CHAPTER 9: ANALYSIS OF HOUSEHOLD SURVEY DATA

Andrés Gutiérrez[1], Pedro Luis do Nascimento Silva[2]

2024-09-23

[1]Comisión Económica para América Latina y el Caribe (CEPAL) - andres.gutierrez@cepal.org

[2]SCIENCE, pedronsilva@gmail.com

Contents

List of Figures

List of Tables

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.

Chapter 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* $Y = \sum_U y_k$ and *mean* $\overline{Y} = \frac{Y}{N}$ of a survey variable y can be estimated by weighted estimators given by $\widehat{Y}_{HT} = \sum_s d_k \, y_k$ and $\overline{y}_H = \frac{\widehat{Y}_{HT}}{\widehat{N}_{HT}} = \frac{\sum_s d_k \; y_k}{\sum_s d_k}$, 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, w_k say, replacing the design weights d_k, for all $k \in s$.

Here $s = \{k_1, \dots, k_n\} \subset U$ denotes the set of units in a sample selected from the population U using a *probability sampling design* $p(s)$, that ensures strictly positive first order inclusion probabilities $\pi_k = Pr(k \in s)$, $\forall \, k \in U$. These inclusion probabilities are assumed known $\forall \, k \in s$, at least to the data producers.

Under the design-based framework and assuming full response, $\widehat{Y}_{HT}$ is unbiased for Y and its sampling variance is given by

$$V_p\left(\widehat{Y}_{HT}\right) = \sum_{k\in U}\sum_{j\in U}\left(\frac{d_k d_j}{d_{kj}} - 1\right) y_k y_j$$

where $d_{kj} = 1/\pi_{kj}$ and $\pi_{kj} = Pr(k, j \in s)$, $\forall\, k, j \in U$. This result assumes that the sampling design $p(s)$ is such that $\pi_{kj} > 0 \;\forall\, k, j \in U$.

Under full response, this variance can be estimated unbiasedly by

$$\widehat{V_p}\left(\widehat{Y}_{HT}\right) = \sum_{k\in s}\sum_{j\in s}\left(d_k d_j - d_{kj}\right) y_k y_j$$

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 π_{kj} (and corresponding pairwise weights d_{kj}) 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 - Binder (1983a). 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 $z_i(\bullet)$ is an *estimating function* evaluated for unit i and θ is a population parameter of interest.

For the case of the population total, take $z_i(\theta) = y_i - \theta/N$. The corresponding population estimation equation is given by $\sum_{i\in U}(y_i - \theta/N) = 0$, and solving for θ gives the population total $\theta_U = \sum_{i\in U} y_i \; = \; Y$. Similarly, take $z_i(\theta) = y_i - \theta$ for the population mean. As a final example, consider the ratio of population totals. Taking $z_i(\theta) = y_i - \theta x_i$, the corresponding population estimation equation is given by $\sum_{i\in U}(y_i - \theta x_i) = 0$. Solving for θ gives the *population ratio* $\theta_U = \sum_{i\in U} y_i / \sum_{i\in U} x_i \; = \; R$.

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* $\sum_{k\in s} d_k\, z_k(\theta) = 0$. Under *probability sampling*, full response and with $d_k = 1/\pi_k$, 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 $\sum_{k\in s} d_k(y_k - \theta) = 0$, and by solving on θ, we recover the Hàjek estimator $\widehat{\theta} = \sum_{k\in s} d_k\, y_k \,/\, \sum_{k\in s} d_k = \overline{y}_H$. In the case of the population ratio, solving $\sum_{k\in s} d_k(y_k - \theta x_k) = 0$ on θ, yields the well-known estimator $\widehat{\theta} = \sum_{k\in s} d_k\, y_k \,/\, \sum_{k\in s} d_k\, x_k = \widehat{R}$.

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

$$V_p(\widehat{\theta}) \doteq \left[J(\theta_U)\right]^{-1} V_p\left[\sum_{k\in s} d_k\, z_k(\theta_U)\right] \left[J(\theta_U)\right]^{-1}$$

where $J(\theta_U) = \sum_{k\in U} \left[\partial z_k(\theta)/\partial\theta\right]_{\theta=\theta_U}$, and θ_U is the solution of the corresponding population estimating equation.

A consistent estimator of this variance is given by:

$$\widehat{V}_p(\widehat{\theta}) = \left[\widehat{J}(\widehat{\theta})\right]^{-1} \widehat{V}_p\left[\sum_{k\in s} d_k\, z_k(\widehat{\theta})\right] \left[\widehat{J}(\widehat{\theta})\right]^{-1}$$

where $\widehat{J}(\widehat{\theta}) = \sum_{k\in s} d_i\, \left[\partial z_k(\theta)/\partial\theta\right]_{\theta=\widehat{\theta}}$.

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 (**?**), 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 n_h PSUs are selected in stratum h, $h = 1, \ldots, H$. Let π_{hi} be the inclusion probability of PSU i stratum h, and by $\widehat{Y}_{hi}$ an unbiased estimator of the total Y_{hi} of the survey variable y for the i-th PSU in stratum h, $h = 1, \ldots, H$. Hence an unbiased estimator of the population total $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi}$ is given by $\widehat{Y}_{UC} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} d_{hi} \widehat{Y}_{hi}$, and the *ultimate cluster* estimator of the corresponding variance is given by:

$$\widehat{V}_{UC}\left(\widehat{Y}_{UC}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(d_{hi} \widehat{Y}_{hi} - \frac{\widehat{Y}_h}{n_h} \right)^2$$

where $d_{hi} = 1/\pi_{hi}$, $\widehat{Y}_h = \sum_{i=1}^{n_h} d_{hi} \widehat{Y}_{hi}$ for $h = 1, \ldots, H$. (See for example, (**?**), 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 (**?**), p. 153.

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

1.4 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.

Chapter 2

Descriptive parameters

When analyzing complex survey data, several descriptive parameters are meaningful and important. For example: poverty and unemployment rates are simple parameters that allow decision-making for governments; also, income distribution can be used to monitor inequality along time.

