CHAPTER 9: ANALYSIS OF HOUSEHOLD SURVEY DATA

Andrés Gutiérrez[[1]](#footnote-20), Pedro Luis do Nascimento Silva[[2]](#footnote-22)

2024-10-10

# Abstract

Analyzing complex household survey data requires knowing and properly applying the foundations of the design-based inference. The researcher will be faced to a small database that contains specific information that will allow her to make conclusions over the whole population.

The purpose of any analysis on this kind of datasets is not referred to make conclusions on the sample itself – which in most of the cases is a small subgroup of the population - but to the domains of interest and the whole population. Having that into account, the first step in any analysis plan should be devoted to defining the sampling design based on the selection mechanisms used to draw the final sample and the findings on the field related to nonresponse and lack of coverage.

The chapter covers three main topics of analysis: descriptive statistics; comparisons and association; and modeling of survey data. On the one hand, we introduce simple descriptive statistics, such as totals, frequencies, means and proportions, quantiles and some graphics; on the other, we delve deeper on complex relationships between the variables of the survey. All these analyses rely on the representativity principle of the design-based inference. This way, the reader will find a strong focus, not only on point estimates, but also on uncertainty measures. The chapter also presents a short discussion on the different approaches that can be used to estimate variances; the best way to visualize the estimates; and NSO practical experiences.

# Introduction

A key concern of every agency that produces statistical information is with the *correct* use of the data that it produces. This is even reflected in the United Nations *Fundamental Principles of Official Statistics*, namely:

* **Principle 3.** To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.
* **Principle 4.** The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

Here we emphasize a particular aspect, aiming to empower users when analysing household survey data. The computer revolution, with the resulting ease of access computers, created favorable conditions for the increased use of statistical data, including those resulting from household sample surveys. Sometimes this data is used for purely descriptive purposes. Other times, however, its use is made for analytical purposes, involving the testing of hypothesis or the construction of models, when the objective is to draw conclusions that are also applicable to populations other than the one from which the sample was extracted. In such cases, standard statistical software may provide biased or misleading results. This chapter’s purpose is to present the relevant models, methods and software to enable users to account for the complex survey design frequently used to conduct household sample surveys when analysing the resulting data.

What makes such data special for those who intend to use them for analytical purposes? The answer is that they are obtained through complex sample surveys of finite populations that often involve: *stratification*, *clustering* of units of analysis, *unequal probabilities of selection*, and *weighting adjustments* to compensate for non-response and/or improve precision.

Standard data analysis methods and software typically ignore these aspects, and may produce biased estimates of both the target parameters and the variances of these estimates. In this chapter we analyze the impact of simplifications made when using standard data analysis methods and software, and present the necessary adjustments to these procedures in order to appropriately incorporate the aspects highlighted here into the analysis.

In section 9.1, a short discussion on the fundamental principles of the design-based inference is presented, emphasizing that conclusions taken from probability sample surveys should be based on a pair: the point estimate and it associated margin of error (or any related measure). In section 9.2, we begin the journey with simple descriptive statistics: means, ratios, proportions and other typical descriptive parameters. Section 9.3 is devoted to more complex parameters that allow comparisons of the phenomenon of interest between subgroups for continuous and discrete variables. In this section we present standard tests to compare means and measure the degree of association between variables, and also consider the problem of correlation and association. Section 9.4 focuses on modelling survey outcomes. We first involve the reader in an discussion on the role of weighting when estimating regression coefficients. Then, we introduce some proper approaches to estimate complex parameters in linear and logistic regression models. Finally, section 9.5 presents a summary of ideas and tools for survey data visualization showing the best practices for creating graphics and maps in a context where uncertainty measures of estimates are important.

Most of the sections in the chapter present national experiences that will guide the reader on how NSOs are dealing with the different stages of the analysis of household survey data.

The purpose of this chapter is defining and explaining basic concepts of the design-based paradigm in household surveys to analyze complex household survey data. In section 9.1, a short discussion on the fundamental principles of the design-based inference is presented, emphasizing that conclusions taken from this kind of surveys should be based on a pair: the point estimate and it associated margin of error (or any related measure). In section 9.2, we begin the journey with simple descriptive statistics: means, ratios, proportions and other parameters are part of this discussion. Section 9.3 is devoted to more complex parameters that allow comparisons of the phenomenon of interest between subgroups for continuous and discrete variables. In this section we present standard tests to compare means and measure the degree of association between variables. This section also deals with the problem of correlation and association. Section 9.4 focuses on modelling survey outcomes. We first involve the reader in an discussion on the role of weighting when estimating regression coefficients. Then, we introduce some proper approaches to estimate complex parameters in linear and logistic regression models. Finally, section 9.5 presents a summary on survey data visualization showing the best practices for creating graphics and maps in a context where uncertainty measures of estimates are important.

Most of the sections in the chapter present national experiences that will guide the reader on how currently NSOs are dealing with the different stages of the analysis of household survey data.

# 1 The golden pair: sample design and estimator

Accounting for the sampling design is crucial for analyzing complex survey data. We must ensure that PSU, strata, and weights are available in the survey dataset to enable adequate analysis. Alternatively, when such information is not available, the dataset should at least contain replicate weights, or the analyst should have clear guidance on how to compute both point and variance estimates.

A well-described survey design facilitates statistical analysis, supports effective data interpretation, and enables meaningful insights into complex phenomena. Missing or incorrect design information may lead to biased estimates and misleading conclusions.

## 1.1 Parameters and estimators

Under a design-based approach this section presents the basic principles of inductive inference and how, using the *sampling weights* (from chapter VIII), one can get consistent estimators for population parameters of interest. We adopt the notation introduced in chapter VIII for presenting the expressions required here.

The *population total* and *mean* of a survey variable can be estimated by weighted estimators given by and , respectively. When the survey weights are calibrated and/or non-response adjusted, the above expressions may still be used, but with the calibrated or non-response adjusted weights, say, replacing the design weights , for all .

Here denotes the set of units in a sample selected from the population using a *probability sampling design* , that ensures strictly positive first order inclusion probabilities . These inclusion probabilities are assumed known , at least to the data producers.

Under the design-based framework and assuming full response, is unbiased for and its sampling variance is given by

where and . This result assumes that the sampling design is such that .

Under full response, this variance can be estimated unbiasedly by

While the above formula for variance estimation is general and covers the vast majority of sample designs used in the practice of household sample surveys, it is not used in practice because the second order inclusion probabilities (and corresponding pairwise weights ) are generally unknown to survey data analysts. In fact, even data producers do not compute such pairwise weights, since there are more efficient methods for variance estimation that do not require having such weights.

## 1.2 Uncertainty in household surveys

As the sample is typically a small subset of the population, it is important to obtain not only point estimates for the parameters of interest, but also the corresponding uncertainty measures and/or confidence intervals. In this subsection we present some approaches for variance estimation: approximate formulas from *Taylor linearization* and/or the *ultimate cluster* approach for variances under multi-stage cluster sampling. We also introduce replication methods and generalized variance functions, which are essential when PSU or strata are missing from the sample dataset.

A unifying idea of sampling theory is that of estimating equations - David A. Binder ([1983b](#ref-Binder1983)). Many population parameters can be written/obtained as solutions for *population estimating equations*. A generic population estimating equation is given by $\_{i U} z\_i () = 0 $, where is an *estimating function* evaluated for unit and is a population parameter of interest.

For the case of the population total, take . The corresponding population estimation equation is given by , and solving for gives the population total . Similarly, take for the population mean. As a final example, consider the ratio of population totals. Taking , the corresponding population estimation equation is given by . Solving for gives the *population ratio* .

The idea of defining population parameters as solutions to population estimating equations allows defining a general method for obtaining corresponding sample estimators. It is a matter of using the *sample estimating equations* . Under *probability sampling*, full response and with , the sample sum in the left hand side is unbiased towards the population sum in the corresponding population estimating equation. Solving the sample estimating equation yields consistent estimators for the corresponding population parameters.

