Chapter 12
Complex Surveys

Suppose we wanted to estimate the prevalence of diabetes among adults in the US, as well as the effects of diabetes risk factors in this broad target population, both with minimum bias—that is, in such a way that the estimates were truly representative of the target population. Observational cohorts that might be used for these purposes are usually convenience samples, and are often selected from subsets of the population at elevated risk. This would make it difficult to generalize sample diabetes prevalence to the broader target population. We might be more comfortable assuming that sample associations between risk factors and diabetes were valid for the broader population, but the assumption would be hard to check (Problem 12.1).

Observational studies as well as randomized trials use convenience samples for compelling reasons, among them reducing cost and optimizing internal validity. But when unbiased representation of a well-defined target population is of paramount importance, special methods for obtaining and analyzing the sample must be used. Crucial features of such a study are

- All members of the target population must have some chance of being selected for the sample.
- The probability of inclusion can be defined for each element of the sample.

Using data from a sample which meets these two criteria, we could in principle compute unbiased estimates of the number and percent prevalence of diabetes cases in the US adult population, as well as of the effects of measured diabetes risk factors. Surveys implemented by the National Center for Health Statistics (NCHS), including the National Health and Nutrition Examination Survey (NHANES), the National Hospital Discharge Survey (NHDS), and the National Ambulatory Medical Care Survey (NAMCS), are prominent examples of surveys that meet these criteria. Data sets based on these surveys are publicly available on the NCHS website www.cdc.gov/nchs/.

In this chapter, we give only a brief overview of the design and implementation of these surveys, which are complicated and expensive undertakings. Our primary purpose is to provide guidance for secondary analyses using complex survey data.

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_12,

Fortunately, Stata and other statistical packages make it straightforward and transparent to account properly for the special features of the sampling design in regression analyses using complex survey data.

12.1 Overview of Complex Survey Designs

To provide unbiased estimates of population parameters, complex survey data are *weighted* in inverse proportion to the known probability of inclusion. In addition, to reduce costs, a *complex sampling design* is often used. In many cases, this means initially sampling clusters, known as primary sampling units (PSUs), rather than individuals; only at some later stage are individual study participants selected. This is in contrast to a simple random sample (SRS), in which individuals are directly and independently sampled. Finally, complex samples are often *stratified*, in that the PSUs are sampled within mutually exclusive strata of the target population.

Inverse Probability Weighting

A primary feature of complex surveys is *inverse probability weighting* (IPW). Introduced in Sect. 11.9.3 for dealing with missing data, IPW is the way complex surveys use well-defined probability of inclusion to obtain representative estimates, as we explain below in Sect. 12.2.

One advantage of IPW is that it accommodates *unequal* probability of inclusion in the survey sample. In part, unequal inclusion probabilities arise naturally from variability in the size of primary and secondary clusters. In addition, subgroups of special interest may be sampled at higher rates, so that they comprise a larger proportion of the sample than they do of the target population. The rationale is to ensure adequate precision of estimates both within the subgroup and in contrasting the subgroup to other parts of the larger population, by increasing their numbers in the sample. IPW ensures that overall estimates properly reflect the population proportions comprised by the over-sampled subgroups.

Cluster Sampling

From Chap. 7, it should be clear that the initial sampling of clusters may affect precision, because outcomes for the observations within a cluster are positively correlated in most cases. The change in precision means that for many purposes a larger sample will be required to achieve a given level of statistical certainty. Nonetheless, the complex survey design is cost-effective, because cluster sampling can be implemented in concentrated geographic areas, rather than having to cover the entire area where the target population is found. Moreover, some of the information required to define probability of inclusion need only be obtained for the selected clusters. Especially for nationally representative samples, the savings can be considerable.

In *multistage* designs, there may be several levels of cluster sampling; for example, counties may initially be sampled, and then census tracts within counties, city blocks with census tracts, and households within blocks. Only at the final stage are individual study participants sampled within households. The rationale is again to reduce costs by making the survey easier to implement.