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.

2.1.1.1 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:

$$\hat{N}_{\omega} = \sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}$$

Similarly, the size estimate in a subpopulation, defined by a dichotomous variable $I(y_i = d)$, which takes the value of one if individual i belongs to category d in the discrete variable, is given by the following expression:

$$\hat{N}^d_\omega = \sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} I(y_i = d)$$

2.2 Means, proportions, and ratios

After conducting the graphical analysis of trends in the continuous survey variables, it is necessary to obtain point estimates of the measured parameters. These estimates can be obtained either generally for the entire population or disaggregated by domains of interest, depending on the research needs. In the context of household surveys, point estimates refer to the estimation of totals, averages, ratios, means, etc. As mentioned by Heeringa et al. (2017), the estimation of totals or averages for a variable of interest in the population, along with the estimation of its variance, has played a crucial role in the development of probability sampling theory. These estimates allow for unbiased and accurate results, providing valuable insights into what is happening in the studied households and enabling informed public policy decision-making.

2.2.1 Total Estimates

Once the sampling design is defined, which was done in the previous section, the estimation processes for the parameters of interest are carried out. For the estimation of totals with complex sampling designs that include stratification ($h = 1, 2, ..., H$) and subsampling in PSUs (assumed to be within stratum h) indexed by $\alpha = 1, 2, ..., a_h$, the estimator for the total can be written as:

$$\hat{y}_\omega = \sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} y_{h\alpha i}$$

Where $n_{h\alpha}$ is the sample size of households or individuals in PSU α of stratum h; a_h is the sample size of PSUs within stratum h; H is the total number of strata in the sampling design. Finally, $y_{h\alpha i}$ and $\omega_{h\alpha i}$ correspond respectively to the observation of the variable of interest and the weight (expansion factor) of element i associated with PSU α within stratum h. The unbiased variance estimator for this total estimator $\hat{y}_\omega$ is:

$$\widehat{var}\left(\hat{y}_\omega\right) = \sum_{h=1}^{H} \frac{a_h}{(a_h - 1)} \left[\sum_{\alpha=1}^{a_h}\left(\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} y_{h\alpha i}\right)^2 - \frac{\left(\sum_{\alpha=1}^{a_h} \omega_{h\alpha i} y_{h\alpha i}\right)^2}{a_h}\right]$$

As can be seen, calculating the total estimate and its estimated variance is complex. However, these calculations can be performed in `R` using the `svytotal` function. The confidence interval is given by the following expression:

$$\hat{y}_{\omega} \pm 1.96 \times \sqrt{\widehat{var}\left(\hat{y}_{\omega}\right)}$$

2.2.2 Estimation of Averages

The estimation of the population mean or average is a very important parameter in household surveys. According to Gutiérrez (2016), an estimator of the population mean can be written as a nonlinear ratio of two estimated finite population totals, as follows:

$$\hat{\bar{y}}_{\omega} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}y_{h\alpha i}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}} = \frac{\hat{y}_{\omega}}{\hat{N}_{\omega}}.$$

It should be noted that if y is a binary variable, the weighted mean estimates the population proportion. On the other hand, since $\hat{\bar{y}}_{\omega}$ is not a linear statistic, there is no closed-form formula for the variance of this estimator. For this reason, resampling methods or Taylor series expansions must be used. In this particular case, using Taylor series, the variance of the estimator is as follows:

$$var\left(\hat{\bar{y}}_{\omega}\right) \doteq \frac{var\left(\hat{y}_{\omega}\right) + \hat{\bar{y}}_{\omega}^{2} \times var\left(\hat{N}_{\omega}\right) - 2 \times \hat{\bar{y}}_{\omega} \times cov\left(\hat{y}_{\omega}, \hat{N}_{\omega}\right)}{\hat{N}_{\omega}^{2}}.$$

As can be observed, calculating the variance estimation involves complex components to compute analytically, such as the covariance between the estimated total and the estimated population size.

2.2.3 Proportion Estimation

The estimation of a proportion for a binary response variable requires a direct extension of the ratio estimator shown in the previous chapter. As mentioned by Heeringa et al. (2017), by recoding the original response categories into a single indicator variable y_i with possible values of 1 and 0 (e.g., yes = 1, no = 0), the estimator for a proportion is defined as follows:

$$\hat{p}_{\omega}^{d} = \frac{\hat{N}_{\omega}^{d}}{\hat{N}_{\omega}} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}\ I(y_i = d)}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}}$$

By applying Taylor linearization to the above estimator, its variance is given by the following expression:

$$var\left(\hat{p}_{\omega}^{d}\right) \doteq \frac{var\left(\hat{N}_{\omega}^{d}\right) + (\hat{p}_{\omega}^{d})^{2} var\left(\hat{N}_{\omega}\right) - 2\,\hat{p}_{\omega}^{d}\, cov\left(\hat{N}_{\omega}^{d}, \hat{N}_{\omega}\right)}{(\hat{N}_{\omega})^{2}}$$

It is common to observe that many statistical packages opt to generate 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 one, the limits of the traditional confidence interval, based on the sampling design, may fall outside the permissible range for proportions. This would have no interpretation due to the nature of the parameter. For this reason, to address this issue, alternative confidence interval estimates based on the sampling design can be used, as proposed by Rust et al. (2007) and Dean and Pagano (2015). Thus, the confidence interval using the $Logit\,(p)$ transformation is given by:

$$IC\left[logit\left(p^{d}\right)\right] = \left\{ ln\left(\frac{\hat{p}_{\omega}^{d}}{1-\hat{p}_{\omega}^{d}}\right) \pm \frac{t_{1-\alpha/2,\,gl} \times se\left(\hat{p}_{\omega}^{d}\right)}{\hat{p}_{\omega}^{d}\left(1-\hat{p}_{\omega}^{d}\right)} \right\} \tag{2.1}$$

