers (see note, Table 6.1).

adjust for several factors, one is the sampling rate, that is, the size of the cohort sample relative to the size of the population in that age cohort. The weighted distribution reflects the actual distribution of these cohorts in the U.S. population aged 52 and older in 2007, whereas the unweighted distribution reflects the number of participants in the study.

The standard deviations (*SD*s) in Table 6.2 show that there is substantial variation in these weights. The weights range from a low of 1,593 to a high of 21,655; there are 7,653 unique values among the 12,236 persons in the sample. This variation reflects the numerous factors that are taken into consideration in the construction of these weights. For example, the average weight for African Americans (3,956) is substantially lower than the other racial/ethnic groups because this group was oversampled (non-Hispanic Whites, 6,008; Latinos, 5,106; and "other," 6,588). However, among African Americans, males have a substantially higher average weight (4,698) than do females (3,551). The end result is that each respondent is almost unique with regard to the number of persons represented in the population.

The impact of ignoring unequal selection probabilities on parameter estimates can be seen by considering a characteristic that is related to cohort, such as education, which has risen historically with each successive cohort. The difference between the unweighted and weighted estimates is about a quarter of a year (0.291) of education, a modest difference, but substantially greater than the robust *SE* for education (see Table 6.1).

This impact also can be seen for characteristics that are associated with age because age is fixed by cohort. The estimate of age itself is lower by 3 years in the weighted than in the unweighted data, reflecting the relatively large weights for the two most recent cohorts. Age also can be used to illustrate the impact of the sample design on variance estimates. Its design-adjusted *SE* (see below) is given in Table 6.1 as *SE* = 0.182; the biased estimate under the assumption of an SRS is *SE* = 0.086, approximately half. A difference of this magnitude is extraordinarily consequential for tests of statistical significance and CIs, increasing the chance of making a Type I error.

Two age-related characteristics further reveal the magnitude of the bias that can occur with estimates using SRS statistical techniques with data obtained from a complex sample.

Being widowed and being retired both increase with age, so it is reasonable to suppose that the estimates of both would decrease from the unweighted to the weighted estimates. As shown in Table 6.1, compared with the weighted estimate, the unweighted estimate of the proportion widowed is 21.6% larger ($100 \times (0.192 - 0.158)/0.158$); this value is 21.3% for proportion retired ($100 \times (0.658 - 0.543)/0.543$).[6]

The impact of unequal selection probabilities is not limited to the estimation of characteristics of the sample. A comparison of reports of having lifetime exposure to major acts of discrimination demonstrates the potential impact. These experiences were assessed by asking whether eight events occurred "at any point in your life," such as being unfairly dismissed from a job; unfairly prevented from moving into a neighborhood; and unfairly stopped, searched, questioned, physically threatened, or abused by the police. These events were counted. Although the modal response is 0, the maximum score of 8 is also observed, indicating that exposure in general is low, but some persons have experienced numerous discriminatory acts. The average level of exposure is highest among persons in their fifties and declines steadily through age 80 before leveling off. As a result of this age trend, the unweighted mean is considerably lower ($\overline{X} = 0.467$) than the weighted mean ($\overline{X} = 0.530$, $SE = 0.011$), such that the design-adjusted 95% CI [0.507, 0.553] does not include the unweighted estimate.

These dynamics are important to grasp because they illustrate the potential bias in parameter estimates that can result from not taking the sample design into account during data analysis. For these reasons, the weighted data are better suited to making univariate parameter estimates that can be generalized to the population.

We now turn to the characteristics of the sample compared with the weighted sample estimates of the population. Table 6.1 shows more women than men, which reflects the average age of 69 when the greater longevity of women has become apparent. The gender difference is slightly smaller in the weighted data, consistent with the lower average age. The overwhelming majority of respondents are non-Hispanic Whites, a distribution that like gender in part reflects the age composition of the sample. This preponderance is even more extreme in the weighted data and reflects the fact that the African American and Latino populations are younger than the non-Hispanic White population. About two of three respondents are married in both the unweighted and weighted data, but as just discussed, the proportion widowed in the population estimate is less than in the sample, whereas the reverse is the case for separated and divorced. Two thirds of the sample is retired and a quarter is employed, but these values are substantially different from the population estimates based on the weighted data. The sizeable numbers of widowed and retired persons indicate that many HRS participants already have undergone some of the major transitions in life course trajectories that occur later in life. The sample is well educated for these birth cohorts, with somewhat more than a high school education on average. Average HH income is above the national median HH incomes of persons aged 65 and older ($27,798), but it is less than the value estimated for the population with the weighted data. Although there is substantial wealth, the sample encompasses people who are considerably less well off financially, including those who are in debt.

