Chapter 19
Statistical analysis of survey data

James R. Chromy
Research Triangle Institute
Research Triangle Park, North Carolina, USA

Savitri Abeyasekera
The University of Reading
Reading, UK

Abstract

The fact that survey data are obtained from units selected with complex sample designs needs to be taken into account in the survey analysis: weights need to be used in analyzing survey data and variances of survey estimates need to be computed in a manner that reflects the complex sample design. This chapter outlines the development of weights and their use in computing survey estimates and provides a general discussion of variance estimation for survey data. It deals first with what are termed “descriptive” estimates, such as the totals, means, and proportions that are widely used in survey reports. It then discusses three forms of “analytic” uses of survey data that can be used to examine relationships between survey variables, namely multiple linear regression models, logistic regression models and multi-level models. These models form a set of valuable tools for analyzing the relationships between a key response variable and a number of other factors. In this chapter we give examples to illustrate the use of these modeling techniques and also provide guidance on the interpretation of the results.

Key Words: complex survey design, analytic statistics, regression, logistic regression, hierarchical structures, multi-level modeling.

I. Introduction

1. Household surveys utilize complex sample designs to control survey costs. In addition, complete sampling frames that list all individuals or all households are usually not available. Even when population registries are available, the cost of implementing a household interview survey based on a simple random sample design would be prohibitively high. The Living Standards Measurement Study (LSMS) Surveys discussed in chapter 23 are a good example of many of the complex features of household survey designs.

2. A typical household survey design structure is shown in Table 1. Most sample designs for household surveys use complex sample designs involving stratification, multi-stage sampling, and unequal sampling rates as indicated above. Weights are needed in the analysis to compensate for unequal sampling rates and adjustments for non-response lead to more unequal weighting. The complex sample design needs to be taken into account in estimating the precision of survey estimates.

•

Table 1. Typical household survey design structure

Features	Possible definitions	Implications
Strata	Regions Community type (urban vs. rural)	May reduce standard errors of estimates. Control distribution of sample may lead to disproportionate sampling
First-stage sampling units	Census enumeration areas or similar geographic areas. Villages in rural strata	Facilitate clustering of the sample to control costs. Facilitate development of complete frames of housing unit addresses only in sampled areas. Selected with probability proportional to size.
Second-stage sampling units	Housing unit addresses	May contain none, one, or more than one household or unrelated person. Selected with equal probability within first-stage sampling units.
Third-stage sampling units (when not all household members are automatically included in the sample)	Household members	Sample selected from roster of household members obtained from a responsible adult household member. May lead to unequal weighting in order to account for household size.
Observational units	Households Household members Agricultural or business enterprises operated by the household members. Special files for subgroups, e.g., adults in the work force Events or episodes pertaining to household members Repeated measures over time (panel surveys)	May require more than one analytic file for special purpose analyses.

3. This next section of this chapter outlines the development of weights for use in survey analysis and the use of weights for the production of simple "descriptive" estimates, such as the

totals, means and proportions/percentages that are widely presented in survey reports. It also provides an overview of variance estimation for such estimates based on complex sample designs.

4. The rest of the chapter then focuses on three forms of "analytic" uses of survey data that explore the way in which a key response or dependent variable - e.g., academic performance of a school-going child, poverty level of a household - is affected by a number of factors, often referred to as explanatory variables, or regressor variables. Multiple linear regression models are suitable when the key response is a quantitative measurement variable, while logistic regression models are applicable when the key response variable is binary, i.e., the response takes only two possible values (e.g., yes/no, present/absent). These regression methods may be applied to a non-nested body of survey data, or to sampling units at a single level of the hierarchy of a multi-stage design. Alternatively, the analysis may need to take account of the different sources of variability occurring at the different hierarchical levels, and then multi-level modeling comes into play. This approach takes account of the correlation structure between sampling units at one level because they occur within units at different levels.

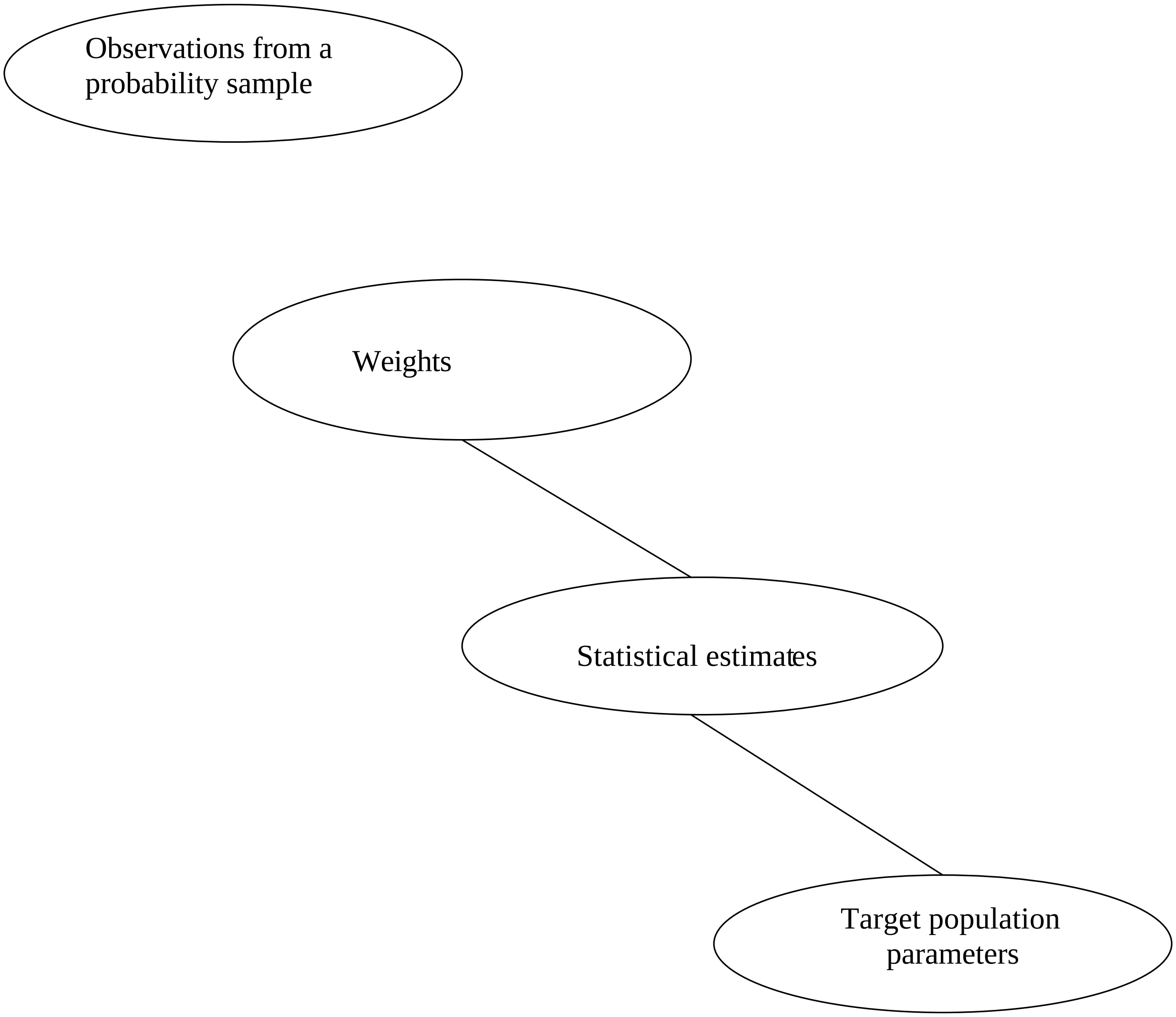
II. Descriptive statistics: weights and variance estimation

5. Household surveys are commonly designed to produce estimates of population totals, population means, or simple ratios of totals or means. Examples of totals might be total population, men in the work force, women in the work force, or the number of children five years old or younger. Examples of means might be average income for persons in the work force, average income of women in the work force, and average income of men in the work force. Ratio estimates might be required to estimate the proportion of households with total income below the poverty level or the average household income for households whose principal wage earner is a female.