The case of the population mean yields the sample estimating equation , and by solving on , we recover the Hàjek estimator . In the case of the population ratio, solving on , yields the well-known estimator .

The variance of estimators obtained as solutions of sample estimating equations can be obtained as:

where , and is the solution of the corresponding population estimating equation.

A consistent estimator of this variance is given by:

where .

This approach implies that by one is able to estimate many population parameters and corresponding variances using essentially well known methods for estimating totals. Its simplicity and generality have enabled the development of software such as the R survey package, the Stata svy functions and others.

## 1.3 Ultimate Cluster Method

The central idea of the *Ultimate Cluster* method for variance estimation for estimators of totals in multi-stage cluster sampling designs, proposed by ([Hansen, Hurwitz, and Madow 1953](#ref-Hansen1953)), is to consider only the variation between information available in the level of PSUs, and assume that these would have been selected with replacement from the PSU population. This idea is simple, but quite powerful, because it allows to accommodate a variety of sampling designs, involving stratification and selection with unequal probabilities (with or without replacement) of both PSUs as well as lower level sampling units. The requirements for the application of this method are that one has unbiased estimators of totals for the variable of interest for each sampled PSU, and that data are available for at least two sampled PSUs in each stratum (if the sample is stratified in the first stage).

Although the method was originally proposed for estimation of variances of estimated totals, it can also be applied in combination with Taylor linearization to obtain variance estimates for estimators of other population quantities that can be obtained as solutions to sample estimating equations.

Consider a multi-stage sampling design, in which PSUs are selected in stratum . Let be the inclusion probability of PSU stratum , and by an unbiased estimator of the total of the survey variable for the -th PSU in stratum , . Hence an unbiased estimator of the population total is given by , and the *ultimate cluster* estimator of the corresponding variance is given by:

where and are the population and sample sets of PSUs in stratum , , for . (See for example, ([Babubhai. V. Shah et al. 1993](#ref-Shah1993)), p. 4).

Although often the selection of primary units can have Primary Cluster estimator presented here may provide a reasonable approximation of the corresponding variance of randomization. This is because sampling plans without replacement are generally more efficient than plans with replacement of equal size. Such an approximation is widely used by sampling practitioners to estimate variances of usual descriptive quantities such as totals and medium (with due adaptation) due to their simplicity, compared to the much greater complexity involved with the employment of variance estimators that attempt to incorporate all steps of plans sampling in several stages. A discussion about Quality of this approximation and alternatives can be found in ([Särndal, Swensson, and Wretman 1992](#ref-SSW92)), p. 153.

In some cases, sample replication methods (*bootstrap*, *jackknife*) can also be used to estimate variances, as we will see later.

## 1.4 Bootstrap Method

The method was proposed by Efron ([1979](#ref-Efron1979)), but the version we consider here is the so-called Rao-Wu-Yue Rescaling Bootstrap, which is adequate for stratified multi-stage sampling designs commonly used in household surveys - see Rao, Wu, and Yue ([1992](#ref-Rao1992)). This method is now widely used for variance estimation with complex survey data. To implement this method, you need to follow the sequence of steps outlined below.

Step 1. Select a simple random sample with replacement of size of PSUs in each of the design strata. Each selected PSU takes with it all the subordinate sampling units and their data.

Step 2. Repeat Step 1 times, and denote by the number of times the PSU of stratum was selected for the sample in replicate .

Step 3. Calculate the *bootstrap* weight of unit within PSU of stratum as .

Step 4. For each replica , calculate an estimate of the target parameter using the *bootstrap* weights .

Step 5. Estimate the variance using:

where is the average of the replica estimates.

Whenever the original sampling weights receive non-response adjustments or are calibrated, the corresponding non-response adjustments and/or calibration of the basic weights must be repeated for each replica, so that the variance estimates adequately reflect the effects of the calibration and non-response adjusments on the uncertainty of the point estimates.

This method is more computationally costly, but provides good estimates of variance, including for quantiles and other parameters of complex nonlinear forms. It also makes it easier for users who do not have access to software capable of calculating complex variance expressions based on Taylor linearization, etc. The survey package allows you to generate *bootstrap* replicas and also estimate variances using this method.

## 1.5 Using software to generate valid inferences

In this part, we advocate to using specialized statistical software to generate efficient estimation processes. Those packages support complex survey data analysis by specifying the survey design using appropriate commands or functions.

# 2 Descriptive parameters

The most frequent analysis of complex household survey data consists in estimating some descriptive population parameters for a range of survey variables. Such *descriptive analysis* generally involve estimating frequencies, proportions, means, and totals. But other target parameters such as selected quantiles of numeric variables, poverty and inequality measures, and a range of indicators such as those required for monitoring the Sustainable Development Goals are becoming part of regular set of estimates needed from household sample surveys.

## 2.1 Frequencies

### 2.1.1 Point Estimation

The accurate estimation of absolute sizes and proportions in household surveys is essential for obtaining representative data that reflects the demographic and socioeconomic reality of a population. These figures serve as the basis for public policy decision-making, resource allocation, and the design of social programs.

The ability to understand the distribution of specific categories, such as poverty status, employment status, education level, among others, provides valuable information to address inequalities and promote equitable development.

### 2.1.2 Size Estimates

In this section, the processes for estimating categorical variables will be carried out. First, one of the most important parameters is the size of a population, which represents the cardinality of that set; in other words, the total number of individuals that comprise it. In terms of notation, the population size is estimated as follows:

where is the sample of households or individuals in PSU of stratum ; is the sample of PSUs within stratum ; and is the weight (expansion factor) of unit within PSU in stratum .

Similarly, the size estimate in a subpopulation, defined by a dichotomous variable , which takes the value one if unit from PSU in stratum belongs to category in the discrete variable , is given by the following expression:

## 2.2 Totals, means, proportions, and ratios

For single numeric survey variables, the simplest estimates are for totals and means. Ratios are often used to obtain summaries that relate two numeric variables. Estimates for such parameters can be obtained either for the entire population or disaggregated by domains of interest, depending on the research needs.

As mentioned by Steven G. Heeringa, West, and Berglund ([2017a](#ref-Heeringa_West_Berglund_2017)), the estimation of population totals or averages for a variable of interest, along with the estimation of corresponding variances, has played a crucial role in the development of probability sampling theory. Estimators of population means, proportions and ratios are all dependent on estimating component population totals, as we show in the sequence.

### 2.2.1 Estimating totals

Once the sampling design is defined, which was done in the previous section, the estimation process for the parameters of interest is carried out. For the estimation of totals with complex sampling designs that include stratification and subsampling in PSUs (assumed to be within stratum ) indexed by , the estimator for the population total can be written as:

Under full response, the Ultimate Cluster variance estimator for was provided in section 9.1. Calculating the total estimate and its estimated variance is complex, but now these calculations can be easily performed using the svytotal function from the survey package in R. The confidence interval of level is given by the following expression:

with denoting the quantile of the Gaussian distribution leaving an area of to its right.

### 2.2.2 Estimating averages

The estimation of the population means or averages is a very important task in household surveys. According to Gutiérrez ([2016](#ref-Gutierrez_2016)), an estimator of the population mean can be written as the ratio of two estimated finite population totals, as follows:

Since is a nonlinear statistic, there is no closed-form formula for exact the variance of this estimator. For this reason, either resampling methods or Taylor series approximations must be used. The latter may be achieved remembering that for the survey mean the sampling estimating equation requires defining , therefore we can apply the variance estimator given in section 9.1 with .

### 2.2.3 Estimating proportions

When is a binary variable, the weighted mean estimates the population proportion. As mentioned by Steven G. Heeringa, West, and Berglund ([2017b](#ref-Heeringa2017)), by recoding the original response categories into simple indicator variables with possible values of 1 and 0 (e.g., 1=Yes, 0=No), the estimator for a proportion is defined as follows:

We can apply Taylor linearization to obtain the approximate variance of the above estimator by defining the estimating function as .