Therefore, the confidence interval for p^{d} would be:

$$IC\left(p^{d}\right) = \left\{ \frac{exp\left[ln\left(\frac{\hat{p}_{\omega}^{d}}{1-\hat{p}_{\omega}^{d}}\right) \pm \frac{t_{1-\alpha/2,\,gl} \times se\left(\hat{p}_{\omega}^{d}\right)}{\hat{p}_{\omega}^{d}\left(1-\hat{p}_{\omega}^{d}\right)}\right]}{1 + exp\left[ln\left(\frac{\hat{p}_{\omega}^{d}}{1-\hat{p}_{\omega}^{d}}\right) \pm \frac{t_{1-\alpha/2,\,gl} \times se\left(\hat{p}_{\omega}^{d}\right)}{\hat{p}_{\omega}^{d}\left(1-\hat{p}_{\omega}^{d}\right)}\right]} \right\} \tag{2.2}$$

2.2.4 Relationship Between Variables

In many household survey analyses, it is not sufficient to examine individual variables in isolation. For instance, analyzing the average income of men and women in a country is informative, but comparing the income difference between men and women is crucial for addressing the gender pay gap. This section provides computational tools for estimating ratios and explores hypothesis testing for differences in means, including more complex contrasts.

2.2.4.1 Estimation of Ratios

A particular case of a non-linear function of totals is the population ratio. This is defined as the quotient of two population totals for continuous characteristics of interest. In household surveys, there are times when estimating such a parameter is necessary. For example, estimating the ratio of expenditures to income, the number of men per woman, or the number of pets per household in a specific country.

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:

$$\hat{R}_\omega = \frac{\hat{y}_\omega}{\hat{x}_\omega} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} y_{h\alpha i}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} x_{h\alpha i}}$$

However, because the ratio estimator is a quotient of two estimators (i.e., a quotient of two random variables), calculating the variance of this estimator is not straightforward. To address this, Taylor linearization methods are applied as demonstrated by Gutiérrez (2016).

2.3 Percentiles and inequality measures

In household surveys, it is always necessary to estimate dispersion measures of the studied variables. For example, to understand how disparate incomes are in a given country, which helps inform public policy decisions. Therefore, studying these parameters is crucial. Below is the estimator for the standard deviation:

$$s_\omega\left(y\right) = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i}\left(y_{h\alpha i} - \hat{\bar{y}}_\omega\right)^2}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} - 1}$$

Non-central location measures (percentiles) are calculated to determine characteristic points of the data distribution beyond central values. Key non-central location measures include the median, quartiles, and percentiles. In most household surveys, not only are totals, means, and proportions estimated; for some indicators, it is necessary to estimate other parameters, such as medians and percentiles.

The median is a measure of central tendency that, unlike the mean, is not easily influenced by outliers, and is thus considered a robust measure. The median is the value that divides the population into two equal parts, implying that half of the population's observations fall above the median and the other half fall below.

On the other hand, the estimation of income percentiles in a given country can 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, using an estimation of the cumulative distribution function (CDF) of the population. Specifically, the CDF for a variable y in a finite population of size N is defined as follows:

$$F\left(x\right)=\frac{\sum_{i=1}^{N}I\left(y_{i}\leq x\right)}{N}$$

Where $I\left(y_i \leq x\right)$ is an indicator variable that takes the value 1 if y_i is less than or equal to a specific value x, and 0 otherwise. An estimator of the CDF in a complex sampling design is given by:

$$\hat{F}_{\omega}\left(x\right)=\frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_{h}}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}I\left(y_{i}\leq x\right)}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_{h}}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}}$$

Once the CDF is estimated using the survey design weights, the q-th quantile of a variable y is the smallest value of y such that the CDF is greater than or equal to q. As is well known, the median is the value where the CDF is greater than or equal to 0.5. Thus, the estimated median is the value where the estimated CDF is greater than or equal to 0.5. Following the recommendations of Heeringa et al. (2017), to estimate quantiles, one first considers the order statistics denoted as $y_{(1)},\ldots,y_{(n)}$ and finds the value of j $(j=1,\ldots,n)$ such that:

$$\hat{F}_{\omega}\left(y_{j}\right)\leq q\leq\hat{F}_{\omega}\left(y_{j+1}\right)$$

Thus, the estimation of the q-th quantile $y_{(q)}$ in a complex sampling design is given by:

$$\hat{y}_{(q)}=y_{j}+\frac{q-\hat{F}_{\omega}\left(y_{j}\right)}{\hat{F}_{\omega}\left(y_{j+1}\right)-\hat{F}_{\omega}\left(y_{j}\right)}\left(y_{j+1}-y_{j}\right)$$

For the variance estimation and confidence intervals of quantiles, Kovar et al. (1988) present results from a simulation study where they recommend using the *Balanced Repeated Replication* (BRR) technique. The previously mentioned estimators and procedures for estimating percentiles and their variances are implemented in `R`. Specifically, the median estimation can be done using the function `survey_median`.

2.3.1 Estimation of 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 (G) is the most commonly used indicator for this purpose. The

Gini coefficient ranges from 0 to 1, where $G = 0$ indicates perfect equality in wealth distribution, and higher values reflect increasing inequality.

Following the estimation equation proposed by Binder and Kovacevic (1995), the estimator for the Gini coefficient is given by:

$$\widehat{G}_{\omega}(y) = \frac{2 \times \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} \omega^{*}_{h\alpha i} \widehat{F}_{\omega}^{h\alpha i} y^{h\alpha i} - 1}{\widehat{\bar{y}}_{\omega}}$$

where $\omega^{*}_{h\alpha i}$ is a normalized weight, defined as:

$$\omega^{*}_{h\alpha i} = \frac{\omega_{h\alpha i}}{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i}}$$

In this formula, $\widehat{F}_{h\alpha i\omega}$ represents the estimated cumulative distribution function (CDF) for individual i in cluster α of stratum h, and $\widehat{\bar{y}}_{\omega}$ is the estimated mean.

Osier (2009) and Langel and Tillé (2013) provide important computational details for estimating the variance of this complex estimator.

2.4 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.