6. Household surveys produce national estimates, but may also be designed to yield estimates for geo-political regions or for other cross-sectional domains. Furthermore, household surveys may be repeated to obtain periodic estimates (e.g., annual or five-year estimates); these might be viewed as temporal domains. As long as the statistics produced consist simply of estimates of totals, means, or rates even when produced for population domains (cross-sectional or temporal), we characterize the analysis required to produce these estimates as "descriptive". Descriptive statistics include the estimates themselves as well as some measure of the precision of those estimates. Descriptive reports may include standard errors of estimates or interval estimates based on those standard errors. Estimation of the standard errors requires an analysis that takes account of the household survey sample design. Interval estimates require not only the appropriate design-based estimates of standard errors, but also require knowledge of the degrees of freedom used in computation of the standard error estimates. These types of fairly simple descriptive statistics constitute the majority of the official statistics published to describe the results of household surveys.

7. Survey weights[1] and statistical estimation based on those weights provide the link between the observations from a probability sample of households and summary measures or population parameters about the household population. Figure 1 illustrates the link. The population of all households is sometimes called the target population or the universe. Without the application of both probability sampling and weighting, there is no supporting statistical theory to provide a link between the sample observations and the target population parameters.

Figure 1. Application of weights and statistical estimation


Observations from a probability sample
Weights
Statistical estimates
Target population parameters

[1] Design-based weights are generally developed as the inverse of the selection probability for selected observational units. The survey weights provided on analysis files for household surveys are usually design-based weights that have been adjusted for non-response. Often additional adjustments are applied to achieve post-stratification or calibration to agree with known, or much more precise, marginal totals. In addition, some form of weight trimming may be applied to limit the unequal weighting effect when large weights are due to unforeseen sampling or field data collection events. The term "survey weights" is used to differentiate them from strict "design-based weights".

•

8. Any analysis that ignores the sample design and the weights must be based on assumptions. If the sample is designed to generate equal probability sample, then the weights for estimating means, rates, or relationships among variables may be safely ignored. Kish (1965, pp. 20-21) called these designs *epsem* designs and noted that even complex multi-stage samples can be designed to be *epsem* for sampling units at the final or near final stage of the design. As noted later, adjustments for non-response may create unequal weights even if the design was initially *epsem*. If post-stratification or multi-dimensional calibration is applied to the data through adjustments to the weights, these processes will almost always create unequal weights adjustments and, therefore, unequal weights.

9. Some analysts are, however, willing to make the assumptions that would allow analysis of household survey data without weights or with equal weights. These assumptions are most tenable when applying models to the data to study relationships between a dependent variable and a number of independent explanatory variables.

10. For the theoretical case of surveys with complete response from all sample members, the use of design-based weights computed as the inverse of each observational unit's probability of selection provides for unbiased estimates of population totals and other linear statistics (Horvitz and Thompson, 1952). In practice, household surveys always encounter some non-response, which can lead to bias in estimates if these observations are dropped from the analysis without taking any other action (see chapter 8). Techniques have been developed which attempt to reduce the bias due to non-response. The simplest approach involves partitioning the sample into weighting classes so that within these classes the differences between the population parameters for respondents and non-respondents are believed to be much smaller or to be ignorable (Rubin 1987). Ratio adjustments to the weights are then performed within the weighting classes so that each class is represented in the adjusted estimates in the same proportion as it would have been represented in the selected sample.

11. The process of probability sampling does not necessarily guarantee that the selected sample's distribution on known characteristics will be identical to that of the total population. Stratification before sample selection can ensure this condition to hold for some characteristics, but may not be possible for others if the classification variable is not available on the frame used to select the sample. Rather than conducting complex ratio adjustments for each estimate produced from the household survey data, post-stratification is often incorporated as a one-time weight adjustment, which then automatically applies to all estimates produced using the adjusted weights. The simplest approach to post-stratification adjustment uses a partitioning of the sample similar to that used for weighting class non-response adjustment.

12. Calibration methods which control the weighted sample distribution in several dimensions simultaneously are sometimes used for weight adjustment for non-response, for post-stratification, or for both (Deville and Särndal, 1992; Folsom and Singh, 2000).

13. Extremely large weights can inflate the variance of household survey estimates through a design effect (see chapters 6 and 7). Sometimes these weights are arbitrarily reduced or trimmed, particularly if the large weight is not a result of the planned sample design.

14. The final weights attached to an analytic file produced from a household survey may contain the following factors:

- The design-based weight computed as the reciprocal of the overall probability of selection;
- A non-response adjustment factor;
- A post-stratification adjustment factor;
- A weight-trimming factor.

15. These factors should be documented so that any analyst can review them. The adjustment factors applied to the initial design-based weights involve some subjective and sometimes arbitrary judgments in the definition of weighting classes, the selection of control totals for post-stratification adjustment, and in the extent of weight trimming applied to control the design effect. When unexpected results or apparent anomalies occur in the survey estimates, it is not uncommon to thoroughly review the weighting process as well as all other aspects of the total survey design and implementation.

16. In general the analytic uses of household survey data provides special challenges due to complex survey designs which include the use of weights and a design structure. Design effects due to complex survey design are discussed in several of the chapters in this handbook. Chapter 20, in particular, addresses the impacts of complex survey design on the results of analysis. For more thorough discussions of complex survey analysis or for more detail on selected topics, the reader may to wish refer to texts that focus on these topics (Skinner, Holt and Smith, 1989; Korn and Graubard, 1999; and Chambers and Skinner, 2003). Chapter 20 of this handbook also provides more technical discussion regarding the analysis of complex surveys, and chapter 21 discusses software and provides examples of approaches to analyzing survey data with real data examples.

17. <u>Non-linear statistics</u>**.** Even simple statistics such as means become non-linear in complex surveys. To estimate a population mean from a complex survey, it is necessary to estimate a population total for the variable of interest, say family income, and to estimate the size of the population, say total number of families. The mean is then estimated as the ratio of the two estimates. Mean family income would be estimated as:

$$Estimate\,of\ mean\ family\,income = \frac{Estimate\,of\ total\ family\,income}{Estimate\,of\ total\,number\,of\ families}$$

This estimated mean turns out to be a non-linear function (a ratio) of two linear statistics. In complex surveys, the sample size (number of observations of a particular type) is itself a random variable. These types of non-linear estimates are not unbiased for small samples, but are consistent in the trivial sense that if the sample size were increased to the finite population size, the non-linear estimate would exactly equal the comparable finite population value (Cochran, 1977, pp. 21, 153, 190). If we allow ourselves to consider the finite population as arising from a hypothetical infinite population, then we can consider letting the sample size increase without limit. In this case we can claim model-consistency when the non-linear estimate converges in probability to the super-population parameter as the sample size increases (e.g., Skinner, Holt, and Smith, 1989, pp. 17-18).

18. Standard errors of non-linear statistics can only be expressed approximately using first-order Taylor series approximations. Estimates of the standard errors of non-linear statistics can be obtained using the first-order Taylor series approximations or replication methods such as balanced repeated replication or jackknife replication.

19. The same types of arguments carry over to analysis using "linear" models when the required linear functions of both the dependent and independent are first estimated at the full population level.

20. In summary, the use of weights leads to unbiased linear estimates and consistent non-linear estimates. In practice, the use of consistent estimates is considered satisfactory for controlling estimation bias. Other types of biases and non-sampling errors such as those arising due to non-response, to interviewer error, or to respondent error are usually of much more practical significance, particularly when sample sizes get large.

21. <u>Sample design structure in household surveys</u>**.** In general, both the population and the sample design can have some structure. In household survey sample designs, a nested structure is generally imposed on the sampling frame as was discussed in the preceding section and illustrated in Table 1. While the structure does not influence the construction of first-order statistical estimates such totals, means, ratios, or model coefficients, it does affect second-order statistics (variance estimates) that allow analysts to estimate the standard errors of the first-order statistics and to construct tests of statistical significance concerning specified hypotheses.