Many statistical packages provide proportion estimates and standard errors on a percentage scale. As is well known in the specialized literature, when the estimated proportion of interest is close to zero or to one, the limits of the traditional symmetric normal confidence intervals may fall outside the permissible range for proportions. This would have no interpretation due to the nature of the parameter.

To address this issue, alternative confidence interval estimates, as proposed by Rust, Hsu, and Westat ([2007](#ref-Rust2007ConfidenceIF)) and Dean and Pagano ([2015](#ref-DeanPagano2015)) are available. One alternative based on using the logit transformation of the estimated proportion is:

### 2.2.4 Estimating ratios

In many household survey analyses, it is not sufficient to examine individual variables in isolation. For example, ODS indicator N.17.6.1 is defined as the ratio of the number of broadband subscriptions per 100 inhabitants in a country or region. Ratio estimators are obtained simply by the ratio of the corresponding estimators of totals (or means) in the numerator and denominator. Another example is estimating the ratio of expenditures to income or the ratio of a particular type of expenditure (say food) over total expenditures in a household budget survey.

Since the ratio is the quotient of two totals, both the numerator and the denominator are unknown quantities and thus need to be estimated. The point estimator for a ratio in complex surveys is the quotient of the estimators for the totals, as defined by:

For variance estimation, all you need to do is specify the estimating function as , when and are the numerator and denominator variables, respectively.

## 2.3 Variances and standard deviations

Sometimes the interest lies in estimating the variance or standard deviation of a numeric survey variable . This can be accomplished using the following estimators:

and for the standard deviation.

## 2.4 Correlations

Pearson correlation coefficients are useful for assessing the relationship between two numeric survey variables, say and . These can be estimated using

## 2.5 Percentiles and inequality measures

Non-central location measures are helpful to determine location and spread of the data distribution beyond central values. Key non-central location measures include the quartiles and other quantiles or percentiles. As an example, the estimation of income percentiles in a given country may help define the onset of public policy. For example, a tax could be imposed on individuals in the top 10% of the income distribution, or transport subsidies could be provided to those in the bottom 15% of the income distribution.

Quantile estimation is based on results related to weighted total estimators, by first estimating the population cumulative distribution function (CDF). The CDF for a variable in a finite population of size is defined as follows:

Where is an indicator variable that takes the value 1 if is less than or equal to a specific value , and 0 otherwise. An estimator of the CDF in a complex sampling design is given by:

Once the CDF is estimated using the survey design weights, the -th quantile of a variable is the smallest value of such that the CDF is greater than or equal to . As is well known, the median is the value where the CDF is greater than or equal to . Thus, the estimated median is the smallest value where the estimated CDF is greater than or equal to . Following Steven G. Heeringa, West, and Berglund ([2017b](#ref-Heeringa2017)), to estimate quantiles, one first considers the order statistics denoted as and finds the value of such that:

Hence, the estimator of the -th quantile is given by:

For the variance estimation and confidence intervals of quantiles, Kovar, Rao, and Wu ([1988](#ref-kovar1988bootstrap)) present results from a simulation study where they recommend using the *Balanced Repeated Replication* (BRR) technique.

### 2.5.1 Estimating the Gini coefficient

Economic inequality is a common issue worldwide, with particular focus from international institutions. Measuring economic inequality among households is of great interest, and the Gini coefficient () is the most commonly used indicator for this purpose. The Gini coefficient ranges from 0 to 1, where indicates perfect equality in wealth distribution, and higher values reflect increasing inequality.

Following the estimation equation proposed by David A. Binder and Kovacevic ([1995](#ref-binder1995estimating)), the estimator for the Gini coefficient is given by:

where is a normalized sampling weight, defined as

and represents the estimated CDF for individual in cluster of stratum .

Osier ([2009](#ref-osier2009variance)) and Langel and Tillé ([2013](#ref-Langel_Tille_2013)) provide important computational details for estimating the variance of this complex estimator.

## 2.6 NSO – Practical example

In this subsection a NSO will share how they do disseminate its results on basic descriptive statistics, how they publish the resulting tables and how do they deal with the suppression of estimates that do not reach expected quality.

# 3 Associations between categorical variables

Household sample surveys often collect data on categorical variables, and assessing whether pairs of these variables are associated becomes of interest. This section will introduce the reader on the main methods currently used to describe and infer associations for pairs of categorical variables.

We start by defining some notation. Let and denote two categorical variables, having and classes respectively. In order to formulate hypothesis tests for the independence between and , we need to consider a *superpopulation model*. We assume that the pairs correspond to observations from identically distributed random vectors , that have joint distribution specified by

with .

If a census could be carried out to collect data on and from every unit in the population, we could calculate the population counts of units having classes for given by:

and the corresponding population proportions as , where denotes the total number of units in the population.

Under the superpopulation model, the population proportions could be used to estimate (or approximate) the unknown probabilities . Since in most instances we will have samples, not censuses, the population proportions must be estimated using weighted estimators provided in the previous sections.

## 3.1 Cross-tabulations and contingency tables

Cross-tabulations organize population frequency distribution estimates for two or more categorical variables to help explore relationships between them. Tests of independence can be used to assess whether the cross-classified variables are related or independent. This type of analysis is important in many research and decision-making settings.

In the specialized literature, cross-tabulations are also referred to as contingency tables. Here a table is a two-dimensional array with rows indexed by and columns indexed by . Such tables are widely used in household survey analysis as they summarize the relationship between categorical variables in terms of frequency counts. A contingency table aims to succinctly represent the association between different categorical variables.

First we consider the case of a two-way contingency table. For most household sample surveys, a typical tabular output comprises the weighted frequencies that estimate the population frequencies, as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 |  |  | row marg. |
| 1 |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| col. marg. |  |  |  |  |

where the estimated frequency in cell is obtained as

and , and .

The estimated proportions from these weighted sample frequencies are obtained as follows:

, and .

Two-way tables can also display the estimates of population relative frequencies, as shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 |  |  | row marg. |
| 1 |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| col. marg. |  |  |  |  |

## 3.2 Testing for independence

Using the estimated contingency tables, it is possible to perform independence tests to verify whether and are associated. Following Steven G. Heeringa, West, and Berglund ([2017b](#ref-Heeringa2017)), the null hypothesis that and are independent is defined as:

Hence, to test the independence hypothesis we compare the estimated proportions with the estimated expected population proportions under the null . If there is a large difference between them, then the independence hypothesis would not be supported by the data. Therefore, the following Pearson Rao-Scott adjusted test statistic ([Rao and Scott 1984](#ref-Rao1984)) is defined:

where estimates the cell frequencies under the null hypothesis and is an estimate of the generalized design effect given by

where denotes the estimated design effect for the estimator .

Under the null hypothesis of independence, the large sample distribution of is .

As mentioned by Steven G. Heeringa, West, and Berglund ([2017b](#ref-Heeringa2017)), it was Fay ([1979](#ref-Fay1979)), along with Fellegi ([1980](#ref-Fellegi1980)), who began proposing corrections to Pearson’s chi-square statistic based on a generalized design effect. Rao and Scott ([1984](#ref-Rao1984)) later expanded the theory of generalized design effect corrections for these statistical tests, as did Thomas and Rao ([1987](#ref-thomas1987small_sample)). The Rao-Scott adjustment requires the calculation of generalized design effects, which are analytically more complex than Fellegi’s approach. Nevertheless, Rao-Scott adjusted statistics are now the standard for analyzing categorical survey data in software systems such as R, Stata and SAS.

The Rao-Scott adjusted Likelihood Ratio statistic is defined as:

Under the null hypothesis of independence, the large sample distribution of this test statistic is also .

When the number of degrees of freedom for the sample is not very large, two adjusted versions of the above test statistics might be preferable, since they enable taking this into account. The F-adjusted test statistic for independence based on Pearson’s is calculated as follows:

where denotes the degrees of freedom in the design.