Estimating the Regression Equation and Standard Errors

Multiple linear regression parameter estimates and their *SE*s are affected by the design of complex samples, which has substantial implications for tests of statistical significance and for the calculation of CIs, and by extension, inferences that are based on these statistics. The same equation is estimated for data obtained with a complex sample as for ordinary least squares (OLS) regression under an assumption of an SRS (hereafter referred to OLS/SRS) (see Equation 5.14),

$$\hat{Y} = a + b_f X_f + b_{i1} X_{i1} + b_{i2} X_{i2} + \cdots + b_{i+} X_{i+}, \tag{6.1}$$

where $\hat{Y}$ is the predicted value of the focal dependent variable, a is the intercept with the y-axis when all of the independent variables equal 0, b_f is the expected average change in Y for a 1-unit increase in the focal independent variable X_f when $X_{i1}, X_{i2}, \ldots, X_{i+}$ are held constant, and the same interpretation applies to $b_{i1}, b_{i2}, \ldots, b_{i+}$.

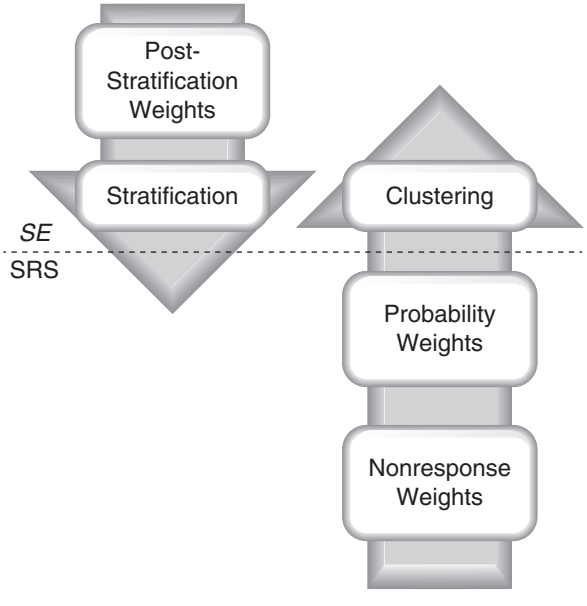
The use of sample weights to obtain unbiased parameter estimates, however, changes the method of estimation. Weighted least squares estimation is used instead of OLS estimation, such that the contribution of each observation to the residual sum of squares is proportional to its population weight (Heeringa et al., 2010).[7]

The cumulative influence of design elements on estimates of *SE*s are summarized in Figure 6.1. Stratification decreases *SE*s somewhat as does the use of poststratification weights. Other weights increase *SE*s as does clustering. The net effect of these opposing influences more often than not is to inflate *SE*s relative to an SRS of the same size of the same population.

These effects are quantified as the **design effect** (DEFF), which is the ratio of the variance under the actual sample design to the variance under a hypothetical SRS of the same size from the same population. Values less than 1 indicate that the actual sample design is more efficient than the hypothetical SRS, whereas the more usual values that are greater than 1 indicate that the actual complex sample design is less efficient. Efficiency refers to how much a statistic fluctuates from sample to sample and is related to the precision of the statistic as an estimate of the parameter. Taking the square root of the DEFF results in the **design factor** (DEFT), which is the ratio of corresponding *SE* estimates; it shows how much the sample design changes the *SE*.

Major software packages provide procedures for correcting *SE*s based on the sample design. For the statistical details of these, the interested reader is referred to Heeringa and colleagues (2010). One approach is linear approximation, also known as Taylor series linearization, which is based on an expansion of the first- and higher-order derivatives of the formula for the population parameter (Menard, 2010). The variance estimate for this approximation is used to determine the variance of the estimate itself (Korn & Graubard, 1999). The estimated variance is calculated from the variation among PSUs; stratum variance estimates are pooled to compute the overall variance estimate (Kneipp & Yarandi, 2002).

Figure 6.1 Impact on Standard Errors (*SE*) of Complex Sample Design Relative to a Simple Random Sample (SRS)



Note: The net effect of a ***complex sample design*** typically is to increase ***SE***s relative to the ***SE***s that would be obtained with an SRS of the same size of the same population. Stratification and poststratification weights tend to decrease ***SE***s somewhat, but these effects usually are more than offset by the increase in ***SE***s that results from clustering and from probability and nonresponse weights.

The second approach entails what is known as resampling in which replicated subsamples are selected from the sample, point estimates are made for each subsample, and then, the overall variance of the statistic is estimated from the variability of the subsamples. The balanced repeated replication (BRR) method draws multiple half-sample replicates for designs with exactly two PSUs per stratum. The jackknife method draws multiple subsamples by deleting a small and different portion of the total sample (e.g., a PSU) from each subsample.