22. The full expression for variance of estimates based on stratified multi-stage samples has components for each stage of the sample design. For example if stratification is employed at the first stage only, an estimate $\hat{T}$ of a population total T based on a three-stage design with area segments, households, and household members might have a variance of the form:

$$Var(\hat{T}) = \sum_{h} \left(fpc_{h1} \frac{S_{h1}^{2}}{n_{h1}} + fpc_{h2} \frac{S_{h2}^{2}}{n_{h2}} + fpc_{h3} \frac{S_{h3}^{2}}{n_{h3}} \right)$$

where the terms within stratum h are defined as follows. The fpc_{hi} terms are finite population correction factors at the area segment selection ($i=1$), housing unit selection ($i=2$), and persons selection stages ($i=3$), respectively. The S_{hi}^{2} terms are variance components based on the weighted data at the three stages of sampling. The n_{hi}'s are the sample sizes of segments ($i=1$), households ($i=2$), and persons ($i=3$) within stratum h. In practice, it is not unusual for some of these variance components to be difficult to estimate or not to be estimable; this can occur due to sub-samples of size 1 or for other reasons. Cochran (1977, p. 279) notes that if the finite population correction factor at the first stage can be ignored (assumed to be 1), then estimates of the variance can be based on a much simpler analog to this formula that involves only the first stage of sampling. The assumption of a first-stage finite population factor of 1 is often described as "with replacement" sample design variance estimate to approximate the variance for a "without replacement" sample design.

23. To make this work for linear estimates of population totals for the three-stage design above when the observational units are persons, we can define a new variable:

$$Z_{hi} = n_{h1} \sum_j \sum_k w_{hijk} Y_{hijk}$$

where w_{hijk} and Y_{hijk} are the weight and the observed variable for person *k* of household *j* of area segment *i* within stratum *h*. Then, a reasonable estimate of the variance can be obtained as:

$$\text{var}(\hat{T}) = \sum_h \frac{\sum_i (Z_{hi} - \overline{Z}_h)^2}{n_{h1}(n_{h1} - 1)}$$

This works because with this formulation because the estimate of the population total can be written as:

$$\hat{T} = \sum_h \overline{Z}_h$$

With appropriate choice of Z_{hi}, the variances of non-linear as well as linear statistics can be estimated using first-order Taylor series approximations.[2] This extends to the parameter estimates in regression or logistic regression. Note that variance contributions from subsequent stages need not be estimable for this to work.

24. If the first-stage finite population correction is appreciably less than 1, this formulation will overestimate the variance and lead to overstating the standard error of survey estimates. A small overestimate would lead to conservatively wide confidence intervals or may lead to fewer declarations of statistical significance when conducting hypothesis tests. In that sense, the assumption of a first-stage finite population correction of 1 is said to be conservative statistically since it will help protect against false declarations of statistical significance. It should be noted that the application both Taylor series-based and replication-based software is simplified by the assumption of a finite population correction factor of 1 at the first sampling stage (see chapter 21 of this publication).

III. Analytic statistics

[2] Woodruff (1971) shows how linearized variables can be developed to facilitate the computation of complex Taylor series variance approximations.

25. In this section we move from consideration of simple descriptive estimates to what are termed "analytic statistics", that is statistics that examine the relationships among variables. In fact, the moment data users wish to compare estimates among domains, the nature of the required statistics becomes "analytic". Simple analytic statistics may be based on differences among domains, e.g., a comparison of the proportion of households with total income below the poverty level in two geo-political subdivisions or a comparison of crop production over the last two years. Sometimes the estimates in a simple comparison are independent of one another so that the standard error of the difference can be determined strictly from the standard error of the individual estimates. Under these circumstances, the standard error of the estimated difference between two domain means can be derived as:

$$se(\bar{y}_1 - \bar{y}_2) = \sqrt{\{se(\bar{y}_1)\}^2 + \{se(\bar{y}_2)\}^2}$$

This formula for the standard error of a difference assumes that the two estimates are independent and, as a result, their estimates are uncorrelated. This form of the standard error of differences is convenient for data users, because they can derive the standard error of a difference from published standard errors of the individual estimates. However, with complex sample designs, domain estimates are often correlated. The variance of the difference of two domain estimates then includes a covariance term:

$$se(\bar{y}_1 - \bar{y}_2) = \sqrt{\{se(\bar{y}_1)\}^2 + \{se(\bar{y}_2)\}^2 - 2\operatorname{cov}(\bar{y}_1, \bar{y}_2)}$$

26. The covariance term is generally positive, and hence it leads to a lower standard error of the difference estimate than the independent case discussed above. Household surveys can be designed to take advantage of the covariance term in the standard errors of estimates of differences; longitudinal panel surveys achieve a high positive covariance among annual estimates by utilizing a common, continuing sample of individuals or households. Because the standard error of the difference cannot be derived from the published standard errors of the individual estimates, it becomes necessary to anticipate what comparisons are of greatest interest and to publish their standard errors also.

27. For strictly descriptive statistics about finite populations, the standard error of descriptive estimates is correctly reduced by the application of a finite population correction factor. In the simplest case of simple random sampling, the finite population correction factor is:

$$fpc = 1 - \frac{n}{N}$$

where *n* is the sample size and *N* is the population size. If the purpose of the analysis is analytic, then, even in the simplest case of statistical significance of the observed difference between two domain means, the use of the finite population correction factor is inappropriate (Cochran, 1977, pp. 34-35). This is because the form of the statistical significance test requires one to hypothesize whether both domain populations could have arisen from a common infinite hypothetical population (a single super-population).[3] The use of finite population correction factors in a structured complex design is discussed later.

IV. General comments about regression modeling

28. The methods covered in the rest of this chapter involve a modeling technique which models the variation in a key response variable or dependent variable, and identifies which subset of a set of potential explanatory variables contributes most significantly to this variation. Choice of this "best" subset can be made by the application of appropriate variable selection procedures, or by using a sensible sequential procedure to explore a number of different models with close attention to the suitability, from a practical viewpoint, of the variables that enter or are removed from the model at each step of the analytical procedure.

29. We would like to stress that the techniques discussed in this chapter should be considered as being supplementary to, rather than in place of, the simpler methods of analysis. Initial exploration of the data using simple descriptive summaries (means, standard deviations, etc.), graphical procedures (scatter plots, bar charts, box plots, etc.) and relevant data tabulations is very valuable and should form the first stage of the data analysis. Sometimes this may be all that is needed. Often however, the survey objectives demand further analysis of the data, and then modeling techniques are likely to become important.

30. The modeling methods discussed here are particularly relevant where the approach is holistic, for example, when the analytic objective is to understand the rationale of existing farming systems and the way in which households manage their limited resources to meet both production and consumption needs. The emphasis throughout is on practical application of an appropriate modeling technique, with an appreciation of possible difficulties faced in developing country field situations. Analysis limitations are highlighted to ensure that the approaches discussed are applied only after careful thought is given to the appropriateness of the method being applied for the research setting in mind.

31. Regression models are used to develop a better understanding of the relationship between a dependent variable and a set of independent or explanatory variables. It is usually impossible to assign "cause" to any observed relationships between a dependent variable and an explanatory variable, except in the case of well-designed controlled and randomized experiments.[4] With this caution in mind, a great deal can be learned from applying regression models to the observed data obtained from household surveys.

32. As opposed to controlled experiments which employ randomization and control of auxiliary variables, household survey data are usually observational with little or no control over

[3] Cochran (1977, p. 39) states that use of the finite population correction factor is not appropriate for statistically testing for differences among domain means. The interpretation of this guideline becomes more ambiguous when applied to complex designs involving both stratification and clustering; Chromy (1998) discusses the problem with regard to sampling of students within schools when schools are stratified and sampled at high rates. Graubard and Korn (2002) provide a recent review of this issue.

[4] Randomized experiments can be embedded in surveys. Often, these are methodological experiments in a pretest sample or supplemental samples for an ongoing survey. Social experiments can also be conducted by recruiting subjects for a social experiment using a household survey sample.

other factors that may influence the relationships among variables. Regression methods can sometimes remove the effects of these uncontrolled confounding variables, so that less biased estimates of the true relationship can be obtained.

33. Regression modeling is often exploratory in nature. A number of different models may be developed to explain the behavior of a dependent variable of interest. The explanatory variables used in the model are restricted to those that are available on the survey data file; as a result the variables selected to explain the variation in a dependent variable may only be strong correlates of the actual causative factor. There may be competing correlates of the causative factor none of which logically seem to be related to the dependent variable. Analysts of household surveys should be guided by the substantive (e.g., social or economic) theory in choosing explanatory variables and in determining the form of the relationship (e.g., linear vs. non-linear).