The F-adjusted teststatistic for independence based on the Rao-Scott adjusted Likelihood Ratio statistic is calculated as:

## 3.3 Tests for group comparisons

To determine whether the means of two groups are significantly different we will introduce t-test and contrasts adjusted for the sampling design.

### 3.3.1 Hypothesis Test for the Difference of Means

A hypothesis test is a statistical procedure used to evaluate evidence in favor of or against a statement or assumption about a population. In this process, a null hypothesis () is proposed, representing the initial statement that needs to be tested, and an alternative hypothesis (), which is the statement opposing the null hypothesis. These statements may be based on some belief or past experience and will be tested using the evidence gathered from the survey data. If it is suspected that the parameter is equal to a particular value , the possible combinations of hypotheses that can be tested are:

One of the two hypotheses will be considered true only if the statistical evidence, which is obtained from the sample, supports it. The process of selecting one of the two hypotheses is called a Hypothesis Test.

In general, some important parameters can be expressed as a linear combination of measures of interest. The most common cases are differences in means, weighted sums of means used to construct economic indices, etc. Thus, consider a function that is a linear combination of descriptive statistics, as shown below:

where the are known constants. An estimator of this function is given by:

And its variance is calculated as follows:

As seen in the variance expression for the estimator, it requires the variances of the individual estimators, as well as the covariances of pairs of estimators.

Of particular interest is analyzing the difference in population means. In order to formulate the hypothesis tests for this case, we need to consider a *superpopulation model*. We assume that correspond to observations from identically distributed random variables having means if unit belongs to domain , with . Then we can define the difference in population means between domains 1 and 2 as . As an example, consider that is the average household income for households with male heads of household, and is the average household income for households with female heads.

This difference in means can be unbiasedly estimated by:

where is the sample estimator of (). Considering the parameter of interest in this section, the hypotheses to be tested are as follows:

To test one of these hypothesis, the following test statistic is used, which follows a t-student distribution with degrees of freedom, calculated as the difference between the number of PSUs (Primary Sampling Units) and the number of strata.

Where:

If a confidence interval for the difference in means is desired, it would be constructed as follows:

### 3.3.2 Contrasts

In many cases, it is necessary to compare more than two population means at the same time. For example, comparing the average household incomes across three regions to identify which regions experienced a greater impact of some shock or policy on households. In such cases, the difference of means we studied before is not sufficient, as it only allows for pairwise comparisons of population means. Using contrasts enables one to address these types of problems.

Based on the definitions discussed in this chapter, a contrast is a linear combination of parameters in the form:

Where is a matrix or vector of known constants, and is a vector of parameters.

XXX The example below is not well chosen, since in the end it simply compares two means again - reconsider the example or remove it altogether.

Alternatively, present the theory above already in terms of contrasts, and then the two-populations case is a simple yet good example. XXX

Next, we will perform the calculation of a hypothesis contrast. Suppose we have the estimates shown in the table, where the goal is to compare the average income by region. As a first example, we will compare two populations: the North and South regions. Specifically, we are interested in the income difference (). Since the population is divided into five regions and the contrast will only be constructed for two of them (North and South), it is defined as follows:

As can be observed, in this case, the contrast vector is . Mathematically, the estimator for this specific contrast is defined as:

Estimation table for regions.

| Region | Income | Standard error (se) | Lower bound (ci\_l) | Upper bound (ci\_u) |
| --- | --- | --- | --- | --- |
| North | 552.3637 | 55.35987 | 443.8603 | 660.8670 |
| South | 625.7740 | 62.40574 | 503.4610 | 748.0870 |
| Center | 650.7820 | 61.46886 | 530.3053 | 771.2588 |
| West | 517.0071 | 46.22077 | 426.4161 | 607.5982 |
| East | 541.7543 | 71.66487 | 401.2938 | 682.2149 |

To continue with the example, we take the estimated mean incomes for the North and South regions and calculate the difference:

The next step is to calculate the variance-covariance matrix and extract the variances for the North and South regions:

|  | North | South | Center | West | East |
| --- | --- | --- | --- | --- | --- |
| North | 3064.715 | 0.000 | 0.000 | 0.000 | 0.000 |
| South | 0.000 | 3894.476 | 0.000 | 0.000 | 0.000 |
| Center | 0.000 | 0.000 | 3778.420 | 0.000 | 0.000 |
| West | 0.000 | 0.000 | 0.000 | 2136.359 | 0.000 |
| East | 0.000 | 0.000 | 0.000 | 0.000 | 5135.854 |

Since the sampling is independent in each region, the covariances in the matrix are zero. To calculate the standard error of the difference (contrast), we will use the properties of variance, as follows:

Therefore, the estimated standard error for this contrast is:

## 3.4 NSO – Practical example

In this part an NSO will share its experiences on dealing with statistical comparisons among groups and how do they present the results in tables.

# 4 Regression: modelling survey data

Modelling survey data is a common task among researcher; some of them include the features of the sampling design in computing standard error of the estimated regression parameters. In this section we will deal with the problem of weighting in regression models and present a parsimonious solution.

## 4.1 To weight or not to weight?

We present the pros and cons of including the complex design features in the estimation of regression parameters and their associated standard errors. We present some adjustment to the sampling weights to fit these kind of models (senate sampling weights, normalized sampling weights, Pfeffermann model weights).

## 4.2 Some inferential approaches to modelling data

When modelling survey data, one should deal with two sources of variability: the one devoted to the complex sampling design and the one that comes from the very model. Combining these sources into a valid set up requires of some advanced methods. We will mention some of them: pseudo likelihood, combined inference.

## 4.3 Linear models

### 4.3.1 Basic Definitions

As noted by Steven G. Heeringa, West, and Berglund ([2017a](#ref-Heeringa_West_Berglund_2017)), the first authors to empirically discuss the impact of complex sampling designs on regression model inferences were Kish and Frankel ([1974](#ref-kish1974inference)). Later, Fuller ([1975](#ref-fuller1975regression)) developed a variance estimator for regression model parameters based on Taylor linearization with unequal weighting of observations under stratified and two-stage sampling designs.

As is well known, the use of regression model theory requires certain statistical assumptions to be met, which can sometimes be challenging to verify in practice. In this regard, B. V. Shah, Holt, and Folsom ([1977](#ref-shah1977inference)) discuss some aspects related to the violation of these assumptions and provide appropriate methods for making inferences about the estimated parameters of linear regression models using survey data.

Similarly, David A. Binder ([1983a](#ref-binder1983variances)) obtained the sampling distributions of estimators for regression parameters in finite populations and related variance estimators in the context of complex samples. Skinner, Holt, and Smith ([1989](#ref-skinner1989analysis)) studied the properties of variance estimators for regression coefficients under complex sample designs. Later, Fuller ([2002](#ref-fuller2002regression)) provided a summary of estimation methods for regression models containing information related to complex samples. Finally, Pfeffermann ([2011](#ref-pfeffermann2011modelling)) discussed various approaches to fitting linear regression models to complex survey data, presenting empirical support for the use of the “*q-weighted*” method, which is recommended in this document.

A simple linear regression model is defined as , where represents the dependent variable, is the independent variable, and and are the model parameters. The variable is known as the random error of the model and is defined as .

Generalizing the previous model, multiple linear regression models are defined by allowing the dependent variable to interact with more than two variables, as presented below:

Another way to write the multiple regression model is:

Where, and .

The subscript refers to the sample element or respondent in the dataset. Steven G. Heeringa, West, and Berglund ([2017a](#ref-Heeringa_West_Berglund_2017)) present some considerations for regression models, which are described below:

* , meaning that the expected value of the residuals conditioned on the covariates is zero.
* (homogeneity of variance), meaning that the variance of the residuals conditioned on the covariates is constant.
* (normality of errors), meaning that the residuals conditioned on the covariates follow a normal distribution. This property also extends to the response variable .
* (independence of residuals), meaning that the residuals in different observed units are not correlated with the values given by their predictor variables.