Chapter 3

Comparisons and association

Elaborate analyses of household survey data must be adjusted for the complex survey design to account for clustering, stratification, and weighting. This section will introduce the reader on the main methods currently used to compare subgroups and make conclusions based on a valid inferential context.

3.1 Cross-tabulations

Contingency tables and tests of independence are essential tools in household survey analysis, as they help explore relationships between categorical variables. These tables organize population estimates based on two or more characteristics, revealing patterns and associations. Independence tests evaluate whether the variables are related or independent. This analysis is crucial in research and decision-making, as it provides insights into the dependence between factors, influencing strategies based on accurate estimates.

3.1.1 Contingency Tables

In specialized literature, contingency tables are also referred to as cross-tabulations. Generally, a table is assumed to be a two-dimensional array with $r = 1, \dots, R$ rows and $c = 1, \dots, C$ columns. These 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.

For an unexpanded sample, these tables are defined using unweighted frequencies as shown below:

Variable 2	**Variable 1**		
	0	1	**Row Marginal**
0	n^{00}	n^{01}	n^{0+}
1	n^{10}	n^{11}	n^{1+}
Column Marginal	n^{+0}	n^{+1}	n^{++}

For weighted analyses based on an expanded sample, the two-way table presents the population estimate of the frequencies as follows:

Variable 2	**Variable 1**		
	0	1	**Row Marginal**
0	$\hat{N}_{\omega}^{00}$	$\hat{N}_{\omega}^{01}$	$\hat{N}_{\omega}^{0+}$
1	$\hat{N}_{\omega}^{10}$	$\hat{N}_{\omega}^{11}$	$\hat{N}_{\omega}^{1+}$
Column Marginal	$\hat{N}_{\omega}^{+0}$	$\hat{N}_{\omega}^{+1}$	$\hat{N}_{\omega}$

Thus, considering the subscript $i \in (r, c)$ represents the individuals classified in cell (r, c), the estimator for the frequency in this cell is given by the following expression:

$$\hat{N}_{\omega}^{rc} = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i \in (r,c)}^{n_{h\alpha}} \omega_{h\alpha i}$$

The estimators for the other frequencies in the table, including row and column marginals, are defined similarly. The estimated proportions from these weighted sample frequencies are obtained as follows:

$$\hat{p}_{\omega}^{rc} = \frac{\hat{N}_{\omega}^{rc}}{\hat{N}_{\omega}}$$

On the other hand, it is also possible to create tables that report estimates of relative frequencies, or percentages, in the population. This analysis is, of course, conducted using weighted data based on the expanded sample. The two-way table with the population estimate of proportions is presented below:

Variable 2	**Variable 1**		
	0	1	**Row Marginal**
0	$\hat{p}_{\omega}^{00}$	$\hat{p}_{\omega}^{01}$	$\hat{p}_{\omega}^{0+}$
1	$\hat{p}_{\omega}^{10}$	$\hat{p}_{\omega}^{11}$	$\hat{p}_{\omega}^{1+}$
Column Marginal	$\hat{p}_{\omega}^{+0}$	$\hat{p}_{\omega}^{+1}$	$\hat{p}_{\omega}$

In the same way as for absolute frequencies, considering that the subscript $i \in (r, c)$ represents the individuals classified in cell (r, c), the estimator for the proportion associated with this cell is given by the following expression:

$$\hat{p}_{\omega}^{rc} = \frac{\hat{N}_{\omega}^{rc}}{\hat{N}_{\omega}} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i\in(r,c)}^{n_{h\alpha}}\omega_{h\alpha i}}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}}$$

3.2 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.

3.2.1 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 (H_0) is proposed, representing the initial statement that needs to be tested, and an alternative hypothesis (H_1), which is the statement opposing the null hypothesis. This statement may be based on some belief or past experience and will be tested using the evidence gathered from the sample data. If it is suspected that the parameter θ is equal to a particular value θ_0, the possible combinations of hypotheses that can be tested are:

$$\begin{cases}H_0: & \theta=\theta_0\\ H_1: & \theta\neq\theta_0\end{cases}\qquad\begin{cases}H_0: & \theta=\theta_0\\ H_1: & \theta>\theta_0\end{cases}\qquad\begin{cases}H_0: & \theta=\theta_0\\ H_1: & \theta<\theta_0\end{cases}$$

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 j descriptive statistics, as shown below:

$$f\left(\theta_1, \theta_2, ..., \theta_j\right) \;=\; \sum_{j=1}^{J} a_j\theta_j$$

In this case, a_j are known constants. An estimator of this function is given by:

$$\hat{f}_{\omega}\left(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_j\right) = \sum_{j=1}^{J} a_j \hat{\theta}_j \tag{3.1}$$

And its variance is calculated as follows:

$$var\left(\sum_{j=1}^{J} a_j \hat{\theta}_j\right) = \sum_{j=1}^{J} a_j^2 var\left(\hat{\theta}_j\right) + 2 \times \sum_{j=1}^{J-1}\sum_{k>j}^{J} a_j a_k \, cov\left(\hat{\theta}_j, \hat{\theta}_k\right) \tag{3.2}$$

As seen in the variance equation of the estimator, it incorporates the variances of the individual components' estimates, as well as the covariances of the estimated pairs.