34. When substantive theory does not suggest strong theoretical relationships or when several competing explanatory variables may be suggested by the substantive theory, variable selection approaches from standard (non-survey) packages can be applied to identify potential explanatory variables. Forward and backward variable selection approaches are available in many non-survey software packages that help identify explanatory variables that have linear relationships to the dependent variable. If the non-survey package allows, the use of survey weights even for this exploratory analysis is highly recommended. Survey weights may be normalized to sum the total sample to provide better estimates of error and more nearly correct tests of statistical significance (see chapter 21 for examples of this approach). After using non-survey statistical packages or programs to perform variable selection, it is a good practice to evaluate the model using a software package that uses the survey weights and recognizes the household survey design.

35. Model variables may be categorical, counts, or continuous measurement variables. Linear regression models are used when the dependent variables are counts or continuous measurements; logarithmic transformations are advocated for count data. When the dependent count variable includes values of zero, the logarithmic transformation fails, but procedures such as the PROC LOGLINK (SUDAAN 2001) can be used to fit the expected value of the logarithm of a count variable. Logistic regression is used when the dependent variable is a categorical variable defined at two levels; multinomial regression models may also be applied to categorical dependent variables with more than two levels. For discussion purposes, we classify explanatory variables as categorical or continuous, because count and continuous (measurement) variables are treated essentially the same way in a modeling context. Survey data may also be analyzed using survival models and other multivariate techniques not discussed in this chapter.

36. The use of categorical explanatory variables, which define study domains, is analogous to constructing simple domain comparisons without using models. The use of models allows the analyst to simultaneously adjust for other possible explanatory variables. This is often called adjusting for covariates. When there is no adjustment for covariates, regression model coefficients reproduce simple domain comparisons and estimate the domain differences that exist in the population. When other variables are included in the model as covariates, the regression model coefficients estimate the domain differences that would hypothetically exist if the covariates were held at the same levels in all domains.

37. Regression model coefficients for continuous explanatory variables can also be obtained with or without adjustment for other covariates. Decisions about adjusting or not adjusting for covariates should be guided by the purpose of the analysis. Unadjusted estimates describe an empirical relationship between dependent and explanatory variables as they exist in the population. Adjusted estimates describe the same relationship if other variables were hypothetically held constant. If the other variables included in the model are also good predictors of the dependent variable, they can improve the precision of the predicted values for set levels of the key predictors under study. Choice of methods of analysis should depend on the purpose of the analysis.

38. Only simple models for continuous explanatory variables are discussed in the examples below. When the explanatory variables are continuous, the analyst should investigate the relationship of the dependent variable with potential explanatory variables. Simple plots can show that a linear relationship is inadequate to properly relate variables. Depending on the observed plots, additional terms (quadratic or cubic terms) can be added to better capture the relationship. The dependent variable can then have linear relationships with an explanatory variable, with its square, and with its cubic or higher terms. Residual plots after having included some of the potential explanatory variables can be used to determine whether other variables or higher order (squared or cubic terms) of included variables may be influencing the model fit. For explanatory variables with a wide range of values and differing effects on the dependent variable over that range, spline models that allow the relationship to change over subsets of the range are often useful. The effects of age when the survey sample includes youth, middle-aged, and elderly persons can often benefit from the use of regression spline models.

39. Other diagnostic procedures include the examination of the goodness of fit of proposed models and the examination of the statistical significance of regression parameters for added variables. Procedures from standard (non-survey) procedures can be adapted to weighted survey data. The concept of explained variation can be used with weighted survey data and linear regression. Contingency table approaches can be used to evaluate the fit of logistic regression models. Korn and Graubard (1999, chapter 3) provide a good discussion of the adaptation of diagnostic procedures to general survey data analysis.

40. The development of regression models based on the observed data clearly involves the concept of exploratory data analysis (Tukey, 1977). This type of analysis can lead to useful insights about the data and the relationship among observed variables, but the statistical significance of findings from such “unplanned” analysis should remain a topic for future confirmation or for validation by the study of other survey data.

V. Linear regression models

41. For the purposes of discussing linear and logistic regression models (sections V and VI), it is convenient to assume that sampling is “with replacement” at the first stage. We further assume that the analytic file of observation data includes index variables for strata, designated by *h*, and for primary sampling units (PSUs), designated by *i*. Additional structure variables do not need to be identified when we are willing to use the “with replacement” design assumption at the

first stage of sample selection as discussed in section II above. The full implications of using a complex household sample design are incorporated into the estimates of model coefficients and their standard errors only if we use a statistical package that properly accounts for the household survey design including the analytic weights and the design structure (strata and PSUs). When we discuss multi-level models, the focus will change to incorporating the design structure into the model and the analysis will permit estimation of effects related to the structure variables.

42. A linear regression model that involves one continuous explanatory variable and one categorical explanatory variable can be expressed as:

Model 1

$$y_{hij} = ax_0 + b_1 x_{1hij} + \sum_{d=1}^{D} g_d x_{2dhij} + e_{hij}$$

43. In model 1, observations are represented by the observed dependent variable, y_{hij}; an intercept variable, x_o, always set to 1; an observed continuous explanatory variable, x_{1hij}; and a set of indicator variables, x_{2dhij}, defining D levels of a categorical variable. The regression model parameters $a, b_1, and\ g_d\ (d = 1,2,...,D)$ are termed regression coefficients and are estimated by the analysis. The final term in the model is the error term and measures the deviation from the model associated with the j-th observation associated with the i-th PSU of the h-th stratum. This is a main effects model, since it contains no interaction effects.

44. Depending on the software being applied, the set of indicator variables can be specified as a single variable in a model statement; it may be necessary to define the variable as categorical and specify the number of levels with program statements or commands. The program then defines a vector of indicator variables. An indicator variable, say x_{2dhij} is set to be 1 if observation hij belongs to category d, and to be 0, otherwise. To avoid linear dependence among the explanatory variables, the analysis program re-parameterizes the indicators for the categorical variable. This is typically done by dropping the final category of the categorical variable; this category then becomes the reference category.[5] Table 2 shows some of the effects that can be estimated for model 1 when the dependent variable is household income from wages, the continuous explanatory variable is number of wage earners in the household, and the categorical variable defines 4 regional domains (north, south, east and west) of the country.

[5] It is also possible to estimate the coefficients of categorical variables by adding a linear constraint such as requiring that the sum of the effects is zero or that sum of the weighted effects is zero.

Table 2. Interpreting linear regression parameter estimates when the dependent variable is household earnings from wages for model 1

Effect (as usually identified in program output)	Coefficient of:	Estimate of:	Interpretation
Intercept	$x_0 = 1$	a	Salaried household income at reference cell or zero levels: 0 wage earners in West
Wage earners in household	x_{1hij}	b_1	Change in household salaried income per additional wage earner (adjusted for region)
Region:			Regional differences in household earnings from wages (adjusted for wage earners in household):
North (*d=1*)	$x_{21hij} - x_{24hij}$	$b_2 = g_1 - g_4$	North vs. West
South (*d=2*)	$x_{22hij} - x_{24hij}$	$b_3 = g_2 - g_4$	South vs. West
East (*d=3*)	$x_{23hij} - x_{24hij}$	$b_4 = g_3 - g_4$	East vs. West
West (reference domain, *d=4*)	$x_{24hij} - x_{24hij} = 0$	$g_4 - g_4 = 0$	No estimate

45. The estimated regression coefficients for the domain variables are defined with regard to the difference between a domain and the reference domain. The statistical significance test of an estimated coefficient for the domain north actually tests whether north and west could be random samples from the same common super-population (See Figure 2). If the coefficient for north is statistically significant from 0, the analyst can conclude that it is highly unlikely (5 per cent chance or less) that household wages for north and west are samples from the same super-population after adjusting for number of wage earners in the household. Statistical programs allow the users to specify different reference sets either by ordering the categories (so that the desired reference category is last) or by explicit specification. This can be a useful device in obtaining meaningful regression parameter estimates. Other comparisons can also be estimated through functions of the estimated coefficients.

46. Table 3 shows some estimable model 1 functions based on estimates of the parameters shown in Table 2. Table 3 shows model 1 estimates of household income from wages by region and number of wage earners in the household. This could easily be extended to 3 or more wage earners per household.