Stratification

An additional feature of many complex surveys is that clusters may be selected from within mutually exclusive and exhaustive *strata*, usually geographic, which cover the entire target population. To the extent that subsets of the target populations are more similar within than across strata, this can increase the precision of estimates of population means and totals.

Example: NHANES

NHANES is a series of complex, multistage probability samples representative of the civilian, noninstitutionalized US population. Interviews and physical exams are used to ascertain a wide range of demographic, risk-factor, laboratory, and disease outcome variables. In NHANES III, conducted between 1988 and 1994, the PSUs were primarily counties. Thirteen large PSUs were selected with certainty, and the remaining 68 were selected with probability proportional to PSU population size, two from each of 34 geographic strata. At the second stage of cluster sampling in NHANES III, area segments, often composed of city or suburban blocks, were selected. In the first half of the survey, special segments were defined for new housing built since the 1980 census, so that no portion of the target population would be systematically excluded; in the second half, more recent information from the 1990 census made this unnecessary. The third stage of sampling was households, which were carefully enumerated within the area segments. At the fourth and final stage, survey participants were selected from within households.

At each stage, sampling rates were controlled so that the probability of inclusion for each participant could be precisely estimated. Children and people over 65 as well as African Americans and Mexican Americans were over-sampled. Almost 34,000 people were interviewed and of these roughly 31,000 participated in the physical exam. Data from NHANES III have been used in many epidemiologic and clinical investigations.

12.2 Inverse Probability Weighting

We pointed out that in selecting a representative sample, every member of the target population has to have some chance of being included in the sample. To put it another way, no part of the target population can be systematically excluded.

In addition, we said that for every element of the sample, the probability of inclusion must be known. Essentially this is what is meant by a so-called *probability sample*. Analysis of such samples makes use of information about probability of inclusion to produce unbiased estimates of the parameters of the target population.

To see how this works, consider a SRS of size 100, drawn at random from a target population of size 100,000. In this simple case, each member of the sample had a one-in-a-thousand chance of being included in the sample. The so-called *sampling fraction*, another term for the probability of inclusion, would be 0.001 for this sample, and constant across observations. Furthermore, we could think of each member of the sample as representing 1,000 members of the target population. If we wanted to estimate the percent prevalence of diabetes in the target population, the proportion with diabetes in the sample would work fine in this case, for reasons that we explain below. Likewise, the average age of the sample would be an unbiased estimate of mean age in the population.

Now consider the more interesting case of estimating the *number* of diabetics in the population. Suppose there were five diabetics in the sample. Since each represents 1,000 members of the target population, an unbiased (though obviously noisy) estimate of the population number of diabetics would be 5,000. Essentially this would be a *weighted sum* of the number of the diabetics in the sample, where each gets weight 1,000, or the number in the population that each sample participant represents. Formally, the weight is the reciprocal of the sampling fraction of 0.001. Note that the overall sum of these sample *inverse probability weights* equals the population size.

> *Definition*: *Inverse probability weights* are the reciprocal of the probability of inclusion, and are interpretable as the number of elements in the target population which each sampled observation represents.

Next, consider the more typical case where the probability of inclusion varies across participants. To make this concrete, suppose that women and men both number 100,000 in the target population, but that the sample includes 200 women and 100 men, for sampling fractions of 0.002 and 0.001, respectively. In this sample, each man represents 1,000 men in the population, but each woman represents only 500 women. Thus the IPW for each man in the sample would be 1,000, and for each woman, 500.

In this case, to estimate means for the whole target population, we would need to use *weighted* sample averages. These would no longer equal their unweighted counterparts, in which men would be under represented. The formula for the weighted average is

$$\mathrm{E_w}[Y] = \frac{\sum_i w_i y_i}{\sum_i w_i}, \tag{12.1}$$

where $\mathrm{E_w}[Y]$ denotes the weighted average of the outcome variable Y, y_i is the value of Y for participant i, and w_i is the corresponding probability weight.