Some public use survey data sets do not contain complete information on the sample design to protect the confidentiality of participants and provide instead a replicate sample weight, which is an adjusted weight, usually with the resampling methods of BRR or the jackknife method.

Thomas and Heck (2001) conclude that using specialized software procedures is by far the most accurate method of accounting for the effects of clustered samples. They note three other less precise corrective steps that can be considered: (1) using a known DEFT value to adjust *SE*s

upward, (2) manipulation of the effective sample size by adjusting the relative weight downward based on a known DEFF value, or (3) simply using a more conservative critical value for alpha (e.g., .01 or .001 instead of .05). Given the marked superiority of specialized software and the ease with which it can be implemented for many statistical procedures, the only reasons to use any of these alternatives would be the absence of information about one or more of the variables that define the sample, which does occur on occasion with the secondary analysis of existing data, or the lack of suitable software for a specific statistical procedure. Ignoring the effects of the sample design is not a viable alternative.

Inferences to the Population

The parameter estimates of Equation 6.1 and their standard errors are used to draw inferences about the probable true population values based on the following model (see Equation 5.15):

$$Y_j = \alpha + \beta_f X_{f_j} + \beta_{i1} X_{i1_j} + \beta_{i2} X_{i2_j} + \cdots + \beta_{i+} X_{i+_j} + \varepsilon_j, \tag{6.2}$$

where the subscript *j* refers to the *j*th observation, α and the βs refer to the population parameters, and the error term ε captures other systematic influences on *Y*, random variation, and measurement error.

Once again, our primary focus from the perspective of the elaboration model concerns the null hypothesis for the focal relationship, $H_0 : \beta_f = 0$. As is the case for OLS/SRS, this hypothesis is tested in two steps. The first step is the overall test for the regression equation, $H_0 : \beta_f = \beta_{i1} = \beta_{i2} = \cdots = \beta_{i+} = 0$. This hypothesis is tested with an overall modified *Wald* test statistic, which follows an *F* distribution (Heeringa et al., 2010). It is analogous to the overall *F* test statistic used with OLS/SRS. However, instead of the usual *degrees of freedom (df)* of *k*, $n - k - 1$, where *k* is the number of variables in the model, the *design degrees of freedom* (*ddf*) are fixed to the number of clusters minus the number of strata. Thus, the *df* for the regression equation are *k*, *ddf*. Alternately, an adjusted *Wald* test can be used with $df = k$, $dff - k + 1$. As Heeringa and associates (2010) point out, sample designs with large *df* permit more precise estimation of the true variance parameters of the reference distribution.

If the null hypothesis for the regression equation is rejected, then we conclude that at least one of the coefficients probably differs from 0. Because the overall test is uninformative about the individual coefficients, the null hypothesis $H_0 : \beta_f = 0$ is tested for the coefficient. As in OLS/SRS, the *t* statistic is used for this purpose:

$$t = \frac{b_f - 0}{SE(b_f)}, \tag{6.3}$$

where $SE(b_f)$ is obtained in the manner described above; the *ddf* are used to determine the statistical significance. Likewise, the robust *SEs* are used to calculate a design-based CI, $b_f = \pm t_{\alpha/2}[SE(b_f)]$.

In the elaboration method of analysis, a model-building strategy is used in which sets of variables are sequentially added to the model to determine their effect on the estimate of the focal relationship. The test of the null hypothesis that the coefficients for all of the m variables that are added equal 0 is the modified partial *Wald* test statistic, which is distributed as F with df equal to m, ddf, (Heeringa et al., 2010) or an adjusted partial *Wald* test with $df = m$, $dff - m + 1$. This test requires nested models in which the variables in the restricted model are a subset of those in the expanded model and the model is estimated on the same sample, that is, the n is constant. It is equivalent of the incremental F test with OLS/SRS. If the null hypothesis cannot be rejected, then the restricted model is preferred on the basis of parsimony. If it is rejected, then the coefficients for the individual variables are evaluated using the t test or CIs just described.

Explanation of the Dependent Variable

Like OLS/SRS, regression that adjusts for a complex sample design yields a value of R^2 that gives the correlation between the dependent variable and the optimal linear combination of independent variables, that is, the correlation between Y and $\hat{Y}$. However, this is a weighted version of R^2 such that the squared differences contributing to the sums of squares are weighted by the observation's sample weight (Heeringa et al., 2010). Its significance is given by the test of the regression equation.

The change in R^2 between nested models gives the contribution of the added variables to explaining Y. It too is weighted. The significance of the increment in R^2 is provided by the modified partial or adjusted *Wald* test statistic just described.