Table 3. Estimable household incomes from wages (model 1)

Region	For households with:	
	1 Wage Earner	2 Wage Earners
North	$\hat{a} + \hat{b}_1 + \hat{b}_2$	$\hat{a} + 2\hat{b}_1 + \hat{b}_2$
South	$\hat{a} + \hat{b}_1 + \hat{b}_3$	$\hat{a} + 2\hat{b}_1 + \hat{b}_3$
East	$\hat{a} + \hat{b}_1 + \hat{b}_4$	$\hat{a} + 2\hat{b}_1 + \hat{b}_4$
West	$\hat{a} + \hat{b}_1$	$\hat{a} + 2\hat{b}_1$

47. Let's examine the assumptions that the analyst must make in using model 1 for studying household earnings from wages. Perhaps the most critical assumption is that household earnings from wages are linearly related to number of wage earners. The linearity assumption states the change in household earnings from wages increases by the same amount when increasing from 0 to 1 wage earners, from 1 to 2 wage earners, from 2 to 3 wage earners, etc. This assumption appears doubtful. Since categorical variables require fewer assumptions about the form of the relationship between the explanatory variable and the dependent variable, the analyst might decide to convert the number of wage earners into a categorical variable and thus use a model with only categorical variables.[6] A variant of model 1 could be written as:

Model 2

$$y_{hij} = ax_0 + \sum_{d=1}^{D_1} g_{1d} x_{1dhij} + \sum_{d=1}^{D_2} g_{2d} x_{2dhij} + e_{hij}$$

48. For model 2, the analyst might define as few as 2 wage earner categories or a much larger number depending on the distribution of the number wage earners in the households. To limit the number of parameters to estimate, the analyst may settle on 4 categories:

- Category 1: no wage earners;
- Category 2: one wage earner;
- Category 3: two wage earners;
- Category 4: three or more wage earners.

49. This model is still a main effects model, but number of regression parameters has now increased from 5 to 7. Table 4 shows the interpretation of estimated regression coefficients under model 2. This model no longer requires the analyst to assume a linear relationship of household wage earnings to number of wage earners in the household. However, since there are no interaction terms in the model, the model does assume the following:

[6] For additional discussion of methodology for assessing the goodness of fit of a linear regression model and for some other alternatives for non-linear relationships, readers may refer to Korn and Graubard (1999, pp. 95-100).

- The “wage earners in household” effect is the same in all four regions;
- The “region effect” is the same all for levels of “wage earners in household.”

Table 4. Interpreting linear regression parameter estimates when the dependent variable is household earnings from wages for model 2

Effect (as usually identified in program output)	Coefficient of:	Estimate of:	Interpretation
Intercept	$x_0 = 1$	a	Household earnings from wages at the reference levels (No wage earners and West)
Wage earners in household			Change in household earnings from wages income per additional wage earner (adjusted for region):
One (d=1)	$x_{11hij} - x_{14hij}$	$b_1 = g_{11} - g_{14}$	One vs. none
Two (d=2)	$x_{12hij} - x_{14hij}$	$b_2 = g_{12} - g_{14}$	Two vs. none
Three or more (d=3)	$x_{13hij} - x_{14hij}$	$b_3 = g_{13} - g_{14}$	Three vs. none
None (reference domain, d=4)	$x_{14hij} - x_{14hij} = 0$	$g_{14} - g_{14} = 0$	No Estimate
Region:			Regional differences in household earnings from wages (adjusted for number of wage earners in household):
North (*d=1*)	$x_{21hij} - x_{24hij}$	$b_4 = g_{21} - g_{24}$	North vs. West
South (*d=2*)	$x_{22hij} - x_{24hij}$	$b_5 = g_{22} - g_{24}$	South vs. West
East (*d=3*)	$x_{23hij} - x_{24hij}$	$b_6 = g_{23} - g_{24}$	East vs. West
West (reference domain, *d=4*)	$x_{24hij} - x_{24hij} = 0$	$g_{24} - g_{24} = 0$	No estimate

50. Most regression packages will allow you to test for interactions among categorical variables. In this case there will be nine degrees of freedom for interaction. While interpreting the effects of regression models with two categorical main effects and an interaction is possible, we would recommend a different approach. First test for interaction; in this case model 2 could be augmented to include interaction between “wage earners in household” and “region”. If the statistical test for interaction indicates that interactions are present, incorporate the full model

with 16 estimable parameters by implementing a simpler model with a single categorical variable defined at 16 levels. Call this model 3 and write it as:

Model 3

$$y_{hij} = a\,x_0 + \sum_{d=1}^{16} b_{1d}\,x_{1dhij} + e_{hij}$$

51. The sixteen levels of the new categorical variable and their estimates (in parenthesis) are

- North, one wage earner ($\hat{a} + \hat{b}_1$);
- North, two wage earners ($\hat{a} + \hat{b}_2$);
- North, three or more wage earners ($\hat{a} + \hat{b}_3$);
- North, no wage earners ($\hat{a} + \hat{b}_4$);
- South, one wage earner ($\hat{a} + \hat{b}_5$);
- South, two wage earners ($\hat{a} + \hat{b}_6$);
- South, three or more wage earners ($\hat{a} + \hat{b}_7$);
- South, no wage earners ($\hat{a} + \hat{b}_8$);
- East, one wage earner ($\hat{a} + \hat{b}_9$);
- East, two wage earners ($\hat{a} + \hat{b}_{10}$);
- East, three or more wage earners ($\hat{a} + \hat{b}_{11}$);
- East, no wage earners ($\hat{a} + \hat{b}_{12}$);
- West, one wage earner ($\hat{a} + \hat{b}_{13}$);
- West, two wage earners ($\hat{a} + \hat{b}_{14}$);
- West, three or more wage earners ($\hat{a} + \hat{b}_{15}$);
- West, no wage earners ($\hat{a}$).

52. With the 16-th category defined as the reference cell, the model 3 intercept estimate $\hat{a}$ corresponds to the estimated household earnings from wages for that cell (west, no wage earners). The estimate of household earnings from wages for each of the other 15 cells is estimated as the 16-th cell estimate plus the estimated regression coefficient for that cell. These 16 estimates could also be obtained from direct estimates. If the survey weights and the design structure are applied in appropriate survey software, the estimates and their estimated standard errors should be identical under the two approaches (model 3 or direct estimation). There is no gain in applying model 3 over developing 16 direct estimates.

53. If the sample sizes for some of the 16 cells are small, the precision of the estimates for these "small sample" cells will be poor. Using a main effects model (model 1 or 2) produces more precise estimates for the cells with small sample sizes by "borrowing" sample size from the marginal estimates and making a few more assumptions (as discussed above) about how the finite population derives from the hypothetical super-population (Figure 2).

54. Analysts generally use models to adjust for a number of explanatory variables. Suppose that an analyst wishes to adjust for city or community characteristics such as urbanicity (per cent urban). The analysis may show that the region effect is reduced after taking account of, and standardizing for, per cent urban. In a main effects linear model, adjusting for per cent urban (as either a continuous or categorical explanatory variable) provides estimates of region effects assuming the same (standard) per cent urban distribution within each region. Without adjustment for covariates, the model (or direct estimates) represents regional parameters as they exist; with a model adjustment for covariates, the model represents regional parameters as they would be if the covariate effects were removed. Korn and Graubard (1999, pp. 126-140) discuss the use of predictive margins as a method of standardization.

VI. Logistic regression models

55. When the dependent variable is categorical, linear regression approaches do not apply. Although multinomial modeling procedures are available, we will be discussing only the binary (two-level) categorical variables that can be analyzed using logistic regression models. In this sense, logistic regression is a special simpler case of multinomial regression.

56. For a two-category or binary dependent variable coded as 0 and 1, linear regression approaches will work but can produce predicted values outside the range of 0 to 1. Linear regression might be used as a preliminary step with a binary dependent variable to identify explanatory variables that are good predictors of the dependent variable, particularly if the software packages available to the analyst have variable selection procedures built into the linear regression software but not into the logistic regression software.

57. Numerical methods are used to fit the parameters of logistic regression models; therefore, they may sometimes have difficulty in converging to a solution. Users should be alert any warnings given by the software when problems occur with convergence; generally, these cases can be resolved by simplifying the model.