Of particular interest is analyzing the difference in population means, which can be written as follows:

$$\bar{y}^1 - \bar{y}^2$$

Where $\bar{y}_1$ is the mean of the first population, for example, the average household income obtained by fathers, and $\bar{y}_2$ is the mean of the second population, which in this example could be the average income of mothers in a household. This parameter can be unbiasedly estimated by:

$$\hat{\bar{y}}_{\omega}^{1} - \hat{\bar{y}}_{\omega}^{2}$$

Where $\hat{\bar{y}}_{\omega}^{i}$ is the sample estimator of $\bar{y}^i$ $(i = 1, 2)$. Considering the parameter of interest in this section, the hypotheses to be tested are as follows:

$$\begin{cases} H_0 : \bar{y}_1 - \bar{y}_2 = 0 \\ H_1 : \bar{y}_1 - \bar{y}_2 \neq 0 \end{cases} \qquad \begin{cases} H_0 : \bar{y}_1 - \bar{y}_2 = 0 \\ H_1 : \bar{y}_1 - \bar{y}_2 > 0 \end{cases} \qquad \begin{cases} H_0 : \bar{y}_1 - \bar{y}_2 = 0 \\ H_1 : \bar{y}_1 - \bar{y}_2 < 0 \end{cases}$$

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

$$t = \frac{\hat{\bar{y}}_{\omega}^{1} - \hat{\bar{y}}_{\omega}^{2}}{se\left(\bar{y}_1 - \bar{y}_2\right)} \sim t_{df}$$

Where:

$$\widehat{se}\left(\hat{\bar{y}}_{\omega}^{1}-\hat{\bar{y}}_{\omega}^{2}\right)=\sqrt{\widehat{var}\left(\hat{\bar{y}}_{\omega}^{1}\right)+\widehat{var}\left(\hat{\bar{y}}_{\omega}^{2}\right)-2\ \widehat{cov}\left(\hat{\bar{y}}_{\omega}^{1},\hat{\bar{y}}_{\omega}^{2}\right)}$$

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

$$\hat{\bar{y}}_{\omega}^{1}-\hat{\bar{y}}_{\omega}^{2}\pm t_{df}\ \widehat{se}\left(\hat{\bar{y}}_{\omega}^{1}-\hat{\bar{y}}_{\omega}^{2}\right)$$

3.2.2 Contrasts

In many cases, it is necessary to compare more than two populations at the same time. For example, comparing the average household incomes across three regions to identify which regions experienced a greater impact on households. In such cases, the difference of means we studied in previous chapters is not sufficient, as it only allows for pairwise comparisons of populations. Therefore, using contrasts becomes a good alternative 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:

$$f = A*\theta = f\left(\theta^{1},\theta^{2},...,\theta^{j}\right) \quad = \quad \sum_{j=1}^{J} a_{j}\theta^{j}$$

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

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 ($f = \bar{y}^{North} - \bar{y}^{South}$). 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:

$$f = A*\theta = 1\times\bar{y}^{North} + (-1)\times\bar{y}^{South} + 0\times\bar{y}^{Center} + 0\times\bar{y}^{West} + 0\times\bar{y}^{East}$$

As can be observed, in this case, the contrast vector is $A = [1, -1, 0, 0, 0]$. Mathematically, the estimator for this specific contrast is defined as:

$$\hat{f}_\omega = A \times \hat{\theta} = [1, -1, 0, 0, 0] \times \begin{bmatrix} \hat{\bar{y}}_\omega^{North} \\ \hat{\bar{y}}_\omega^{South} \\ \hat{\bar{y}}_\omega^{Center} \\ \hat{\bar{y}}_\omega^{West} \\ \hat{\bar{y}}_\omega^{East} \end{bmatrix}$$

Table 3.4: 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:

$$f = A \times \theta = 552.4 - 625.8 = -73.4$$

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:

$$se(\hat{f}_\omega) = se\left(\hat{\bar{y}}_\omega^{North} - \hat{\bar{y}}_\omega^{South}\right) = \sqrt{var\left(\hat{\bar{y}}_\omega^{North}\right) + var\left(\hat{\bar{y}}_\omega^{South}\right) - 2\,cov\left(\hat{\bar{y}}_\omega^{North}, \hat{\bar{y}}_\omega^{South}\right)}$$

Therefore, the estimated standard error for this contrast is:

$$se(\hat{f}_\omega) = \sqrt{3064.715 + 3894.476 - 2 \times 0} = \sqrt{6959.191}$$

3.3 Tests of Independence

Based on the estimated tables, it is possible to perform independence tests to verify whether there is an association between two categorical variables. Two variables are independent if the structural behavior of one variable does not depend on the other, and vice versa. Heeringa et al. (2017) state that, under simple random sampling, two categorical variables are independent if the expected proportion in row r and column c, denoted as π^{rc}, follows the relationship:

$$\pi^{rc} = \frac{n^{r+} \times n^{+c}}{(n^{++})^2}$$

Thus, one way to verify if there is independence between the variables of interest is to directly compare the estimated proportions $\hat{p}_\omega^{rc}$ with the expected proportions π^{rc}. If there is a large difference between them, then the independence hypothesis would not be supported by the collected data. Therefore, the following Rao-Scott χ^2_{RS} statistic (Rao and Scott, 1984b), which follows a chi-square distribution with $(R-1) \times (C-1)$ degrees of freedom, is defined:

$$\chi^2_{RS} = \frac{\chi^2_{Pearson}}{GDEFF} \tag{3.3}$$

Where:

$$\chi^2_{Pearson} = n^{++} \left(\sum_r \sum_c \frac{(\hat{p}_\omega^{rc} - \pi^{rc})^2}{\pi^{rc}} \right)$$

Additionally, $GDEFF$ is an estimate of the generalized design effect by Rao–Scott, defined as:

$$GDEFF = \frac{\sum_r \sum_c (1-p_{rc})\, d^2(p_{rc}) - \sum_r (1-p_{r+})\, d^2(p_{r+}) - \sum_c (1-p_{+c})\, d^2(p_{+c})}{(R-1)(C-1)}$$

As mentioned by Heeringa et al. (2017), it was Fay (1979), along with Fellegi (1980), who began proposing corrections to Pearson's chi-square statistic based on a generalized design effect. Rao and Scott (1984a) later expanded the theory of generalized design effect corrections for these statistical tests, as did Thomas and Rao (1987). The Rao-Scott method requires the calculation of generalized design effects, which are analytically more complex than Fellegi's approach. Rao-Scott corrections are now the standard for analyzing categorical survey data in software systems such as Stata and SAS.