Once the linear regression model and its assumptions are defined, it can be deduced that the best unbiased linear estimator is defined as the expected value of the dependent variable conditioned on the independent variables , as:

Additionally,

It is also established that:

Thus, the response variable has the following distribution:

### 4.3.2 Estimation of Parameters in a Regression Model with Complex Samples

Once the assumptions of the model and the distributional characteristics of the errors are established, the next step is the process of parameter estimation. As an illustrative and introductory example, if instead of observing a sample of size from the elements of the population, a complete census had been conducted, the finite population regression parameter could be calculated as follows:

Now, when estimating the parameters of a linear regression model considering that the observed information comes from surveys with complex samples, the standard approach to estimating regression coefficients and their standard errors is altered. The main reason for this change is that data collected through a complex survey generally does not have an identical distribution, and the assumption of independence cannot be maintained since the sample design is constructed with dependencies (as most complex designs include stratification, clustering, unequal selection probabilities, etc.).

In this context, when fitting regression models with such datasets, using conventional estimators derived from traditional methods (such as maximum likelihood, for example) will induce bias because these methods assume the data are independently and identically distributed and come from a specific probability distribution (binomial, Poisson, exponential, normal, etc.). Instead, according to Wolter ([2007](#ref-Wolter_2007)), robust non-parametric methods based on Taylor linearization or variance estimation methods using replication (Jackknife, bootstrapping, etc.) are used to eliminate bias by including the sampling design in the analyses.

For illustrative purposes, the estimation of the parameter and its variance for a simple linear regression will be shown. The extension to multiple regression parameter estimation is algebraically complex and beyond the scope of this book. Below is the estimation of the slope and its variance in a simple linear regression model:

As can be seen in the above equation, the parameter estimator is a ratio of totals; therefore, its variance is given by:

As a generalization, according to Kish and Frankel ([1974](#ref-kish1974inference)), the variance estimation of coefficients in a multiple linear regression model requires weighted totals for the squares and cross-products of all combinations of and . Below is the estimation of these variances:

### 4.3.3 The Pfeffermann Weighting Approach

Steven G. Heeringa, West, and Berglund ([2017a](#ref-Heeringa_West_Berglund_2017)) addresses the problem of how to correctly weight regression models and whether expansion factors should be used to estimate regression coefficients when working with complex survey data. In this context, it is essential to know that two primary paradigms exist in the specialized literature:

* **The design-based approach**, illustrated in this document, seeks to make inferences about the entire finite population, and the use of expansion factors ensures that regression parameter estimates are unbiased. However, using survey weights does not protect against model misspecification; if the researcher fits a poorly specified model using expansion factors, unbiased estimates of the regression parameters in a model that does not correctly describe the relationships in the finite population are being computed.
* **The population-based modeling approach**, which argues that the use of expansion factors in estimation should not be necessary if the model is correctly specified. Under this approach, including survey weights only serves to increase the variance of the estimators, inducing larger-than-necessary standard errors.

The choice between these two approaches should depend on the sensitivity of inferences to different estimation methods. It is often recommended to use statistical software to fit regression models with and without survey weights to evaluate the sensitivity of the results. If the use of weights produces substantially different estimates and conclusions, it suggests that the model may be misspecified, and weighted estimates should be preferred. However, if the use of weights does not significantly alter the regression parameter estimates and only considerably increases standard errors, it could indicate that the model is well-specified, and the use of weights may not be necessary.

An intermediate solution between these two approaches is given by Pfeffermann ([2011](#ref-pfeffermann2011modelling)), who proposed a variant (called the *q-weighted approach*) with a slightly different specification of the expansion factors, detailed as follows:

1. Fit a regression model to the final survey weights using the predictor variables in the regression model of interest.
2. Obtain the predicted survey weights for each case as a function of the predictor variables in the dataset.
3. Divide the survey expansion factors by the predicted values from the previous step.
4. Use the new weights obtained for fitting the regression models.

This method adjusts the survey weights based on the fitted model, balancing between design-based and model-based approaches to reduce variance while accounting for complex survey design.

### 4.3.4 Model Diagnostics

When fitting statistical models to household survey data, it is essential to perform quality checks to ensure the validity of the conclusions drawn. Most academic texts provide a detailed overview of the assumptions and considerations necessary for a properly defined model. Below are some of the key aspects to consider:

* **Model fit**: It is important to determine whether the model provides an adequate fit to the data.
* **Distribution of errors**: Examine whether the errors are normally distributed.
* **Error variance**: Check whether the errors have constant variance.
* **Error independence**: Verify that the errors can be assumed to be uncorrelated.
* **Influential data points**: Identify if any data points have an unusually large influence on the estimated regression model.
* **Outliers**: Detect points that do not follow the general trend of the data, known as outliers.

#### 4.3.4.1 Coefficient of Determination

The coefficient of determination, also known as the multiple correlation coefficient (), is a common measure of goodness-of-fit in a regression model. This coefficient estimates the proportion of variance in the dependent variable explained by the model and ranges between 0 and 1. A value close to 1 indicates that the model explains a large proportion of the variability, while a value near 0 suggests the opposite.

The calculation of this coefficient for a population is done as follows:

Where:

* : This is the total sum of squares, representing the total variability in the dependent variable.
* : This is the sum of squared errors, representing the variability not explained by the regression model.

For surveys with complex sampling designs, the weighted estimator of is given by:

Where is the weighted sum of squared errors, defined as:

This estimator adjusts the calculation to reflect the characteristics of the sampling design, such as stratification and unequal selection probabilities, ensuring that survey weights are considered when evaluating the goodness-of-fit of the model.

#### 4.3.4.2 Standardized Residuals

In model diagnostics, analyzing residuals is crucial. These analyses provide, under the assumption that the fitted model is adequate, an estimate of the errors. Therefore, a careful study of the residuals should help the researcher conclude whether the fitting process has not violated the assumptions or if, on the contrary, one or more assumptions are not met, requiring a review of the fitting procedure.

To analyze the residuals, Pearson residuals ([Steven G. Heeringa, West, and Berglund 2017a](#ref-Heeringa_West_Berglund_2017)) are defined as follows:

Where is the expected value of , and is the survey weight for the i-th individual in the complex sample design. Finally, is the variance function of the outcome. These residuals are used to perform normality and constant variance analyses.

If the assumption of constant variance is not met, the estimators remain unbiased and consistent, but they are no longer efficient. That is, they are no longer the best in the sense that they no longer have the smallest variance among all unbiased estimators. One way to analyze the assumption of constant variance in the errors is through graphical analysis. This is done by plotting the model residuals against or the model residuals against . If these plots reveal any pattern other than a constant cloud of points, it can be concluded that the error variance is not constant.

#### 4.3.4.3 Influential Observations

Another set of techniques used for model analysis involves examining influential observations. An observation is deemed influential if, when removed from the data set, it causes a significant change in the model fit. It is important to note that an influential point may or may not be an outlier. To detect influential observations, it is essential to clarify what type of influence is being sought. For instance, an observation may be influential for parameter estimation but not for error variance estimation. Below are some statistical techniques for detecting influential data points:

1. **Cook’s Distance**: This diagnostic measures whether the i-th observation is influential in the model estimation by being far from the data’s center of mass. Various authors consider an observation influential when this value exceeds 2 or 3.
2. **Statistic**: This statistic measures the change in the estimated regression coefficient vector when the observation is removed. The i-th observation is considered influential for if with . Alternatively, can be used, where is the 97.5th percentile.
3. **Statistic**: This statistic measures the change in the model fit when a particular observation is removed. In this case, the i-th observation is considered influential in the model fit if with .

#### 4.3.4.4 Inference on Model Parameters

Once the proper fit of the model has been evaluated using the methodologies discussed above, and the distributional properties of the errors—and consequently the response variable —have been verified, the next step is to assess whether the estimated parameters are significant. This involves determining whether the covariates used to fit the model add value in explaining and/or predicting the study variable and the phenomenon of interest.