58. A logistic regression model that involves one continuous explanatory variable and one categorical explanatory variable can be expressed as:

Model 4

$$\log\left(\frac{p(\underline{x}_{hij})}{1-p(\underline{x}_{hij})}\right) = ax_0 + b_1 x_{1hij} + \sum_{d=1}^{D} g_d x_{2dhij} + e_{hij}$$

59. Except for the dependent variable, the terms in model 4 are defined the same way as in model 1. To understand the logistic transformation, consider an example where $p(\underline{x}_{hij})$ is a function of the explanatory variables; designate it by *p* for convenience. Further assume that *p* is the probability that a household with a given set of values for the explanatory variables has an income level below the established poverty level. Then *p/(1-p)* is called the odds of being in poverty, and *log(p/1-p))* is the log odds of *p*, sometimes called *logit(p)*. Model 4 tries to relate the log odds of *p* to the *x*'s. The observations are single households where we observe not the probability of being in poverty, but the actual current status: in poverty or not in poverty. Also since the dependent variable is a log odds of *p*, each parameter $[a, b_1, and\, g_d\, (d=1,...,D)]$ is also on the log odds of *p* scale; furthermore, the relationship between the log odds of *p* and the *x*'s is assumed to be linear (compare to model 3 above).

60. Re-parameterization of categorical explanatory variables and the definition of reference categories is the same as for linear regression discussed above. Regression model parameters in the output of the logistic regression program look like those for linear regression, but they have different interpretations. Table 6 summarizes the interpretation of the usual parameter estimates for model 4. Note that there are five estimated parameters (an intercept, a , and four b 's).

Table 6. Interpreting logistic regression parameter estimates when the dependent variable is an indicator for households below the poverty level for model 4

Effect (as usually identified in program output)	Coefficient of:	Estimate of:	Interpretation
Intercept	$x_0 = 1$	a	The log odds of being in poverty at reference cell or zero levels: 0 wage earners in West
Wage earners in household	x_{1hij}	b_1	Change in log odds of the of being in poverty per additional wage earner (adjusted for region)
Region:			Regional differences in the log odds of being in poverty (adjusted for wage earners in household):

North (*d=1*)	$x_{21hij} - x_{24hij}$	$b_2 = g_1 - g_4$	North vs. West
South (*d=2*)	$x_{22hij} - x_{24hij}$	$b_3 = g_2 - g_4$	South vs. West
East (*d=3*)	$x_{23hij} - x_{24hij}$	$b_4 = g_3 - g_4$	East vs. West
West (reference domain, *d=4*)	$x_{24hij} - x_{24hij} = 0$	$g_4 - g_4 = 0$	No estimate

61. Note also that the logistic model parameters predict the log odds of being in poverty and do not directly predict the probability of being in poverty. Consider b_2 in Table 6. It is expressed as a difference in log odds:

$$b_2 = \log\left(\frac{p(North)}{1-p(North)}\right) - \log\left(\frac{p(West)}{1-p(West)}\right)$$

By the properties of logarithms, it can also be expressed as the log of an odds ratio:

$$b_2 = \log\left(\frac{\dfrac{p(North)}{1-p(North)}}{\dfrac{p(West)}{1-p(West)}}\right)$$

Standard output from logistic regression procedures routinely also provides the odds ratios, since they can be readily computed as:

$$e^{b_2} = \left(\frac{\dfrac{p(North)}{1-p(North)}}{\dfrac{p(West)}{1-p(West)}}\right)$$

In addition, individual household probabilities of being in poverty can be determined from the model as:

$$p(\underline{x}_{hij}) = \frac{1}{1+e^{-\log it[\ p(\underline{x}_{hij})}}$$

62. When citing the results of logistic model fitting, writers sometimes interpret an odds ratio of 2 as indicating that the probability of the event (poverty) in one domain (e.g., north) is twice the probability of the event (poverty) in the other domain (e.g., west). While this type of statement is approximately true for rare events (*p* near 0), it is far from true for more common events.

VII. Use of multi-level models

63. We now turn to a discussion concerning multi-level modeling, and begin by emphasizing the need to recognize the survey data structure. Of relevance here is the structure imposed by surveys that are designed to be multi-stage. For example agro-ecological regions in a country may form strata, and from each, a number of administrative units may be selected. The latter will form the primary sampling units. Secondary units are then selected from each primary unit; subsequent units are selected from the secondary units, and so on. This leads to a hierarchical data structure. It can involve the use of stratification variables at one or more of the levels.

64. For example, a survey concerning farming households in a region may involve using the administrative divisions of the region as primary units, then choosing villages from each division and then selecting households from each village, perhaps ensuring that different wealth categories of households are included. Here attention must be paid to the different sources of variability in the data collected at the household level. It incorporates variation between the administrative divisions, variation between villages, and variation between households within villages. Often data are also collected at each level of the hierarchy, here at the household level, at the village level and at the administrative division level**.** It is then important to recognize and note which variables are measured at the village level (e.g., existence of an extension officer; government subsidies for fertilizer), and which are measured at the household level (e.g., socio-economic characteristics of the household).

65. For data analysis purposes, separate ‘flat’ spreadsheet files may be prepared to hold the village level information and the household level information, using some key identifier to link these files. This is appropriate if the analysis objectives require data at village level to be analysed separately from data at the household level. It is not however suitable if the analysis needs to combine village information with household level information. A relational database is much more desirable, i.e., a database which allows data at different levels to be stored in one file, together with links that permit data at one level to be related to data at another level. The analysis must pull together the information from the multiple levels so that (for example) the inter-relationships between the different levels can be explored to enable an overall interpretation.

66. Multi-level modeling is the key statistical technique of relevance here. This modeling approach (Goldstein, 1995; Snijders and Bosker, 1999; Kreft and Leeuw, 1998) is desirable because it allows relationships across and within hierarchical levels of a multi-stage design to be explored, taking account of the variability at different levels. Inter-correlations between variables at the same level are also taken into account. It also provides, through use of appropriate software, e.g., *MlwiN* (Rashbash et al., 2001), SAS (2001), model based standard errors for estimates from complex survey designs. Such standard errors can serve as reasonable approximations for more exact standard errors that take account of stratification and clustering. It should be noted that *MlwiN* could also take account of sampling weights. This is important since unequal probabilities of selection in a multi-stage sampling design can induce bias in estimators of key parameters. Pfeffermann et al. (1998) and Korn and Graubard (2003) discuss these issues more thoroughly.

67. It is worth highlighting briefly at this point, the consequences of ignoring the hierarchical structure. This happens when the data are aggregated to a higher level or disaggregated to a

lower level. If the analysis is relevant and is required only at one level, there is no problem. However, care must then be taken that any inferences are made only at that level. It will not be possible to make inferences about one particular level of the hierarchy from data analysed at another level. Thus an analysis ignoring the hierarchy will not permit cross-level effects to be explored. Another difficulty arises if data is analysed at its lowest level by regarding the higher-level units as a factor in the analysis. This is inefficient because it does not allow conclusions to be generalized to all higher-level units in the population - they will only apply to the sampled units.

68. We present below an example to illustrate a scenario where the use of multi-level modeling can be beneficial in exploring relationships. Further examples can be found in Congdon (1998), Langford et al. (1998) and Goldstein et al. (1993).

Example 1

69. In a study of factors contributing to successful community-based co-management of coastal resources amongst Pacific Island countries, 31 sites across 5 countries were chosen and 133 interviews conducted with mini-focus groups comprising two to six respondents from different households (World Bank, 2000). The countries, Fiji, Palau, Samoa, Solomon Islands and Tonga, were chosen to represent a range of coastal management conditions. The 31 sites were selected to cover a range of conditions that were believed to influence management success. The study collected “perceptions of success” in terms of trends in perceived catch per unit effort (CPUE), condition of habitats, threats to the site, and an assessment of compliance. The first three indicators were measured on a 5-point scale (5 = improving a lot; 1 = declining a lot), while compliance was measured on a 4-point scale.

70. Data were also collected nationally from the fisheries and environmental ministries in each country, and at site level. Additionally, each focus group, comprising members of several households, was asked to give its perceptions for up to three resources (for CPUE), three habitats, three threats and five management rules for compliance. Thus the information collected during this study resided at four levels: country, site, focus group and specific resource, habitat, threat or rule.