Furthermore, if Y were a binary indicator variable coded $1 =$ diabetic and $0 =$ nondiabetic, then (12.1) also holds for estimating the population proportion

with diabetes. As we pointed out in Sect. 4.3, this equivalence between averages and proportions only holds with the 0–1 indicator coding of Y. In addition, with this coding of Y, the weighted estimate of the total number in the population with diabetes is simply $\sum w_i y_i$—the sum of the weights for the diabetics in the sample.

12.2.1 Accounting for Inverse Probability Weights in the Analysis

Taking account of the IPWs, which are included in the NHANES, NHDS, NAMCS, and other NCHS datasets, is essential for obtaining unbiased estimates. The differences between the weighted and unweighted estimates can be considerable. For example, the unweighted proportion with diabetes among adult respondents in NHANES III is 7.4%, but the weighted proportion is 4.8%. The corresponding unweighted estimate of the number of adult diabetics at the time of NHANES III was 12.5 million, as compared to a weighted estimate of 8.1 million—not a trivial difference for estimating the burden of disease and health services needs. All statistical packages for complex surveys accommodate IPWs.

12.2.2 Inverse Probability Weights and Missing Data

Estimation of population parameters, in particular totals, means, and proportions, can be quite vulnerable to missing data. The potential for bias arises because the non-responders usually differ systematically from responders, especially when the response of interest is sensitive. The nonresponders are not *missing completely at random* (MCAR). However, we might be willing to assume that the data are MCAR within relatively homogeneous demographic subgroups defined by measured covariates. In the framework of Chap. 11, the data are assumed to be *CD-MCAR*.

12.2.2.1 Adjustment of IPWs to Account for Unit Non-Response

In NHANES as in many complex surveys, the inverse probability weights are adjusted to account for missing observations—so-called *unit non-response*—in such a way as to minimize the potential for bias. Under the CD-MCAR assumption, the inverse probability weights are adjusted within relatively homogeneous demographic subgroups. Specifically, for each such subgroup, the weights for the responders are inflated by a fixed factor, determined so that the adjusted weights for the responders sum to the total of the original inverse probability weights for both responders and non-responders. In short, the responders in the subgroup are made to stand in for the non-responders.

In many complex surveys, a second so-called *poststratification* adjustment is made to ensure that the IPWs sum to regional totals for the target population, which are known from the US Census.

12.2.2.2 Multiple Imputation to Account for Item Non-Response

In addition to unit non-response, we also need to be concerned about *item* non-response, or missing responses on particular questions by study participants. The recommended approach to item non-response in complex surveys is *multiple imputation* (Rubin 1987, 1996); see Sect. 11.5.

12.3 Clustering and Stratification

In contrast to accounting for the inverse probability weights, which is required mainly to avoid bias, taking account of the stratification and clustering of observations due to the complex sampling design is required solely to get the standard errors, CIs, and P-values right, and has no effect on the point estimates. Unlike the point estimates, standard errors accounting for the special characteristics of a complex survey do differ from what would be obtained in standard weighted regression routines, sometimes in ways that are crucial to the conclusions of the analysis.

The default standard errors, CIs, and P-values provided by most survey packages including Stata are calculated using so-called *linearization*. These are closely related to the robust standard errors available with many Stata regression commands, and thus account, as with longitudinal and hierarchical data, for clustering. In Stata, the main difference is that in testing whether each estimated regression coefficient differs from zero, the survey routines use a t-test with degrees of freedom equal to the number of PSUs minus the number of strata, rather than the asymptotic Z-test used in GEE. In addition, stratification is taken into account.

Of note, these methods for analyzing survey data do not directly extend to random effects models, introduced in Chap. 7, which represent a different approach to clustered data. Rabe-Hesketh and Skrondal (2006) propose a pseudo-likelihood approach to analyzing multi-level data with a binary outcome, which is implemented in the downloadable `gllamm` package for Stata.

12.3.1 Design Effects

Because of positive correlation within clusters, the standard errors of parameter estimates from a complex survey are often (but not always) inflated as compared

to estimates from a SRS of the same size. This inflation can be summarized by a *design effect:*

> *Definition*: The *design effect* is the ratio of the true variance of a parameter estimate from a complex survey to the variance of the estimate if it were based on data from a simple random sample.