Given the distributional properties of the regression coefficient estimators, a natural test statistic for evaluating the significance of these parameters is based on the t-distribution and is described as follows:

Where is the number of model parameters and is the sample size of the survey. The test statistic above evaluates the hypotheses versus the alternative . Similarly, a confidence interval of for can be constructed, as follows:

Where the degrees of freedom () for the interval in a household survey (complex samples) is given by the number of final stage clusters minus the number of primary stage strata .

#### 4.3.4.5 Estimation and Prediction

According to Neter, Wasserman, and Kutner ([1996](#ref-neter1996applied)), linear regression models are essentially used for two purposes. One is to explain the variable of interest in terms of covariates that may be found in surveys, administrative records, censuses, etc. Additionally, they are also used to predict values of the variable under study, either within the range of values collected in the sample or outside of it. The first purpose has been addressed throughout this chapter, and the second is achieved as follows:

Explicitly, in the model exemplified in this chapter, the expression for predictions would be:

The variance of the estimation is calculated as follows:

## 4.4 Logistic models

To model the probability of discrete variables, we apply the principles of design-based inference.

### 4.4.1 Logistic Regression Model for Proportions

Logistic regression is a regression method that allows the estimation of the probability of success for a binary qualitative variable based on other continuous or discrete covariates. The variable of interest is binary or dichotomous, meaning it takes a value of one (1) if the condition being observed is met and zero (0) otherwise. In this way, the observations are separated into groups according to the value taken by the predictor variable.

If a categorical variable with two possible levels is coded as ones (1) and zeros (0), it is mathematically possible to fit a linear regression model using estimation techniques such as least squares. However, a problem arises with this approach: since the model is a straight line, it can produce estimated values that are less than zero or greater than one, which contradicts the theory requiring probabilities to always fall within the [0,1] range.

The objective of logistic regression is to model the logarithm of the probability of belonging to each group. As a result, assignment is ultimately made based on the obtained probabilities. Logistic regression is ideal for modeling the probability of an event occurring as a function of various factors. Therefore, the approximate probability of the event is represented by a logistic function of the form:

It is important to note that linear regression should not be used when the dependent variable is binary, as it cannot directly estimate the probability of the studied event. Instead, logistic regression is used, where a transformation (logit) is applied to obtain the probability estimates of the studied event. Applying the logit function to both sides yields an expression similar to that of linear regression:

Thus, a linear relationship is assumed between each of the explanatory variables and the logit of the response variable. There are at least three major differences between logistic regression and linear regression. First, in logistic regression, there is no requirement for a linear relationship between the explanatory variables and the variable of interest; second, the residuals of the model do not need to follow a normal distribution; and third, the residuals do not need to have constant variance (homoscedasticity).

Using appropriate techniques that include complex sampling designs in inference, the estimated probability that the variable of interest takes a value of one, which is also the expected value of the variable of interest in a logistic regression model, is:

The variance of the estimated parameters is calculated using the following expression:

Where:

and,

Where . Since the model uses a logarithmic link, confidence intervals are constructed by applying the exponential function to each parameter:

Therefore, the confidence interval is given by the following expression:

It is important to note that interpreting logistic regression coefficients can be challenging due to its non-linear nature. To facilitate interpretation, similarities and key differences with classic linear regression are highlighted. One similarity is that the sign of the estimated equation can be interpreted in the same way in both cases. A positive sign accompanying a covariate indicates an increase in the probability of the event occurring, while a negative sign indicates a decrease. As in linear regression, the intercept can only be interpreted assuming zero values for the other predictors.

However, the interpretation of regression coefficients between logistic and linear models differs significantly. The estimated coefficients in logistic regression correspond to a logarithm of odds, requiring the previously mentioned transformation. According to Gelman and Hill ([2019](#ref-gelmanhill2019)), the exponentiated logistic regression coefficients can be interpreted as odds ratios. If two outcomes present probabilities of , then is called the odds. For example, an odds ratio of 1 corresponds to a probability of 0.5, indicating equally likely outcomes. Doubling the odds further increases the probability to 8/9, and so on.

To determine whether a variable is significant in the model, it is common to use the Wald statistic, which is based on the likelihood ratio. In this case, the full model (with all parameters) is compared to the reduced model (with fewer parameters). The test statistic is given by:

This statistic evaluates the difference in fit between the two models, allowing for the assessment of the significance of the parameters included in the full model.

## 4.5 NSO – Practical example

In this subsection, we will share the experience of an NSO in applying models to household survey data, and the results they present in terms of significance of models and relations among variables.

# 5 Data visualization

In this section we discuss how to present data and estimates resulting from household surveys using graphics. While standard plots can still be used to show distributions and associations from the raw (unweighted) sample data, these can be misleading for the corresponding population distributions and associations. Therefore it is recommended that modified plots that account for survey weights be used instead. In addition, regarding the display of survey estimates, which are subject to sampling error, it is important to convey this message by presenting not only point estimates, but also standard errors or confidence intervals.

Graphs are important for the exploratory analysis of the survey data, for the diagnostics of fitted models and for the presentation of results. However, often the sample data sets are very large. In addition, sampling units typically have different weights. These two difficulties may cause standard graphs to fail in facilitating the analysis or presentation.

## 5.1 Graphs and sampling weights

When the survey units have different sampling weights, these should be taken into account when preparing graphs with their data. The main reason is that weights can be interpreted as the number of population units that each sample unit represents. Hence, it is evident that unequal weights need to be considered in the elaboration of graphs based on such sample data.

## 5.2 Graphs for categorical data

When the data of interest are categorical, their descriptive analysis will be done using contingency tables. Then simple graphs like bar charts can be done using as input contingency tables produced with weighted counts. Ideally one should also aim to display error lines overlaying bars to indicate their respective confidence interval widths, thus conveying the uncertainty of the corresponding point estimates. Obtaining the weighted counts or proportions and their confidence intervals can be easily done using tools from several software packages, e.g. the survey package in R.

## 5.3 Histograms

Histograms serve to present the distribution of a single numeric (continuous) survey variable or response. If one had a census, then the histogram is a powerful tool to describe the underlying distribution, even for very large datasets. When displaying sample data, however, the sampling weights must be taken into account when estimating frequencies or relative frequencies of population units having values in the specified histogram bins. Modern survey analysis tools can easily provide weighted histograms where the sampling weights are incorporated.

Histograms are the precursors to density function estimates, and the later can be thought of as histograms with very large number of bins. The survey package in R provides functions that can plot smoothed density estimates obtained accounting for the sampling weights.

## 5.4 Box Plots

Box plots are often used to present the distribution of continuous variables. They can summarize large datasets by providing a visual display providing easy visual to identify location, range, variability and outliers. They are great also for enabling comparing distributions across specified grouping variables, such as strata, clusters, etc.

The key to producing such graphs from complex sample surveys is to account for the sampling weights when estimating the location measures that drive the plot, namely the quartiles. Once these have been estimated using the methods described earlier, the resulting box plots will be good depictions enabling analysis of the underlying population distributions.

## 5.5 Scatter Plots

Scatter plots are the tool of choice to explore relationships between two continuous variables, potentially revealing patterns or trends in the data. These plots face the two challenges discussed above. First one needs to try and convey in the plot that the different sample observations carry different weights. For small to moderate sample sizes this can be done by plotting circles or dots of varying sizes where the symbol size represents the corresponding observation sampling weight. Plots like these can be obtained using standard bubble plot tools or the scatter plot available in the survey package in R.

The second challenge, present when there is a large dataset to be displayed, has motivated creation of some alternatives. Two ideas are worth noting. The first one is subsampling. One may choose to select a small to moderate subsample from the full dataset to display. Such a sample should be selected with replacement and with probabilities proportional to the observations sampling weights. Then the resulting smaller dataset can be used to produce a standard scatter plot. The subsample obtained in this way behaves approximately as a simple random sample from the survey population - see Lumley ([2010](#ref-Lumley2010)) page 69.