71. It is important however to note that this survey used non-probability sampling; it may therefore be argued that any analytical conclusions may not be generalizable to any clearly defined target population. However, for the purpose of this discussion, suppose that sampling had been on a probability basis and that data at the focus group level are being analysed using a multi-level model - the particular variable of interest being the perception of CPUE trend, obtained by averaging the perception scores across the three resources. The country effect (at the top level of the hierarchy) can be included in the model as a factor (a fixed effect) since it is essentially a stratification variable. However, to enable results to be generalised across all co-managed sites, it is necessary to include sites as a random variable rather than as a fixed effect. Focus groups within sites would also enter the model as a random effect. Including a mixture of fixed effect variables and random effect variables is the essence of multi-level modeling. Such models also allow interactions among site level variables and variables at the focus group level to be explored.

72. To illustrate the way in which a multi-level model can be formulated to answer specific survey questions, we use an example from a Food Production and Security Survey conducted in Malawi in 2000-01 (Levy and Barahona, 2001). The survey was aimed at evaluating a programme aimed to increase food security among rural smallholders through the distribution of a starter pack containing fertilizer, and maize and legume seed.

Example 2

73. The Food Production and Security Survey was a national survey that used a stratified 2-stage sampling scheme with districts as the strata. Four villages were selected from each of Malawi's 27 districts, and about 30 households selected from each village. Selection of villages was limited to those with more than 40 households (to ensure that there were enough households in the village to interview recipients of the starter pack) and less than 250 households (to make the work possible for the team within the time allowed according to resource availability).[7] Within this restriction, the sampling at each stage was done at random. A total of 108 villages and 3,030 households were visited during the survey.

74. The data we consider for multi-level modeling comes from a household questionnaire completed during the survey. The subset of variables we will consider in our illustration are the district, village, household identification number, sex and age of household head, size of the household, whether or not the household had received a Starter Pack and two indices reflecting household assets[8] and income.[9]

75. There are several multi-level models that can be fitted to this data. In formulating the model, the first step is to decide which variables are random and which are fixed effects.

76. In Example 2, district is a stratification variable and would be regarded as a fixed effect. In general, any effect is regarded as fixed if repeats of the sampling process would result in the same set of selections. On the other hand, villages and households have been selected at random, so they form random effects in the model.

77. The basic model for analysing (say) the asset index (AI) is:

Model 5

$$y_{ijk} = \mu + d_k + U_{jk} + \varepsilon_{ijk},$$

where d_k is the district effect (k = 1, 2, …., 27), and indices *i* and *j* correspond to the *i*th household and *j*th village respectively. It is sometimes convenient to think of the district parameter as reflecting the deviation of the mean value of AI for district *k* from the overall mean value of AI across all districts. However, software for modeling use a different parameterization and sets

[7] This limitation on the target population limits inference to the population residing in villages in this size range.

[8] The asset index was a weighted average based on different livestock numbers and household assets, e.g., radio, bicycle, oxcart, etc.

[9] The income index was based on income from a range of different sources.

one of the district effects to zero. The remaining effects then provide a comparison of the asset index of each district with the district whose effect has been set to zero.

78. The U_{jk} and ε_{ijk} in this model, denote random variables representing respectively the variation among all villages within district k (assumed the same for all districts), and the variation among all households in village j in district k (assumed the same for all village and district combinations). The U_{jk} and ε_{ijk} are random variables that are assumed in the model to be normally distributed variables with zero mean and constant variances σ_u^2 and σ_e^2 respectively. They are further assumed to be independent of each other. We may therefore write $U_{jk} \sim N(0, \sigma_u^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma_e^2)$.

79. Fitting this model provides estimates of σ_u^2 and σ_e^2 and estimates for d_k, along with relevant standard errors. The parameter estimates for d_k (k=1, 2,…, 27), allow a comparison of the AI means across the 27 districts.

80. Now suppose that it was of interest to investigate how the variation in the asset index was affected by the size of household (a quantitative variate) and whether or not the household received a starter pack (a binary variable). These would be included in the model in the same way as would be done in standard general linear modeling. The model would be given by:

Model 6

$$y_{ijk} = \mu + d_k + U_{jk} + t_{p(ijk)} + \beta x_{ijk} + \varepsilon_{ijk},$$

where $t_{p(ijk)}$ represents the effect corresponding to the receipt of the starter pack; x_{ijk} represents the size of the household, and β represents the slope describing the relationship of household size (HHSIZE) to y_{ijk}, i.e., the asset index, AI.

81. Here both $t_{p(ijk)}$ and β are regarded as fixed effects, while U_{jk} and ε_{ijk} (as before) are random effects. The form of this model assumes that the relationship of HHSIZE to AI is the same across all villages and districts.

82. The inclusion of both components of variation (village and household) in the above model means that the model takes account of the variability at 2 levels of the hierarchy. This means that standard errors associated with $t_{p(ijk)}$ and β are calculated correctly, as would be the results of tests of significance associated with these parameters. This would not have been the case if a general linear model had been fitted regarding villages as fixed effects. Even if survey software (which take account of sampling weights) is used, standard regression procedures would ignore the correlation structure between households within any one village.

83. There is another important benefit in treating villages as random effects. If villages had been regarded as fixed, then the conclusions from the analysis apply only to the set of villages visited during the survey. Regarding villages as random effects means that the conclusions concerning the relationship of household size to the asset index, the comparison of the asset index across households receiving or not receiving the starter pack, and the comparison across districts,can all be generalized to the wider population of all villages. The interaction between

the district level fixed effect d_k and the starter pack recipient effect $t_{p(ijk)}$ can also be explored by including such an interaction term in the model.

84. A further useful model is obtained by regarding the slope term β in model 6 as a random variable that varies across the villages. This is often referred to as a random coefficient regression model. The model then becomes:

Model 7

$$y_{ijk} = \mu + d_k + U_{jk} + t_{p(ijk)} + \beta_j x_{ijk} + \varepsilon_{ijk}$$

where β_j is assumed $N(\beta, \sigma_\beta^2)$. Further, since β_j is random across villages, it may also be considered to have a covariance with U_{jk}, say $\sigma_{\beta u}$.

85. Thus, in the analysis here, testing the hypothesis that σ_β^2 is zero, effectively tells us whether there is variability in the slope of the AI vs. HHSIZE relationship across villages. If this hypothesis cannot be rejected, then it may be concluded that the form of the relationship is the same for all villages.

86. It is possible to extend this model further to include village level variables, e.g., access to a clean water supply or the degree of availability of advice from agricultural extension officers. Here, the real benefits of multi-level modeling come into play since it would be possible to explore relationships between such village level variables and the household level variables. Thus the study of relationships between variables at different levels of a hierarchical sampling scheme becomes possible through multi-level modeling. The benefits lie in being able to take account of the correlation structure among lower level units when variables at different levels are being analysed together. In the above example, further models could be considered, e.g., models that include gender and age of the household head, and interactions between these and terms previously included in the model.

87. There are of course limitations associated with fitting multi-level models. As with all other modeling procedures, the hypothesized multi-level model is assumed to be "correct" to a reasonable degree and to conform to the sample design. Whether such assumptions are true is of course debatable.

VIII. Modeling to support survey processes

88. Even when a household survey is used strictly to provide descriptive statistics, there may be need for modeling to support other survey processes. Adjustments for non-response are often based directly or indirectly on statistical models; Groves et al. (2002, pp. 197-443) discuss a variety of methods for accounting for non-response all of which must assume some statistical model. Logistic regression models may be used to develop predicted response propensities for the purpose of non-response adjustment or to identify weighting classes based on similar response propensities (e.g., Folsom, 1991; Folsom and Witt, 1994; or Folsom and Singh, 2000). Predictive statistical models may also be used as part of the procedure for imputing missing data

(e.g., Singh, Grau and Folsom, 2002). Finally, statistical models can be used to evaluate methodological experiments embedded in surveys (e.g., Hughes et al., 2002).

IX. Conclusions

89. Our aim in this chapter has been to discuss issues involved in the analysis of survey data. These issues include the use of survey weights and of appropriate variance estimation methods with both descriptive and analytic uses of survey data. The chapter also provides an overview of practical situations where modeling techniques have a role to play in survey data analysis. They are useful tools but their application requires careful thought and attention to their underlying assumptions.