Note that design effects can vary for different parameters estimated in the same survey, because some predictors may be more highly concentrated and some outcomes more highly correlated within clusters than others. Furthermore, design effects in regression may vary with the degree to which the regression effect is estimated by contrasting observations within as opposed to between clusters.

12.4 Example: Diabetes in NHANES

Stata makes it easy to run a regression analysis taking account of the special features of a complex survey. Variables identifying the PSU, IPW, and stratum for each observation are first specified using the `svyset` command. For our NHANES example, the `svyset` command takes the form

```
svyset sdppsu6 [pweight = wtpfqx6], strata(sdpstra6)
```

The regression is then run using the usual Stata commands, in conjunction with the `svy:` command prefix.

Table 12.1 shows three logistic models for prevalent diabetes estimated using data from NHANES III. The predictors are age (per 10 years), ethnicity, and sex. The reference group for ethnicity is whites. The odds-ratio estimates given by unweighted logistic regression (Model 1) differ both quantitatively and qualitatively from the results of the weighted and survey analyses (Models 2 and 3), which are identical. In the unweighted model, women appear to be at about 20% higher risk, but this does not hold up after accounting for probability of inclusion; similarly, the apparently increased risk among African Americans and Mexican Americans is smaller after accounting for the weights.

In addition, the standard errors differ across all three models, in part because the survey model takes proper account of stratification as well as clustering within PSUs. Note that in accommodating IPWs in Model 2, Stata by default uses robust standard errors, which are similar to the `Linearized Std. Err.` estimates given for Model 3.

We can obtain the design effect for each parameter estimate using the Stata `postestimation` command `estat effects, deff`. In the survey logistic model for prevalent diabetes shown in Table 12.1, the design effects are 2.7 for age, 0.93 for African American, 0.41 for Mexican American, 2.0 for other ethnicity, and 1.7 for sex. The increase in precision for the coefficient for Mexican Americans results from the strong concentration of this subgroup in a few PSUs, so that the comparison with whites rests primarily on within-cluster contrasts. In contrast,

Table 12.1 Unweighted, weighted, and survey logistic models for diabetes

```
. * Model 1: Unweighted logistic model ignoring weights and clustering
. logit diabetes age10 aframer mexamer othereth female, or nolog

Logistic regression                                     Number of obs   =      18140
                                                        LR chi2(5)      =    1148.81
                                                        Prob > chi2     =     0.0000
Log-likelihood = -4206.1375                             Pseudo R2       =     0.1202
------------------------------------------------------------------------------
    Diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age10 |   1.679618   .0284107    30.66   0.000     1.624847    1.736235
     aframer |   2.160196   .1651839    10.07   0.000     1.859534     2.50947
     mexamer |   2.784521   .2125535    13.42   0.000      2.39759    3.233896
    othereth |    1.25516   .2297557     1.24   0.214     .8767735    1.796845
      female |   1.200066   .0713788     3.07   0.002     1.068013    1.348447
------------------------------------------------------------------------------

. * Model 2: Weighted logistic model still ignoring clustering
. logit diabetes age10 aframer mexamer othereth female [pweight = wtpfqx6], ///
  or nolog

Logistic regression                                     Number of obs   =      18140
                                                        Wald chi2(5)    =     523.98
                                                        Prob > chi2     =     0.0000
Log-pseudolikelihood =  -28717819                       Pseudo R2       =     0.1124
------------------------------------------------------------------------------
             |               Robust
    Diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age10 |   1.704453   .0420649    21.61   0.000      1.62397    1.788925
     aframer |   1.823747   .1727191     6.34   0.000     1.514785    2.195726
     mexamer |   1.915197   .2029156     6.13   0.000     1.556068    2.357211
    othereth |   1.031416   .2386775     0.13   0.894     .6553287    1.623335
      female |   .9805769   .0992109    -0.19   0.846     .8041933    1.195647
------------------------------------------------------------------------------

. * Model 3: Survey model accounting for weights,stratification,and clustering
. svy: logit diabetes age10 aframer mexamer othereth female, or nolog
(running logit on estimation sample)

Survey: Logistic regression

Number of strata   =        49                  Number of obs      =      18140
Number of PSUs     =        98                  Population size    =  168471391
                                                Design df          =         49
                                                F(   5,     45)    =      80.86
                                                Prob > F           =     0.0000
------------------------------------------------------------------------------
             |             Linearized
    Diabetes | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age10 |   1.704453   .0479719    18.95   0.000     1.610726    1.803635
     aframer |   1.823747   .1840181     5.96   0.000      1.48903    2.233705
     mexamer |   1.915197   .1934747     6.43   0.000      1.56332    2.346276
    othereth |   1.031416    .225949     0.14   0.888     .6641157    1.601857
      female |   .9805769   .0921775    -0.21   0.836      .811784    1.184467
------------------------------------------------------------------------------

. estat effects, deff
------------------------------------------------
             |             Linearized
    Diabetes |      Coef.   Std. Err.       DEFF
-------------+----------------------------------
       age10 |   .5332443    .028145    2.72072
     aframer |   .6008932   .1009011     .933096
     mexamer |   .6498207   .1010208     .415208
    othereth |   .0309323   .2190668    1.98449
      female |  -.0196142   .0940033    1.67026
       _cons |  -5.798575   .2023545    3.05472
------------------------------------------------
```