Subgroup Analysis

The analysis of subgroups with data from complex samples differs from the same analysis conducted with an SRS. West (2008) explains that estimates of *SE*s should reflect theoretical sample-to-sample variation based on the original complex sample design; therefore, the entire sample should be retained rather than deleting the cases that are not in the subgroup of interest. For instance, both men and women should be retained in the sample even when the analysis is restricted to women. Similarly, Lumley (2004) makes it clear that a subpopulation of a survey cannot be treated simply as a smaller survey.

West cites two primary problems with the dropping cases for subgroup analysis. First, sample-to-sample variability in the size of the subgroup should be incorporated into *SE*s, but dropping cases that are not in the subgroup treats the subgroup size as fixed. The second problem identified by West is that the entire first-stage sampling clusters are dropped from the analysis when they do not contain any cases in the subgroup by chance, which results in underestimates of *SE*s because these clusters are not recognized by the software as ever being part of the original sample design. Relatedly, he notes that there may be only a single cluster within a stratum, which is problematic because within-stratum variance between the clusters cannot be estimated.

Instead, the entire sample should be retained with an indicator variable for whether the case is in the subgroup of interest or not. As West explains, this indicator variable enables the software to recognize the full complex design of the sample, treat the subgroup sample size as random, and estimate parameters and *SE*s based on the full complex sample design.[8]

Lumley (2004) likewise maintains that the correct analysis involves keeping the entire sample but assigning zero weight to observations not in the subpopulation. He further notes that this strategy maintains the equivalence between separate regression models in subpopulations and a common regression with interactions.

Incorporating the Sample Design Into Analysis

The information necessary to identify the sample design for recognition by the software is embedded in a set of variables created by a sampling statistician for this purpose. Depending on the design, these variables include a strata variable, a cluster variable, and one or more weight variables. These codes link the person to a particular cluster within a specific stratum. West explains that public use data sets usually contain only one cluster variable, and it is at the first stage because the variance at this level encompasses all of the variance in estimates due to later stages of cluster sampling.

The final sample weight for each individual may be the only weight variable in the data set, although sometimes, other weights are also provided. For example, it's not uncommon to have survey data that can be analyzed at the individual level and at the HH level. The individual weights differ from the HH weights because they additionally take into consideration the number of persons in the HH.

Other specifications that may be necessary include options for handling strata that contain only one cluster, which are problematic because variance estimates are based on variation within the stratum. Also, a finite correction factor may be needed when the sample constitutes a large fraction of the population.

The Question of Weights

Some social scientists contend that sample weights are unnecessary because multivariate analyses typically control statistically for the same set of variables that go into the sample design and calculation of sample weights. Strata are often defined by dimensions of SES and take race/ethnicity into account, variables that are almost always taken into consideration during data analysis. Although these statistical controls serve much the same purpose as sample weights, analyses are limited to the main effects of these variables, leaving out the complex ways in which these factors might interact in the design and execution of the sampling plan. For example, controlling for race/ethnicity and SES will not adjust fully for an especially low participation rate among African American women, whereas a poststratification weight could be used for this purpose. Also, the substitution of control variables in multivariate analysis

does not address the impact of differential selection probabilities on univariate point estimates (e.g., means and proportions).

Equally important, the argument about weights is often taken to imply that the impact of the sample design on variance estimation can be ignored too. Inaccurate *SE*s obtained under an assumption of an SRS have profound implications for tests of statistical significance and CIs, which argues forcefully for adjustments for stratification and clustering to avoid potential inferential errors and misleading study conclusions. In my opinion, this last consideration is of paramount importance and leads to the conclusion that the features of the complex sample, including weights for unequal selection probabilities, should be taken into consideration using software designed for this purpose.

Interpretation

Given all that has been said, it is decidedly anticlimactic to sum up the implications for the interpretation of results in a single sentence. Although the design features of complex samples are extremely consequential for parameter estimates and their *SE*s, when these features are taken into consideration using appropriate design-based software, the interpretation of regression results is the same as it would be for regression conducted with OLS/SRS.

HRS ANALYSIS JOURNAL 6.3

Regression With Complex Samples: Loneliness and Discrimination

The example of regression using survey data presented in this chapter concerns the focal relationship between exposure to discrimination and being lonely. Discrimination refers to the unequal treatment of persons or groups on the basis of some ascribed or perceived trait such as race/ethnicity, gender, age, or mental illness. It is distinguished from related phenomenon such as prejudice and stereotypes in that discrimination refers to behavior as distinct from the possible motivations for that behavior. Among other adverse social corollaries, the target of discriminatory actions is treated as "other"—not one of "us"—which may lead to feeling separated from people in general. For this reason, perceived discrimination is hypothesized to be positively related to loneliness—the feeling of being alone, lacking friends or companions, being socially isolated, and cut off from others.