Additionally, Fisher's F-test for independence allows the analysis of whether two dichotomous variables are associated when the observed sample is too small, and the conditions for applying Pearson's χ^2 test are not met. To use this technique, consider the expressions for the estimated probability and Pearson's χ^2 statistic. Based on these, the likelihood ratio statistic is defined as:

$$G^2 = 2 \times n_{++} \times \sum_r \sum_c p_{rc} \times \ln\left(\frac{p_{rc}}{\hat{\pi}_{rc}}\right)$$

where r is the number of rows, and c represents the number of columns, and the test has $(R-1) \times (C-1)$ degrees of freedom. Applying a correction for the generalized design effect, the likelihood ratio statistic is calculated as:

$$G^2_{(R-S)} = \frac{G^2}{GDEFF}$$

Thus, the F-statistic for independence based on Pearson's chi-square test is calculated as follows:

$$F_{R-S,Pearson} = \frac{\chi^2_{R-S}}{[(R-1)(C-1)]} \sim F_{(R-1)(C-1),df}$$

And, the F-statistic for independence based on the likelihood ratio is calculated as:

$$F_{R-S,LRT} = \frac{G^2_{R-S}}{(C-1)} \sim F_{(C-1),df}$$

where C is the number of columns in the cross-tabulation.

3.4 Correlation

To conclude on the degree of association between variables, we show the proper approach to include sampling weights and complex sampling design.

3.5 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.

Chapter 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 Heeringa et al. (2017), the first authors to empirically discuss the impact of complex sampling designs on regression model inferences were Kish and Frankel (1974). Later, Fuller (1975) 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, Shah et al. (1977) 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, Binder (1983b) obtained the sampling distributions of estimators for regression parameters in finite populations and related variance estimators in the context of complex samples. Skinner et al. (1989) studied the properties of variance estimators for regression coefficients under complex sample designs. Later, Fuller (2002) provided a summary of estimation methods for regression models containing information related to complex samples. Finally, Pfeffermann (2011) 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 $y = \beta_0 + \beta_1 x + \varepsilon$, where y represents the dependent variable, x is the independent variable, and β_0 and β_1 are the model parameters. The variable ε is known as the random error of the model and is defined as $\varepsilon = y - \hat{y} = y - \beta_0 + \beta_1 x$.

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:

$$y = x\beta + \varepsilon = \sum_{j=0}^{p} \beta_j x_j + \varepsilon = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

Another way to write the multiple regression model is:

$$y_i = x_i \beta + \varepsilon_i$$

Where, $x_i = \left[1\, x_{1i} \, \ldots \, x_{pi}\right]$ and $\beta^T = \left[\beta_0 \; \beta_1 \; \ldots \; \beta_p\right]$.

The subscript i refers to the sample element or respondent in the dataset. Heeringa et al. (2017) present some considerations for regression models, which are described below:

- $E\left(\varepsilon_i \mid x_i\right) = 0$, meaning that the expected value of the residuals conditioned on the covariates is zero.
- $Var\left(\varepsilon_i \mid x_i\right) = \sigma^2_{y,x}$ (homogeneity of variance), meaning that the variance of the residuals conditioned on the covariates is constant.

- $\varepsilon_i \mid x_i \sim N\left(0,\, \sigma^2_{y,x}\right)$ (normality of errors), meaning that the residuals conditioned on the covariates follow a normal distribution. This property also extends to the response variable y_i.
- $cov\left(\varepsilon_i,\, \varepsilon_j \mid x_i, x_j\right)$ (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 x, as:

$$E\left(y \mid x\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

$$\hat{y} = E\left(y \mid x\right) = E\left(x\beta\right) + E\left(\varepsilon\right) = x\beta + 0 = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Additionally,

$$var\left(y_i \mid x_i\right) = \sigma^2_{y,x}$$

It is also established that:

$$cov\left(y_i, y_j \mid x_i, x_j\right) = 0$$

Thus, the response variable has the following distribution:

$$y_i \sim N\left(x_i\beta, \sigma^2_{y,x}\right)$$

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 n from the N elements of the population, a complete census had been conducted, the finite population regression parameter β_1 could be calculated as follows:

$$\beta_1 = \frac{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2}$$

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), 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 β_1 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:

$$\hat{\beta}_1 = \frac{\sum_h^H \sum_\alpha^{a_h} \sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} \left(y_{h\alpha i} - \hat{\bar{y}}_\omega\right)\left(x_{h\alpha i} - \hat{\bar{x}}_\omega\right)}{\sum_h^H \sum_\alpha^{a_h} \sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} \left(x_{h\alpha i} - \hat{\bar{x}}_\omega\right)^2}$$

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

$$var\left(\hat{\beta}_1\right) = \frac{var\left(\hat{t}_{xy}\right) + \hat{\beta}_1^2 var\left(\hat{t}_{x^2}\right) - 2\hat{\beta}_1 cov\left(\hat{t}_{xy}, \hat{t}_{x^2}\right)}{\left(\hat{t}_{x^2}\right)^2}$$

As a generalization, according to Kish and Frankel (1974), the variance estimation of coefficients in a multiple linear regression model requires weighted totals for the squares and cross-products of all combinations of y and $x = \{1, x_1, \ldots, x_p\}$. Below is the estimation of these variances:

$$var\left(\hat{\beta}\right) = \hat{\Sigma}\left(\hat{\beta}\right) = \left[\begin{array}{cccc} var\left(\hat{\beta}_0\right) & cov\left(\hat{\beta}_0, \hat{\beta}_1\right) & \cdots & cov\left(\hat{\beta}_0, \hat{\beta}_p\right) \\ cov\left(\hat{\beta}_0, \hat{\beta}_1\right) & var\left(\hat{\beta}_1\right) & \cdots & cov\left(\hat{\beta}_1, \hat{\beta}_p\right) \\ \vdots & \vdots & \ddots & \vdots \\ cov\left(\hat{\beta}_0, \hat{\beta}_p\right) & cov\left(\hat{\beta}_1, \hat{\beta}_p\right) & \cdots & var\left(\hat{\beta}_p\right) \end{array}\right]$$

4.3.3 The Pfeffermann Weighting Approach

Heeringa et al. (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), 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 (R^2), 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:

$$R^2 = 1 - \frac{SSE}{SST}$$

Where:

- $SST = \sum_{i=1}^{N}(y_i - \bar{y})^2$: This is the total sum of squares, representing the total variability in the dependent variable.
- $SSE = \sum_{i=1}^{N}(y_i - x_i\beta)^2$: 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 R^2 is given by:

$$\widehat{R}^2_\omega = 1 - \frac{\widehat{SSE}_\omega}{\widehat{SST}_\omega}$$

Where $\widehat{SSE}_\omega$ is the weighted sum of squared errors, defined as:

$$\widehat{SSE}_{\omega} = \sum_{h}^{H} \sum_{\alpha}^{a_h} \sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} \left(y_{h\alpha i} - x_{h\alpha i} \hat{\beta} \right)^2$$

This estimator adjusts the R^2 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 (Heeringa et al., 2017) are defined as follows:

$$r_{p_i} = \left(y_i - \mu_i \left(\hat{\beta}_{\omega} \right) \right) \sqrt{\frac{\omega_i}{V \left(\hat{\mu}_i \right)}}$$

Where μ_i is the expected value of y_i, and ω_i is the survey weight for the i-th individual in the complex sample design. Finally, $V(\mu_i)$ 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 $\hat{y}$ or the model residuals against X_i. 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. $D_f Beta_{(i)}$ **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 B_j if $\mid D_f Betas_{(i)j} \mid \geq \frac{z}{\sqrt{n}}$ with $z = 2$. Alternatively, $t_{0.025,n-p}/\sqrt{n}$ can be used, where $t_{0.025,n-p}$ is the 97.5th percentile.
3. $D_f Fits_{(i)}$ **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 $\mid DfFits\,(i) \mid \geq z\sqrt{\frac{p}{n}}$ with $z = 2$.

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 y—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:

$$t = \frac{\hat{\beta}_k - \beta_k}{se\left(\hat{\beta}_k\right)} \sim t_{n-p}$$

Where p is the number of model parameters and n is the sample size of the survey. The test statistic above evaluates the hypotheses $H_0 : \beta_k = 0$ versus the alternative $H_1 : \beta_k \neq 0$. Similarly, a confidence interval of $(1-\alpha) \times 100\%$ for β_k can be constructed, as follows:

$$\hat{\beta}_k \pm t_{1-\frac{\alpha}{2},\,df}\, se\left(\hat{\beta}_k\right)$$

Where the degrees of freedom (df) 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 $\left(df = \sum_h a_h - H\right)$.

4.3.4.5 Estimation and Prediction

According to Neter et al. (1996), 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:

$$\hat{E}(y_i \mid x_{obs,i}) = x_{obs,i}\hat{\beta}$$

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

$$\hat{E}(y_i \mid x_{obs,i}) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

The variance of the estimation is calculated as follows:

$$var\left(\hat{E}\left(y_i \mid x_{obs,i}\right)\right) = x'_{obs,i} cov\left(\hat{\beta}\right) x_{obs,i}$$

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 $\beta_0 + \beta_1 x$ 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:

$$\pi(\mathbf{x}) = Pr(y = 1|\mathbf{x}) = \frac{\exp\{\mathbf{x}'\beta\}}{1 + \exp\{\mathbf{x}'\beta\}}$$

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:

$$g(\mathbf{x}) = logit(\pi(\mathbf{x})) = ln\left\{\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right\} = \mathbf{x}'\beta$$

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:

$$\hat{\pi}(\mathbf{x}) = \frac{\exp\{\mathbf{x}'\hat{\beta}\}}{1+\exp\{\mathbf{x}'\hat{\beta}\}}$$

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

$$var\left(\hat{B}\right) = J^{-1}var\left(S\left(\hat{B}\right)\right)J^{-1}$$

Where:

$$S\left(B\right) = \sum_{h}\sum_{a}\sum_{i} w_{hai}D_{hai}^{t}\left[\left(\pi_{hai}\left(B\right)\right)\left(1-\pi_{hai}\left(B\right)\right)\right]^{-1}\left(y_{hai}-\pi_{hai}\left(B\right)\right) = 0$$

and,

$$D_{hai} = \frac{\delta\left(\pi_{hai}\left(B\right)\right)}{\delta B_{j}}$$

Where $j = 0, \ldots, p$. Since the model uses a logarithmic link, confidence intervals are constructed by applying the exponential function to each parameter:

$$\hat{\psi} = \exp\left(\hat{B}_1\right)$$

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

$$CI\left(\psi\right)=\exp\left(\hat{B}_j\pm t_{df,1-\frac{\alpha}{2}}se\left(\hat{B}_j\right)\right)$$

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), the exponentiated logistic regression coefficients can be interpreted as odds ratios. If two outcomes present probabilities of $(\pi, 1-\pi)$, then $\pi/(1-\pi)$ 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:

$$G=-2\ln\left[\frac{L\left(\hat{\beta}\right)_{reduced}}{L\left(\hat{\beta}\right)_{full}}\right]$$

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.

Chapter 5

Data visualization

In this section we delve deeper on how to present the results of the analysis of household surveys using graphics. This part is important because household surveys estimates are subject to error and the researcher should face this challenge on presenting not only point estimates, but also standard errors.

5.1 Weighted Histograms

To visualize the distribution of continuous variables, adjusted for survey weights.

5.2 Bar Charts

To display the distribution of categorical variables with standard errors.

5.3 Box Plots

To show the distribution of continuous variables, including measures of central tendency, variability, and outliers, across different groups or strata.

5.4 Scatter Plots

To explore the relationship between two continuous variables, potentially revealing patterns or trends in survey data.

5.5 Maps

To display the behavior of the interest variable across geographical domains.