The other alternative is to produce so-called *hexagonal binned scatter plots*. This involves dividing the plot surface into a grid of hexagons and combining all the points that fall into a grid cell into a single plotted hexagon whose shading or size indicates the number of points in the bin. With complex household survey data, the number of points in a hexagonal bin should be replaced by the sum of the weights for points in the bin - see Lumley ([2010](#ref-Lumley2010)) page 70.

The third alternative is to avoid the display of the individual data points altogether, but instead produce *smoothed scatter plots*. One idea that can be useful would be to estimate specified quantiles (say the quartiles) of the y-axis (response) variable conditional on the values of the x-axis (predictor) variable, and smooth these across the range of the x-axis. Such plots can easily signal whether the y-variable has any relationship with the x-variable, and suggest the kind of curves that might be useful in summarising or modelling such a relationship - see Lumley ([2010](#ref-Lumley2010)) page 71.

## 5.6 Maps

Maps are the display of choice to present the behavior of the interest variable across geographical domains. Maps that aim to present how a single response variable behaves can be easily obtained by plotting a summary of the response across the domains. Such a summary (say mean or median) should be an estimate for the corresponding population parameter btained accounting for the sample design and weights.

Secondary survey analysts will may find that the limits of what mapping they can do is the level of geographic detail provided with the survey microdata. Many household sample surveys are design to provide precise estimates at some broad geographic level, say the country or its first level geographic subdivisions, such as states or departments. Lower level geographies are seldom disseminated with the survey microdata due to confidentiality protection constraints imposed.

It is therefore important that statistical agencies conducting the household sample surveys and preparing the dissemination of the corresponding microdata consider carefully which level of geographic detail may be included with public use datasets.

One area which still needs further research is that of providing appropriate means to convey the uncertainty of underlying point estimates when mapping these.

## 5.7 NSO – Practical example

In this subsection we will include the experience of a NSO on displaying information through graphics.

# 6 Other modeling scenarios

In this section we indicate the literature and software supporting the fitting of some other models to complex household survey data, including:

## 6.1 Multilevel models

Multilevel models, also known as mixed effects models or hierarchical models, are a statistical technique used in the analysis of household survey data to incorporate a hierarchical or multilevel structure. In these surveys, data are collected at the individual level (e.g., information about the age, gender, and education of each household member) and at the household level (e.g., household income, home ownership, and geographic location). Additionally, these models allow for the analysis of how household-level and individual-level factors influence responses to survey questions. For example, a multilevel model could be used to investigate how household income and the ages of household members influence the consumption of healthy foods.

In multilevel models, two types of effects are considered: fixed effects and random effects. Fixed effects represent the average relationships between variables, while random effects model the variation in these relationships across households. Thus, multilevel models account for heterogeneity in the population and provide more accurate estimates of the variables of interest. Therefore, multilevel models are a valuable tool in analyzing household survey data, as they allow for the examination of how both household-level and individual-level factors influence survey responses, while accounting for the hierarchical structure of the data.

Relevant references on the use of multilevel models in household surveys include Goldstein ([2011](#ref-goldstein2011)), a classic reference for multilevel data analysis that discusses the use of hierarchical models in various contexts, including household surveys; Gelman and Hill ([2019](#ref-gelmanhill2019)), which offers an accessible introduction to the theory and practice of hierarchical models; Rabe-Hesketh and Skrondal ([2012](#ref-rabe2012multilevel)), a practical guide for analyzing multilevel and longitudinal data using statistical software such as Stata; and Browne and Draper ([2006](#ref-browne2006comparison)), which compares Bayesian and frequentist approaches based on likelihood.

### 6.1.1 Model with Random Intercept

In the analysis of multilevel models, two types of estimates are relevant. The first is associated with the regression coefficients, generally referred to as the fixed parameters of the model; the second pertains to the variance estimates, usually called the random parameters of the model. Any multilevel regression analysis should always begin with estimating the variance at both levels for the dependent variable.

The recommended first step in multilevel regression analysis involves decomposing the variance of the dependent variable across the different levels. For example, assuming that the variable of interest is individual income and that there is a hierarchical relationship between individuals and strata, the variance of income can be decomposed into two parts: the variance within strata and the variance between strata. These two components of variance can be obtained from a simple multilevel regression with a null model represented by the following expression:

Here, represents the income of individual in stratum ; is the intercept in stratum ; is the residual for individual in stratum ; is the overall intercept, and is the random effect for the intercept. For this model, it is assumed that:

Additionally,

According to Cai ([2013](#ref-cai2013investigation)), there is sufficient evidence that sampling weights should be used in multilevel modeling to obtain unbiased estimates. Currently, different authors recommend various approaches on how to use sampling weights in hierarchical models. For instance, Pfeffermann et al. ([1998](#ref-pfeffermann1998weighting)) and Asparouhov ([2006](#ref-asparouhov2006general)) advise employing a pseudolikelihood approach to calculate estimates within and between different levels to achieve unbiased estimates.

An important concept in this type of analysis is the intra-class correlation (ICC), which refers to the proportion of the total variance of a variable that is explained by differences between groups or levels (strata) in the model. In other words, the ICC measures the similarity or correlation between observations within the same group or level compared to observations from different groups. This quantity is calculated as follows:

A high ICC indicates that a large proportion of the total variation in the variable is due to differences between groups, suggesting that the groups are distinct from one another and that the group effects should be considered in the model. Conversely, a low ICC indicates that most of the variation in the variable occurs within groups, implying that the group effects are not as significant in explaining variability in the variable.

### 6.1.2 Model with Random Slope

This type of model allows the relationship between the independent variable and the dependent variable to change according to some other explanatory variable. In other words, it permits the slope of the relationship between the variables to differ as groups or subsets of data are considered. In a simple linear regression model, the relationship is modeled as a straight line with a fixed slope. However, in a model with a random slope, the slope can vary according to another explanatory variable.

In these types of models, the relationship between the variables can imply a curve with different slopes for different subgroups. Random slope models are useful in situations where it is expected that the relationship between the variables changes in a nonlinear way or when differences in slope among subgroups need to be modeled. Consider the following model:

where is given as

In this particular case, the slope varies according to the sampling strata, while the intercept remains fixed. This allows for a more accurate capture of how the relationship between spending and income may differ across different groups, providing a better understanding of the patterns in the data.

### 6.1.3 Model with Random Intercept and Slope

Models with random intercepts and slopes are a type of statistical model that allows for modeling the relationship between a response variable and one or more predictor variables, considering both fixed and random effects. In these models, the regression coefficients (i.e., the slope and intercept) are treated as random rather than fixed, meaning these coefficients can vary across units of analysis, which may be individuals, groups, geographical regions, etc. These variations are modeled as random effects incorporated into the regression equation.

Continuing with the context of a household survey, consider the following model:

where the intercept and slope are modeled as:

and

In this model, and depend on the stratum variable, allowing both the intercept and slope to change according to the group of interest. This provides greater flexibility and better captures the heterogeneity in the data, reflecting how relationships between the variables may differ across subgroups.

## 6.2 Survival models

## 6.3 Loglinear models for contingency tables

When analyzing the relationships between variables that generate estimated totals in a contingency table (a rectangular arrangement that organizes data based on two or more categorical variables, showing the frequency or proportion of cases falling into each combination of categories), it is possible to use a log-linear model. This type of model is used to study the association between categorical variables while controlling for or considering potential effects of other covariates. Additionally, it allows for adjusting the observed associations in the contingency table and assessing whether these associations are statistically significant.

Log-linear models in contingency tables can be used to determine whether there is a significant association between categorical variables; adjust the association between variables of interest by accounting for other variables that may influence the relationship; evaluate how the probability of a category in one categorical variable changes given a change in another categorical variable; and estimate the probability of a case falling into a specific category of a categorical variable given the information from other variables.