90. We have discussed the role of survey weights and recognition of the sample structure in developing both descriptive and analytic statistics from survey data. Survey data analysis software that use survey weights and take account of the sample structure may be used to estimate the parameters of both linear and logistic regression models based on survey data. The estimates based on the sample are estimates of what would be obtained from fitting the models to the entire finite population. Furthermore, standard errors of the estimates can also be obtained. The explanatory variables in regression models applied to survey data are almost always observed as they exist in the population rather than randomly assigned according to some experimental design. Analysts need to be clear that regression coefficients based on survey data simply reflect relationships that exist between the dependent variable and the explanatory variables in the population and do not necessarily imply causation. We have discussed how the parameters of regression and logistic regression models relate to simple descriptive statistics and how they may be interpreted for some relatively simple models.

91. Multi-level modeling in particular would generally be regarded as a rather “advanced” technique and is best carried out in consultation with a statistician familiar with the use and limitations of this technique. At present multi-level models appear to be rarely used in analyzing surveys in developing countries, but their use is highly desirable for the insights they can provide concerning inter-relationships between variables at different levels and their ability to take account of variability amongst sampling units at different levels in a multi-stage design.

92. We have shown that the formulation of multi-level models is not too difficult for someone familiar with the application of general linear models (GLM), but again there are assumptions associated with the models that need to be checked by carrying out residual analyses as would be the case with GLMs. The multi-level modeling approach can also be undertaken when the main response of interest is binary, although we have not presented an example of such a model. Care is also needed in deciding which effects are random and which are fixed and how the model specification will help in answering specific survey objectives.

93. However, as with all statistical techniques, the modeling methods discussed in this chapter have various limitations and these need to be recognized in their application. We have urged the use of survey weights and analysis software that recognizes the sample design structure. The difficulty of access to appropriate software that takes account of the sampling design must be recognised. Chapter 21 describes several software packages that pay attention to

sampling design issues with respect to multiple regression and logistic regression procedures. Unfortunately however, these packages do not have facilities for fitting multi-level models. For this purpose, the user needs to turn to more general-purpose statistical software such as SAS (2001), Genstat (2002) or SPSS (2001), or to a specialist software package such as *MLwiN* (Rahbash et al., 2001).

94. This chapter has offered some modeling techniques that can serve as useful tools for survey data analysis. We recommend that survey analysts and researchers seriously consider these methods, where appropriate to survey objectives, during survey data analysis, with a view to extracting as much information as possible from expensively collected survey data.

Acknowledgements

The authors express special thanks to the reviewers and editors for many useful comments and suggestions, and especially to Dr. Graham Kalton for the enhancements to the discussion on survey weighting.

The UK Government's Department for International Development (DFID) is also thanked for providing ideas for this chapter through its funding of many interesting projects involving surveys in the developing world. The material in this paper however remains the sole responsibility of the authors and does not imply the expression of any opinion whatsoever on the part of DFID.

References

Chambers, R.L. and C.J. Skinner, 2003. Analysis of Survey Data. Wiley, Chichester, UK.

Chromy, James R., 1998. "The Effects of Finite Sampling on State Assessment Sample Requirements." Palo Alto, CA: NAEP Validity Studies, American Institutes for Research.

Cochran, W. G., 1977. Sampling Techniques, Third Edition. New York: John Wiley & Sons.

Congdon, P., 1998. A multi-level model for infant health outcomes: maternal risk factors and geographic variation. *The Statistician*, 47, Part 1, 159-182.

Deville, J.C. and C.E. Sarndal, 1992. "Calibration Estimating in Survey Sampling." *Journal of the American Statistical Association* 87:376-382.

Folsom, Ralph E., Jr., 1991. "Exponential and Logistic Weight Adjustments for Sampling and Non-response Error Reduction." Pp. 376-382 in *Proceedings of the Social Statistics Section, American Statistical Association*.

Folsom, Ralph E. and Michael B. Witt, 1994. "Testing a New Attrition Non-response Adjustment Method for SIPP." Pp. 428-433 in *American Statistical Association, Section on Survey Research Methods*.

Folsom, R. E. and A.C. Singh, 2000. "The General Exponential Model for Sampling Weight Calibration for Extreme Values, Non-response, and Post-stratification." in *Proceedings of the Survey Research Methods Section, American Statistical Association*. Indianapolis, Indiana.

GenStat, 2002. *GenStat for Windows. 6th Edition.* Oxford: VSN International Ltd.

Goldstein, H. 2003. *Multi-level Statistical Models*. 3nd Edition. Arnold, London.

Goldstein, H., J. Rasbash, M. Yang, G. Woodhouse, H. Pan, D. Nuttall and S. Thomas, 1993. A multi-level analysis of school examination results. *Oxford review of education*, 19, 425-33.

Graubard, B. I. and E.L. Korn, 2002. "Inferences for super-population……" *Statistical Science* 17:73-96.

Groves, Robert M., Don A. Dillman, John L. Eltinge and Roderick J.A. Little, 2002. "Survey Non-response." New York, NY: John Wiley & Sons, Inc.

Horvitz, D. G., and D.J. Thompson, 1952. "A generalization of sampling without replacement from a finite universe." *The Journal of the American Statistical Association* 47:663-685.

Hughes, Arthur, James Chromy, Katherine Giacoletti and Dawn Odom, 2002. "Impact of interviewer experience on respondent reports of substance use." pp. 161-184 in *Redesigning an Ongoing National Household Survey: Methodological Issues. DHHS Publication No. SMA 03-3768,* edited by Gfroerer, J., Eyerman, J. and Chromy, J., Rockville, Maryland: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Kish, Leslie, 1965. *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kreft, I. and J. de Leeuw, 1998. *Introducing Multi-level Modeling*. Sage, London.

Korn, E. L. and B.I. Graubard, 1999. Analysis of Health Surveys. New York, John Wiley & Sons.

Korn, E. L. and B.I. Graubard, 2003. Estimating variance components by using survey data. *Journal of the Royal Statistical Society B, 66*, 175-190.

Langford, I.H., G. Bentham and A. McDonald, 1998. Multi-level modeling of geographically aggregated health data: a case study on malignant melanoma mortality and UV exposure in the European community. *Statistics in Medicine*, 17, 41-58.

Levy, S. and C.I. Barahona, 2001. *The Targeted Inputs Programme, 2000-01*. Main Report. (Unpublished).

Pfeffermann, D., C.J. Skinner, D.J. Holmes, H. Goldstein and J. Rasbash, 1998. Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistical Society B*, 60, 23-40.

Rasbash, J., W. Brown, M. Healy, B. Cameron, C. Charlton, 2001. *MLwiN Version 1.10.0007*. Multi-level Models Project, Institute of Education, University of London.

Rubin, Donald B., 1987. *Multiple Imputation for Non-response in Surveys*. New York, NY: John Wiley & Sons.

SAS, 2001. *SAS Release 8.2*. SAS Institute Inc. SAS Publishing.

Singh, Avinash, Eric Grau and Ralph Folsom Jr., 2002. "Predictive mean neighborhood imputation for NHSDA substance use data." Pp. 111-134 in *Redesigning an Ongoing National Household Survey: Methodological Issues. DHHS Publication No. SMA 03-3768,* edited by Gfroerer, J., Eyerman, J. and Chromy, J. Rockville, Maryland: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Skinner, C.J., D. Holt and T.M.F. Smith. Editors, 1989. *Analysis of Complex Surveys*. Wiley, New York.

Snijders, T.A.B. and R.J. Bosker, R.J., 1999. *Multi-level Analysis: An Introduction to Basic and Advanced Multi-level Modelling*. Sage, London.

Tukey, J. W., 1977. *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

SPSS, 2001. *SPSS for Windows. Release 11.0.0.* LEAD Technologies, Inc., Illinois, U.S.A.

The World Bank, 2000. Voices from the village: A comparative study of coastal management in the Pacific Islands. *Final Report, Washington, D.C., The World Bank*, 87 pp.

Woodruff, R. S., 1971. "A simple method for approximating the variance of a complicated sample." *Journal of the American Statistical Association* 66:411-414.