These two constructs were assessed as part of the PQ among the HRS sample described earlier. *Loneliness* was measured with a three-item scale asking how often you feel "you lack companionship," "left out," and "isolated from others." Responses were coded as 1 = "hardly ever or never," 2 = "some of the time," and 3 = "often"; thus, a high score indicates that the person frequently feels lonely (Hughes, Waite, Hawkley, & Cacioppo, 2004). Responses were averaged across the three items ($\alpha = .81$) to maintain

the metric of the response categories. Loneliness generally tends to be infrequent ($\bar{X} = 1.498$, $SD = 0.549$), equivalent to an average response midway between "hardly ever or never" and "some of the time"; however, the maximum score of 3 was observed, indicating that some respondents feel alone more often than not.

The focal independent variable is perceived *everyday discrimination,* a five-item scale that asked respondents, "How often in your day-to-day life have any of the following things happened to you," such as "you are treated with less courtesy or respect than other people," and "you receive poorer service than other people at restaurants or stores" (Williams, Yu, Jackson, & Anderson, 1997). Response codes were 0 = "never," 1 = "less than once a year," 2 = "a few times a year," 3 = "a few times a month," 4 = "at least once a week," and 5 = "almost every day." Responses were averaged across items ($\alpha = .80$). Everyday discrimination usually is conceptualized and operationalized differently than major lifetime experiences of discrimination, described earlier, which includes actions such as being unfairly denied a promotion or a bank loan (see also Chapter 10).

Perceived exposure to everyday discrimination is quite low on average ($\bar{X} = 0.680$, $SD = 0.755$): The mode is 1 (one third of the sample)—a response of "never" to all five items. Nevertheless, some respondents report encountering discrimination on a daily basis, given that the maximum score possible is observed.

Respondents also were asked, "What do you think were the reasons why these experiences happened," and could give multiple responses. Of those who cited a reason, half of them reported more than one reason. The most common attribution by far is age, cited by roughly 3 in 10 of all respondents. About 10% of the sample identified the actions as arising in response to their race/ethnicity or gender.

The simple regression of loneliness on everyday discrimination is shown as Model 1 in Table 6.3. The *F* statistic for the regression equation is obtained from the adjusted *Wald* test statistic, as mentioned above, and we see that it has *df* of 1, 56 because there is one variable in the equation and there are 112 clusters and 56 PSU yielding $ddf = 56$[9], such that $df = k$, $dff - k + 1$.

Based on this test, the null hypothesis that everyday discrimination in not associated with loneliness is rejected. Thus, we conclude that it is extremely unlikely that a coefficient of this magnitude or larger would be observed if in fact the two variables are independent of one another in the population. Because this is a simple regression, the test of the regression equation is equivalent to the test of the individual coefficient for discrimination. The *SE* for this coefficient is adjusted for the complex sample design using Taylor series linearization. This robust *SE* is used in the *t* test for the coefficient for discrimination such that $t = 26.33$ (0.260/0.010). And, $t^2 = F$. By comparison, the *SE* estimated incorrectly with OLS/SRS is substantially smaller (0.006), yielding a much larger value of $t = 40.70$. This illustrates the inflation in tests of statistical significance that can result from not adjusting for a complex sample design.

The expected change in Y with a 1-unit increase in X_f is equivalent to about a half an *SD* (0.260/0.549 = 0.474) of loneliness. The difference in the expected value of loneliness between those who never experience discrimination and those who experience it

Table 6.3 Analysis of Complex Samples: Regression of Loneliness on Everyday Discrimination and Social Support

	Loneliness								
	Model 1			*Model 2*			*Model 3*		
Independent Variables[a]	b	*Robust* SE	p	b	*Robust* SE	p	b	*Robust* SE	p
Everyday discrimination	0.260	0.010	***	0.210	0.010	***	0.206	0.010	***
Social support				–0.273	0.012	***	–0.269	0.012	***
Age (–52)							–0.009	0.002	***
Age $(-52)^2$							0.0002	0.00004	***
Female							0.055	0.012	***
Race/ethnicity									
African American							–0.019	0.019	
Latino							0.006	0.023	
"Other"							0.031	0.050	
Education (years)							–0.009	0.002	***
Income (/$1,000)[b]							–0.034	0.008	***
Wealth (/$1,000)[b]							–0.050	0.034	
Marital status									
Divorced/ separated							0.198	0.020	***
Widowed							0.216	0.014	***
Never married							0.236	0.032	***
Employment status									
Retired							0.029	0.016	
"Other"							0.067	0.021	**
Constant	1.321	0.010	***	2.207	0.045	***	2.816	0.248	***

	Loneliness								
	Model 1			*Model 2*			*Model 3*		
Independent Variables[a]	b	*Robust* SE	p	b	*Robust* SE	p	b	*Robust* SE	p
Model Statistics									
F	693.48		***	715.40		***	134.84		***
df	1, 56			2, 55			16, 41		
R^2	.127			.192			.260		
Model Comparisons[c]									
F				487.32		***	58.57		***
df				1, 56			14, 43		
ΔR^2				.065			.067		

Note: $N = 12{,}236$; *SE* = standard error.

a. Reference categories: Race/ethnicity = non-Hispanic White; marital status = never married; employment status = employed.

b. Log transformed.

c. ΔR^2 = Difference in R^2: Model 2 compared with Model 1; Model 3 compared with Model 2; ΔR^2 values may not sum exactly due to rounding errors.