5.6 NSO – Practical example

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

Chapter 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), a classic reference for multilevel data analysis that discusses the use of hierarchical models in various contexts, including household surveys; Gelman and Hill (2019), which offers an accessible introduction to the theory and practice of hierarchical models; Rabe-Hesketh and Skrondal (2012), a practical guide for analyzing multilevel and longitudinal data using statistical software such as Stata; and Browne and Draper (2006), 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:

$$y_{ij} = \beta_{0j} + \epsilon_{ij}$$

Here, y_{ij} represents the income of individual i in stratum j; β_{0j} is the intercept in stratum j; ϵ_{ij} is the residual for individual i in stratum j; γ_{00} is the overall intercept, and τ_{0j} is the random effect for the intercept. For this model, it is assumed that:

$$\tau_{0j} \sim N(0, \sigma^2_\tau)$$

Additionally,

$$\epsilon_{ij} \sim N(0, \sigma^2_\epsilon).$$

According to Cai (2013), 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) and Asparouhov (2006) 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:

$$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2}$$

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:

$$Income_{ij} = \beta_0 + \beta_{1j} Spending_{ij} + \epsilon_{ij}$$

where β_{1j} is given as

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Stratum_j + \tau_{1j}$$

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:

$$Income_{ij} = \beta_{0j} + \beta_{1j} Spending_{ij} + \epsilon_{ij}$$

where the intercept and slope are modeled as:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Stratum_j + \tau_{0j}$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Stratum_j + \tau_{1j}$$

In this model, β_{0j} and β_{1j} 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:

$$\ln(\pi_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY},$$

where π_{ijk} is the conditional probability of occurrence for the combination of categories i and j in the categorical variables X and Y, respectively; μ is the intercept representing the logarithm of the reference probability; λ_i^X and λ_j^Y are the main effects associated with categories i and j in the variables X and Y, respectively; and finally, λ_{ij}^{XY} is the interaction effect between categories i and j 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 X and Y.

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:

$$Q_{wald} = \hat{Y}^t \left(H\hat{V}\left(\hat{N}\right)H^t\right)^{-1}\hat{Y}$$

where

$$\hat{Y} = \left(\hat{N} - E\right)$$

is a vector of $R \times C$ differences between observed and expected cell counts, that is, $\hat{N}_{rc} - E_{rc}$. The matrix $H\hat{V}\left(\hat{N}\right)H^t$ 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, $\hat{V}\left(\hat{N}\right)$, is estimated using resampling methods or Taylor approximation. The matrix H is the inverse of the matrix J given by:

$$J = -\left[\frac{\delta^2 \ln PL\left(B\right)}{\delta^2 B}\right]\Bigg|_{B=\hat{B}}$$

Under the null hypothesis of independence, the Wald statistic follows a chi-squared distribution with $(R-1) \times (C-1)$ degrees of freedom:

$$Q_{wald} \sim \chi^2_{(R-1)(C-1)}$$

The F transformation of the Wald statistic is:

$$F_{wald} = Q_{wald} \times \frac{df - (R-1)(C-1) + 1}{(R-1)(C-1)df} \sim F_{(R-1)(C-1), df-(R-1)(C-1)+1}$$

References

Bibliography

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods*, 35(3):439–460.

Binder, D. A. (1983a). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51(3):279–292.

Binder, D. A. (1983b). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51:279–292.

Binder, D. A. and Kovacevic, M. S. (1995). Estimating some measures of income inequality from survey data: An application of the estimating equations approach. *Survey Methodology*, 21(2):137–145.

Browne, W. J. and Draper, D. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514.

Cai, T. (2013). Investigation of ways to handle sampling weights for multilevel model analyses. *Sociological Methodology*, 43(1):178–219.

Dean, N. and Pagano, M. (2015). Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys. *Journal of Survey Statistics and Methodology*, 3(4):484–503.

Fay, R. E. (1979). On adjusting the pearson chi-square statistic for cluster sampling. In *Proceedings of the Social Statistics Section, American Statistical Association*, pages 402–405, Washington, DC.

Fellegi, I. P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association*, 75:261–268.

Fuller, W. A. (1975). Regression analysis for sample survey. *Sankyha, Series C*, 37:117–132.

Fuller, W. A. (2002). Regression estimation for survey samples (with discussion). *Survey Methodology*, 28(1):5–23.

Gelman, A. and Hill, J. (2019). *Data Analysis Using Regression and Multi-*

level/Hierarchical Models. Cambridge University Press, Cambridge, UK, third edition edition.

Goldstein, H. (2011). *Multilevel Statistical Models.* John Wiley & Sons, Place of Publication.

Gutiérrez, H. A. (2016). *Estrategias de muestreo: diseño de encuestas y estimación de parámetros.* Ediciones de la U, segunda edición edition. Google-Books-ID: Ul-VmE5pkRwIC.

Heeringa, S. G., West, B. T., and Berglund, P. A. (2017). *Applied survey data analysis.* Chapman and Hall CRC statistics in the social and behavioral sciences series. CRC Press.

Kish, L. and Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, 36:1–37.

Kovar, J. G., Rao, J. N. K., and Wu, C. F. J. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, 16(Suppl.):25–45.

Langel, M. and Tillé, Y. (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):521–540.

Neter, J., Wasserman, W., and Kutner, M. H. (1996). *Applied Linear Statistical Models.* McGraw-Hill.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality. *Journal of the European Survey Research Association*, 3(3):167–195.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? why is it a problem? how can we approach it? *Survey Methodology*, 37(2):115–136.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (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, S. and Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata.* STATA Press, College Station, TX.

Rao, J. N. K. and Scott, A. J. (1984a). On chi-squared test for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12:46–60.

Rao, J. N. K. and Scott, A. J. (1984b). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics*, 12(1):46–60.

Rust, K. F., Hsu, V., and Westat (2007). Confidence intervals for statistics for categorical variables from complex samples.

Shah, B. V., Holt, M. M., and Folsom, R. F. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, 41(3):43–57.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989). *Analysis of Complex Surveys*. John Wiley & Sons, New York.

Thomas, D. R. and Rao, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. *Journal of the American Statistical Association*, 82:630–636.

Wolter, K. M. (2007). *Introduction to variance estimation*. Statistics for social and behavioral sciences. Springer, 2nd ed edition.