women are about half of respondents in all PSUs, so that more of the information for the comparison with men comes from less efficient between-PSU contrasts (Problems 12.3 and 12.4).

In summary, accounting for IPW mainly affects the point estimates and secondarily the standard errors, while accounting for stratification and clustering only affects the latter.

12.5 Some Details

12.5.1 Ignoring Secondary Levels of Clustering

We pointed out earlier that NHANES is a *multistage* complex survey, meaning that area segments are selected within PSUs, then blocks with segments and households within blocks, before individuals are finally selected. Thus clusters are nested within clusters. For the NCHS surveys, multistage design is typical.

SUDAAN and recent versions of Stata make it possible to account more completely for the effects of multistage cluster sampling, by specifying identifiers for *secondary sampling units* (SSUs). They also accommodate so-called *finite population correction factors* to account for the fact that both PSUs and SSUs are sampled without replacement from relatively small "populations" of PSUs and SSUs.

However, only the stratum and PSU identifiers are provided with the NHANES data; to protect the confidentiality of survey respondents, no information is provided about area segment or block—the SSUs. Fortunately, in large samples like NHANES, the robust *sandwich* standard error calculations used in `svy` regression commands will properly reflect differences in the degree of correlation between observations sampled from the same or different SSUs within each PSU.

12.5.2 Other Methods of Variance Estimation

NHANES 2000, next in the series after NHANES III, began collecting data in 1999 and continues to sample yearly, using a similar complex multistage design. A nationally representative sample of approximately 5,000 participants is obtained each year. Data for the first two years were available in mid-2003. Although stratum and (psuedo) PSU identifiers have since been made available, they were not provided in 2003, to protect the confidentiality of study participants. Other surveys that do not provide stratum and PSU identifiers include the NHDS, and until recently, the National Ambulatory Medical Care Survey (NAMCS).

12.5.2.1 Relative Standard Errors

For the NHDS, constants for computing *relative standard errors* are provided with the documentation, so that approximate CIs for means and proportions can be calculated, but regression analysis is not possible.

12.5.2.2 Jackknife and Balanced Repeated Replication

Two other methods of variance estimation are implemented in Stata as well as the SUDAAN and WESVAR packages, and are compatible with regression analyses. The *jackknife* method uses a resampling procedure to estimate variability. The complete sample is split into K groups in such a way as to reflect the complex sampling structure but obscure geographic location, and a set of jackknife weights corresponding to each group is provided. In the kth set, the weights for group k are set to zero, and adjusted for the remaining groups, using adjustment methods already described for dealing with nonresponse. The analysis is then carried out $K + 1$ times, once with the original weights and once with each of the K sets of jackknife weights. It should be clear that the group with jackknife weights equal to zero will be omitted from that analysis. Then the variance of the overall estimates is estimated by variability among the jackknife estimates, appropriately scaled (Rust 1985; Rust and Rao 1996).