$^*p \leq .05$. $^{**}p \leq .01$. $^{***}p \leq .001$.

daily is 1.300, on a scale of loneliness with a range of 2. The R^2 value similarly indicates that there is a medium-sized association.

In the next step, social support is taken into consideration because it is often conceptualized as a coping resource that may offset the deleterious effects of exposure to stressors, and discrimination is considered to be a pernicious stressor (Thoits, 2010). If social support offsets exposure, it is because discrimination is positively associated with support; for example, one's victimization may prompt others to come to one's assistance, demonstrating that you are cared for and valued by others—thus, offsetting the effect of everyday discrimination on loneliness. However, discrimination may erode a person's sense of being connected to people in general, including those who otherwise would be sources of social support. Thus, a negative association between these two variables would suggest that discrimination depletes support rather than activates it: People may withdraw because they become tired of helping a person cope with a chronic stressor, for instance. Finally, supportive social ties should diminish feelings of loneliness.

The most important dimension of social support for psychological well-being appears to be socioemotional support: the perception that your basic social needs—affection, esteem, approval, belonging, identity, and security—are satisfied through interaction with

others (Cassel, 1976; Cobb, 1976; Thoits, 1983, 2011). In the PQ, *social support* was measured with three items that were asked separately as they pertain to spouse/partner ($\alpha = .81$), children ($\alpha = .82$), other family ($\alpha = .86$), and friends ($\alpha = .83$). Participants were asked how much these persons "really understand the way you feel about things," "can you rely on them if you have a serious problem," and "can you open up to them if you need to talk about your worries." Responses were coded from 1 = "not at all" through 4 = "a lot" and were averaged across items for each type of relationship. These averages were then summed and divided by the number of sources reported (e.g., two scores were averaged for someone who reported support only from family and friends). Given a maximum score of 4, the mean ($\overline{X} = 3.116$, $SD = 0.530$) indicates that respondents, on average, enjoy considerable support from their family, friends, children, and spouses; However, others derive little support from them as evidenced by scores of 1. In preliminary bivariate analysis, social support is negatively correlated with both everyday discrimination and loneliness; therefore, it has the potential to mediate the effect of discrimination on loneliness.

As in Model 1, the *F* test of the regression equation for Model 2 (Table 6.3) is based on the adjusted Wald test statistic and indicates that we can reject the null hypothesis that the regression coefficients for discrimination and social support are equal to each other and equal to 0. This finding makes it appropriate to examine the individual coefficients, both of which are statistically significant: Everyday discrimination is positively associated with loneliness with social support held constant, and social support is negatively associated with loneliness when discrimination is controlled. It bears repeating that the *t* test for the individual coefficients is based on the robust *SE* that is adjusted for the complex sample using Taylor series linearization. Therefore, it is smaller than the values obtained by incorrectly using OLS/SRS (21.26 vs. 33.10 for discrimination and −22.08 vs. −31.47 for social support), again demonstrating the inflation of tests of statistical significance that accompanies the inappropriate analysis of complex sample data with methods based on an assumed SRS.

Model 1 is nested within Model 2, so that the adjusted partial *Wald* test can be used to assess the statistical significance of adding social support. However, given that only one variable was added, this 1-*df* test is equivalent to the *t* test of the coefficient for social support. The increment in R^2 between Model 1 and Model 2 shows that social support makes a small- to medium-sized contribution to the explanation of loneliness. The total R^2 value is a substantial level of multivariate association between the dependent variable and the two independent variables.

When social support is added in Model 2, the coefficient for everyday discrimination is reduced by roughly 20% ($100 \times [0.260 - 0.210]/0.260 = 19.2\%$), although it remains statistically significant. The interpretation of this change hinges on the inverse association between discrimination and social support, which shows that people who have encountered frequent acts of discrimination are the same people who tend to have the least social support. Thus, rather than counteracting the effects of discrimination, the diminishment of social support may be one of the ways in which discrimination produces loneliness.