The term log-linear essentially describes the role of the link function used in generalized linear models. In the simplest case, with two categorical variables inducing data from counts or proportions in contingency tables, the following statistical model can be formulated:

where is the conditional probability of occurrence for the combination of categories and in the categorical variables and , respectively; is the intercept representing the logarithm of the reference probability; and are the main effects associated with categories and in the variables and , respectively; and finally, is the interaction effect between categories and in the categorical variables. The natural logarithm function is commonly used in these models to transform conditional probabilities and allow for interpretation in terms of log-odds. In summary, the model describes how the conditional probabilities of categorical events are related to each other and how these relationships can be influenced by main and interaction effects in the categorical variables and .

This statistic is applied after a statistical model has been chosen (such as simple linear regression, logistic regression, among others). The Wald chi-squared test statistic for the null hypothesis of independence between rows and columns in a contingency table is defined as follows:

where

is a vector of differences between observed and expected cell counts, that is, . The matrix represents the estimated variance-covariance matrix for the difference vector. In the case of a complex survey design, the variance-covariance matrix of the weighted frequency counts, , is estimated using resampling methods or Taylor approximation. The matrix is the inverse of the matrix given by:

Under the null hypothesis of independence, the Wald statistic follows a chi-squared distribution with degrees of freedom:

The F transformation of the Wald statistic is:

# References

Asparouhov, Tihomir. 2006. “General Multi-Level Modeling with Sampling Weights.” *Communications in Statistics—Theory and Methods* 35 (3): 439–60.

Binder, David A. 1983a. “On the Variances of Asymptotically Normal Estimators from Complex Surveys.” *International Statistical Review* 51: 279–92.

Binder, David A. 1983b. “On the Variances of Asymptotically Normal Estimators from Complex Surveys.” *International Statistical Review* 51 (3): 279–92. <https://doi.org/10.2307/1402588>.

Binder, David A., and Milojica S. Kovacevic. 1995. “Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations Approach.” *Survey Methodology* 21 (2): 137–45.

Browne, William J., and David Draper. 2006. “A Comparison of Bayesian and Likelihood-Based Methods for Fitting Multilevel Models.” *Bayesian Analysis* 1 (3): 473–514. <https://doi.org/10.1214/06-BA117A>.

Cai, Tianji. 2013. “Investigation of Ways to Handle Sampling Weights for Multilevel Model Analyses.” *Sociological Methodology* 43 (1): 178–219.

Dean, Natalie, and Marcello Pagano. 2015. “Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys.” *Journal of Survey Statistics and Methodology* 3 (4): 484–503. <https://doi.org/10.1093/jssam/smv024>.

Efron, Bradley. 1979. “Bootstrap Methods: Another Look at the Jackknife.” *The Annals of Statistics* 7 (1): 1–26.

Fay, R. E. 1979. “On Adjusting the Pearson Chi-Square Statistic for Clustered Sampling.” *ASA Proceedings of the Social Statistics Section*, 402–8.

Fellegi, Ivan P. 1980. “Approximate Joint Estimation of the Parameters of Multinomial Distributions in the Analysis of Data from Complex Surveys.” *Journal of the American Statistical Association* 75 (370): 261–68.

Fuller, Wayne A. 1975. “Regression Analysis for Sample Survey.” *Sankyha, Series C* 37: 117–32.

———. 2002. “Regression Estimation for Survey Samples (with Discussion).” *Survey Methodology* 28 (1): 5–23.

Gelman, Andrew, and Jennifer Hill. 2019. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Third Edition. Cambridge, UK: Cambridge University Press.

Goldstein, Harvey. 2011. *Multilevel Statistical Models*. Place of Publication: John Wiley; Sons.

Gutiérrez, H. A. 2016. *Estrategias de Muestreo: Diseño de Encuestas y Estimación de Parámetros*. Segunda edición. Ediciones de la U.

Hansen, Morris H., William N. Hurwitz, and William G. Madow. 1953. *Sample Survey Methods and Theory*. Vol. 1 and 2. New York: John Wiley; Sons.

Heeringa, Steven G., Brady T. West, and Patricia A. Berglund. 2017a. *Applied Survey Data Analysis*. Chapman and Hall CRC Statistics in the Social and Behavioral Sciences Series. CRC Press.

Heeringa, Steven G, Brady T West, and Patricia A Berglund. 2017b. *Applied Survey Data Analysis, Second Edition*. *Statistics in the Social and Behavioral Sciences*. 2nd edition. Chapman; Hall - CRC.

Kish, Leslie, and Martin R Frankel. 1974. “Inference from Complex Samples.” *Journal of the Royal Statistical Society, Series B* 36: 1–37.

Kovar, J. G., J. N. K. Rao, and C. F. J. Wu. 1988. “Bootstrap and Other Methods to Measure Errors in Survey Estimates.” *Canadian Journal of Statistics* 16 (Suppl.): 25–45.

Langel, Matti, and Yves Tillé. 2013. “Variance Estimation of the Gini Index: Revisiting a Result Several Times Published: Variance Estimation of the Gini Index.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (2): 521–40. <https://doi.org/10.1111/j.1467-985X.2012.01048.x>.

Lumley, Thomas. 2010. *Complex Surveys: A Guide to Analysis Using r*. *Wiley Series in Survey Methodology*. John Wiley; Sons.

Neter, John, William Wasserman, and Michael H. Kutner. 1996. *Applied Linear Statistical Models*. McGraw-Hill.

Osier, Guillaume. 2009. “Variance Estimation for Complex Indicators of Poverty and Inequality.” *Journal of the European Survey Research Association* 3 (3): 167–95. <http://ojs.ub.uni-konstanz.de/srm/article/view/369>.

Pfeffermann, Danny. 2011. “Modelling of Complex Survey Data: Why Model? Why Is It a Problem? How Can We Approach It?” *Survey Methodology* 37 (2): 115–36.

Pfeffermann, Danny, Chris J Skinner, Don J Holmes, Harvey Goldstein, and Jon Rasbash. 1998. “Weighting for Unequal Selection Probabilities in Multilevel Models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60 (1): 23–40.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: STATA Press.

Rao, J. N. K., and A. J. Scott. 1984. “On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data.” *The Annals of Statistics* 12: 46–60.

Rao, J. N. K., C F J Wu, and K. Yue. 1992. “Some Recent Work on Resampling Methods for Complex Surveys.” *Survey Methodology* 18: 209–17.

Rust, Keith F., Valerie Hsu, and Westat. 2007. “Confidence Intervals for Statistics for Categorical Variables from Complex Samples.” In. <https://api.semanticscholar.org/CorpusID:195852485>.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Shah, B. V., M. M. Holt, and R. F. Folsom. 1977. “Inference about Regression Models from Sample Survey Data.” *Bulletin of the International Statistical Institute* 41 (3): 43–57.

Shah, Babubhai. V, Ralph E Folsom, Lisa LaVange, Sara C Wheeless, Kerrie E Boyle, and Rick L Williams. 1993. “Statistical Methods and Mathematical Algorithms Used in SUDAAN.” Research Triangle Institute.

Skinner, Chris J, Daniell Holt, and Tom M F Smith. 1989. *Analysis of Complex Surveys*. New York: John Wiley; Sons.

Thomas, D. R., and J. N. K. Rao. 1987. “Small-Sample Comparisons of Level and Power for Simple Goodness-of-Fit Statistics Under Cluster Sampling.” *Journal of the American Statistical Association* 82: 630–36.

Wolter, Kirk M. 2007. *Introduction to Variance Estimation*. 2nd ed. Statistics for Social and Behavioral Sciences. Springer.

1. Comisión Económica para América Latina y el Caribe (CEPAL) - [andres.gutierrez@cepal.org](mailto:andres.gutierrez@cepal.org) [↑](#footnote-ref-20)
2. SCIENCE, [pedronsilva@gmail.com](mailto:pedronsilva@gmail.com) [↑](#footnote-ref-22)