A related method for variance estimation called *balanced repeated replication* (BRR) is also implemented in Stata as well as SUDAAN and WESVAR, but is beyond the scope of this chapter.

12.5.3 Model Checking

In addition to accounting for clustering, stratification, and inverse probability weighting, we need to do standard model checking in regression analyses using complex survey data. These should include checks for linearity of the effects of continuous predictors, possibly using restricted cubic splines, and for omitted interactions. One useful tool is the Stata `postestimation` command `estat gof`, which extends the Hosmer–Lemeshow goodness of fit test to logistic and probit models for binary responses in survey data.

12.5.4 Postestimation Capabilities in Stata

Other useful `postestimation` commands, including `margins`, for obtaining average causal effects (Sect. 9.3.4), are also available. We also note that the factor

notation used to include categorical variables, quadratic terms, and interactions (Sects. 4.3 and 4.6) carries over without change to the Stata survey regression commands.

12.5.5 Other Statistical Packages for Complex Surveys

In addition to Stata, three other software packages make it straightforward to carry out descriptive as well as regression analyses using complex survey data. These packages include

- SUDAAN (Research Triangle Institute, Research Triangle Park, NC; www.rti.org),
- SAS (SAS Institute, Cary, NC; www.sas.com),
- WESVAR (Westat, Inc., Rockville MD; www.westat.com).

12.6 Summary

Complex surveys, unlike many convenience samples, can provide representative estimates of the parameters of a target population. However, to obtain these estimates and compute valid standard errors, CIs, and P-values, such surveys have to be analyzed using methods that take account of the special features of the design, including multistage cluster sampling, stratification, and the fact that not all members of the population have an equal chance of being included in the sample. A number of software packages make it straightforward to carry out multipredictor regression analyses using complex survey data.

12.7 Further Notes and References

For in-depth treatments of the many topics not covered in our brief overview focusing on regression analyses, leading books about complex surveys include Levy and Lemeshow (1999), Korn and Graubard (1999), Scheaffer (1996), Kish (1995), and Cochran (1977). These books deal comprehensively with the design of complex surveys and the underlying statistical theory. They also cover more specific topics including ratio estimators, variance estimation for subpopulations, and analysis of longitudinal surveys and using multiple surveys.

12.8 Problems

Problem 12.1. Taking HIV infection as an example, explain why it might be more problematic to generalize estimates of prevalence from a convenience sample than to generalize estimates of risk factor effects. For the latter, we essentially have to assume that there is little or no interaction between the risk factor and being represented in the sample. Does this make sense?

Problem 12.2. Show that (12.1) reduces to the unweighted average $\sum y_i / n$ when $w_i \equiv w$.

Problem 12.3. Judging from the logistic model shown in Table 12.1, which was used to assess risk factors for diabetes, design effects greater than 1.0 appear to be more common than design effects less than 1.0. Describe what would happen in these two cases to model standard errors, CIs, and P-values, if we were to analyze the survey data incorrectly, ignoring the clustering. In which case would we be more likely to make a type-I error? In which case would we be likely to dismiss an important risk factor? Can we reliably predict whether the design effect will be greater or less than 1.0?

Problem 12.4. In contrast to the design effects in regression analyses, design effects for means, proportions, and totals are almost always greater than 1.0. Explain why this should be the case.

12.9 Learning Objectives

(1) Describe the rationale for and special features of a complex survey.
(2) Identify what can go wrong if the analysis of a complex survey ignores inverse probability weights, strata, and cluster sampling.
(3) Know how to use data from NHANES III or a similar complex survey to validly estimate the parameters of multi-predictor linear and logistic regression models, with standard errors, CIs, and P-values that properly reflect the complex survey design.