Model 3 in Table 6.3 illustrates the addition of a set of several variables to the regression equation, in this instance, control variables for sociodemographic characteristics that may be related to loneliness on the one hand and be the object of discriminatory acts on the other hand: age, gender, and race/ethnicity. (As mentioned previously, the association of loneliness with age is curvilinear and is modeled as age and age^2; see Figure 3.2.) Other demographic characteristics are added for the same reasons, although their connections to the focal variables are not so readily apparent. Marital statuses other than married and employment statuses other than being employed may result in being treated differently and unequally, and are likely to be associated with loneliness; SES is controlled too (education, income, and wealth), given that being poor is a stigmatized condition.

Based on the test of the regression equation, conducted with an adjusted *Wald* test, the null hypothesis that all of the regression coefficients equal 0 can be rejected. Note again that the *df* are *k* and *ddf* − *k* + *1*, where *k* is the number of variables (16) and the *ddf* = the number of strata (112) minus the number of PSU (56). Under OLS/SRS, the *df* for this model and data are much greater: *k*, *n* − *k* − 1 (16, 12,219).

The adjusted partial Wald test is used to compare the restricted Model 2 with the expanded Model 3. As shown in Table 6.3, the *F* value indicates that the null hypothesis that the coefficients for the variables added to Model 3 all equal 0 can be rejected. Recall that this test is distributed like *F* and has *df* of *m*, *ddf*, where *m* is the number of added variables, in this case, 14, 43. If these data had been collected from an SRS of the same size, then the *df* would be 14, 12,219. The smaller value in the denominator means that the *p* values and CIs are larger for the complex sample. Also, the value of *F* is substantially greater when this model is incorrectly estimated with OLS/SRS (81.54). This inflation of the test statistic opens the door to inferential errors.

Given that the null hypothesis can be rejected, it is appropriate to examine the individual variables added to Model 3. As before, the statistical significance of these coefficients is based on the *t* test, which is the weighted estimate of the coefficient divided by the robust *SE* that is adjusted for the complex sample design. As can be seen in Model 3, Table 6.3, compared with those who are married, those in other marital statuses experience more frequent loneliness, other factors held constant, as do persons in the "other" employment category compared with the employed. Women more often feel alone than do men. There are no significant differences between non-Hispanic Whites and the other racial/ethnic groups (net of the other variables in the model). Age and age^2 continue to define a curvilinear association with loneliness first decreasing until about age 75 and then reversing course and increasing, other things held constant. Higher levels of education and income (but not wealth) are inversely associated with loneliness.

From the perspective of the elaboration model, we are interested in whether the addition of these sociodemographic characteristics alters the estimate of the focal relationship. As can be seen, the change is barely perceptible, smaller than the *SE* of the coefficient, meaning that it is negligible. Thus, apart from social support, the other variables in this analysis do not contribute to the explanation of the focal relationship.

The impact of ignoring the sample design can be seen with the DEFT values for the regression coefficients in Model 3, where the DEFT is the ratio of the *SE* under the actual

HRS design to the SE of a hypothetical SRS of the same size (12,236) of the same population, that is, the 2007 population of the U.S. born before 1924. The mean across the 16 variables in Model 3 is 1.266.[10] Two variables have values less than 1.000: age^2 (0.926) and widowed (0.939). The largest DEFF values are for log wealth (1.831), "other" race/ethnicity (1.519), and everyday discrimination (1.418), that is, both sociodemographic characteristics that are implied in the HRS sample design and a psychosocial variable that is not. Given that these *SE*s are used in the calculation of the *t* tests for these coefficients, as just discussed, we risk making substantial inferential errors for this example if we ignore the sample design.

The R^2 and change in R^2 values are calculated using weights for each observation, as discussed above. The increment in R^2 shows a small- to medium-sized contribution of the set of sociodemographic variables to explaining *Y*. The total R^2 value of Model 3 shows that a quarter of the variance in loneliness is explained by the entire set of control variables, social support, and discrimination—a very good level of explanatory efficacy in the social sciences that nevertheless leaves three quarters of the variance unexplained.

The decrease in R^2 obtained by dropping everyday discrimination from Model 3 is small to medium (0.071), and gives the amount of variance in loneliness that is exclusively due to the focal independent variable over and above what is attributed to the control variables and social support. This demonstrates the importance of understanding the role of discrimination in loneliness among older adults.

In conclusion, the results of this analysis are consistent with the idea that frequent exposure to everyday discrimination leads to feeling alone and that it may do so because it has a corrosive effect on feelings of being supported by significant others. The finding that discrimination is associated with loneliness net of sociodemographic controls substantiates this interpretation because the association has survived the test that it is merely a spurious association. Although these characteristics do little to explain the focal relationship, they do contribute to the explanation of loneliness. Net of the other variables in the model, discrimination continues to be associated with loneliness, suggesting that further specification of mediators of this association would be a productive next step. This interpretation is quite speculative, however, given that this analysis is an incomplete example of a test of the relationship between everyday discrimination and loneliness.

This example illustrates quite clearly that analyzing data from a complex sample as if it were obtained from an SRS can produce substantial inflation in tests of statistical significance, leading to potential inferential errors.

Summary

Much of what is taught in standard statistics courses is predicated on the assumption that the data have been collected from an SRS, but complex samples are encountered far more frequently than SRSs in large-scale social science surveys. Although SRSs have very desirable statistical properties, cost and logistical concerns make complex samples far more practical.

Complex samples usually are stratified, which means that the population is subdivided into strata that are as internally homogeneous as possible with regard to the variables being studied and distinctly different from each other on these variables. Stratification improves the precision of parameter estimates, so that it generates *SE*s that are smaller than would be obtained with an SRS of the same size. In contrast, clustering tends to increase *SE*s because only some of the clusters are sampled, so that observations may be more different from one another than would have been the case had all clusters been sampled. Sample weights also tend to increase the variance of parameter estimates, except for poststratification weights, which slightly decrease variance estimates. When these multiple influences are combined, complex samples usually generate larger *SE*s than SRSs of the same population of the same size, thereby inflating tests of statistical significance and potentially leading to inferential errors.

There are several methods to adjust *SE*s for the complex design, including Taylor series linearization and resampling using either BRR or Jacknife procedures. Tests of statistical significance, such as the adjusted *Wald* test statistic for the regression equation and the *t* test for the individual coefficients, incorporate the sample design into their calculations.

In addition to their influence on *SE*s, complex samples affect the parameter estimates themselves because the elements of the sampling frame typically have unequal selection probabilities. Sample weights adjust for these differences, including oversampling of some segments of the population and the corresponding undersampling of other segments of the population. Weights also may be applied to minimize bias due to nonresponse. In addition, poststratification weights are sometimes used to align the distribution of the sample with the known distribution of the population on the types of characteristics that are used to stratify both samples and society, such as age, gender, race/ethnicity, and SES.

The correct analysis of data obtained from a complex sample necessitates the use of specialized procedures that are available in major statistical software packages that take into consideration stratification, clustering, and sample weights. The method of estimation and tests of statistical significance used in these procedures differ from OLS/SRS. However, the interpretation of results is identical.

Although there is a long history of ignoring the statistical implications of complex samples in survey data analysis, it is no longer scientifically acceptable to do so.

Notes

1 A **robust *SE*** is one that is resistant to errors produced by deviations from assumptions.

2 In Stata, these design features are analyzed with the survey (•svy) estimator commands. First the •svyset command is used to identify strata, clusters, and sample weights and other features of the sample so that they are recognized by the software. Then the •svy prefix is used to put into effect the specific type of statistical procedure. For example, for multiple linear regression, the survey estimator command is, •svy: regress followed by the dependent variable and a list of the independent variables.

3 Participants were added to these enrollment cohorts over time, and the sample also includes "age-ineligible" spouses, that is, husbands and wives born outside the birth cohorts. The entire sample of

persons enrolled is used as the base for the analyses reported here, and cohort is measured with the variable *wtcohort*, which is based on birth date as reported at enrollment in the study. The AHEAD and CODA samples were supplemented with persons drawn from a list of age-eligible people enrolled in Medicare.

4 Other uses of weights are to adjust for sample attrition over time for longitudinal studies.

5 The sample design features for the HRS data were set as follows: •svyset *RAEHSAMP* [pweight=WT_PQ], strata (*RAESTRAT*). These RAND HRS variables correspond to the HRS Tracker variables *SECU* and *STRATUM* and identify the stratum and half sample, respectively. *WT_PQ* is the sample weight for the PQ data created for these analyses from the HRS-supplied weights *KLBWGTR and LLBWG* (renamed in lower case for convenience), such that *WT_PQ* = *klbwgtr*/2 or *llbwg/2*, depending on whether the respondent was sampled for the 2006 or 2008 PQ, respectively.

6 Slight discrepancies in calculations here and elsewhere are due to rounding error; calculations are carried to more decimal points than shown in the text for accuracy.

7 Given a dearth of diagnostic tests for regression with complex sample data, Heeringa and colleagues (2010) recommend using the tools available in usual regression software.

8 This can be accomplished, for example, by using the •subpopulation procedure in Stata. For a complete discussion of this topic, see UCLA Statistical Consulting Group (n.d.).

9 In the Stata •svy: regress procedure, the default is an adjusted *Wald* test with $df = k$, $ddf - k + 1$ instead of $df = k$, ddf. For the comparison of nested models, the $df = m$, $ddf - m + 1$, where m is the number of variables differentiating the restricted and expanded models. The following command yields the unadjusted model: •test *varlist*, nosvyadj.

10 These values were obtained in Stata with the following command •estat effects, deft.