Banco Central de Chile
Documentos de Trabajo

Central Bank of Chile
Working Papers

N° 632

Junio 2011

COMPUTING POPULATION WEIGHTS FOR THE EFH SURVEY

Carlos Madeira

BANCO CENTRAL DE CHILE

CENTRAL BANK OF CHILE

COMPUTING POPULATION WEIGHTS FOR THE EFH SURVEY

Carlos Madeira
Banco Central de Chile

Abstract

This paper explains the trade-off between bias and variance in the choice of expansion factors for the Chilean Financial Household Survey (EFH). I outline several alternatives for the expansion factors assigned to each household in the EFH. The alternatives are based on a full post-stratification procedure using as strata different groups of Chile's geographical regions, the wealth of each town, and the income level of each household. I find that expansion factors based on a small number of strata can accurately represent the age, education and income distribution of Chile with little bias and variance involved.

Resumen

Este trabajo explica las ventajas y desventajas en la elección de factores de expansión para la Encuesta Financiera de Hogares (EFH) del Banco Central de Chile. En general, la construcción de factores de expansión (o pesos poblacionales) con más estratos reduce el sesgo de las estimaciones, pero al costo de un posible aumento de la varianza cuando el número de estratos es demasiado alto. Se exponen diversas alternativas para los factores de expansión atribuidos a cada hogar de la EFH. Las alternativas se basan en un procedimiento de pos-estratificación completo, usando como estratos agregaciones de regiones chilenas, el ingreso de cada comuna, y el ingreso del hogar. Los resultados muestran que factores de expansión construidos con un número bajo de estratos representan de forma precisa la edad, educación, y distribución del ingreso de Chile, con poco sesgo y poca varianza.

1 Introduction

The Encuesta Financiera de Hogares (EFH) is a survey of financial assets and debts of Chilean households that has been implemented by the Central Bank of Chile and Center Microdatos of the University of Chile since 2007. There are now three waves of the EFH project. The first wave of the EFH was implemented in 2007 and collected information from 4021 urban households at the Chilean national level. The second and third waves were implemented respectively in 2008 and 2009, covering a panel of 1207 urban households in the Metropolitan Region of Chile.

The EFH survey seeks to obtain a rigorous estimation of the financial risk incurred by Chilean households. Research applications range include studies of the distribution of households' assets and debt (Banco Central de Chile, 2009, 2010), effects of unemployment cycles on debt risk (Fuenzalida and Ruiz-Tagle, 2009), borrowing constraints (Ruiz-Tagle and Vella, 2010), credit default (Alfaro, Gallardo and Stein, 2010), estimates of the earthquake effects on households' finances (Banco Central de Chile, 2010), and changes in real estate prices (Sagner, 2009).

However, in Chile - as for other countries such as Italy, the UK, Spain, Netherlands, and the USA - most of the complex financial relationships are concentrated in a small number of upper income households. The need to cover this small number of households requires the EFH to cover more upper income households based on a complex sample design. This work examines the challenge of making the sample of 4021 EFH households representative of the 3.850.000 Chilean households at the urban national level. I explain briefly how these "statistical representativity" procedures are built and their importance for data analysis. Finally, I evaluate a set of several "statistical population" methods and suggest one of these to be used in the public EFH dataset provided by the Central Bank of Chile.

This paper is organized as follows. Section 2 explains briefly what expansion factors mean, while section3 considers the most general methods of computing expansion factors and how they can be used to obtain representative statistics of a survey dataset.

In section 4 I explain how the Center Microdatos built the EFH random sample and the special problems it involves. I also outline the basic structure of the expansion factors published by the Center Microdatos and its main flaws. The sample design of the EFH survey allows me to consider several methods of estimating the expansion factors of each EFH household. In section 5 I consider

a set of post-stratification procedures based on three types of household characteristics: 1) their regional area of residence, 2) the income distribution of their urban area (or sub-area) of residence, and 3) the household's income stratum at the Chilean national level. This large set of procedures are then estimated using data from the Chilean population survey (CASEN) in 2003 and 2006. When applied to the survey data expansion factors can be used to compute population weights and obtain representative statistics of any variable at the Chilean national level. Expansion factors can also be correctly used to obtain representative statistics for some of Chile's regions and demographic groups.

Section 6 shows how these expansion factors when applied to the EFH 2007 sample allow us to obtain accurate statistics for the income distribution and house ownership among Chilean households. I compare several types of expansion factors in terms of how well they reproduce Chilean national statistics from other datasets. The expansion factors that represent both more accurate statistics and less variance are the ones based on simple information, such as the broad regional area of residence and the income placement of the urban area of residence in terms of the Chilean national population. I then show that the estimated expansion factors are also reasonably accurate in reproducing key statistics of the distribution of education and age in Chilean households.

I provide therefore a clear recommendation of which type of expansion factor future researchers should use in most of their applications. At the same time I have written an easy software procedure that allows future researchers to use other types of expansion factors for projects involving smaller regions of Chile and to examine how sensitive their results are to changes in population weights.

Finally, I present a summary of the conclusions in Section 7 and provide broad lines for future improvements and research.

2 How to understand expansion factors and population weights

2.1 What are expansion factors?

Each household in the a survey represents a different number of statistically equivalent households. Survey researchers call this the statistical representativity or "expansion factor" of each observation. For example, if the survey directors go to city A and interview one household in every 1000, then

the expansion factor of city A households should be 1000. Also, the sum of the expansion factors for all observations in the sample dataset should be equivalent to the target population universe of the survey. Therefore, if for instance the target population includes all the households of Chile, then the sum of the expansion factors for all observations in the dataset should be equivalent to the household population of Chile.

Many researchers erroneously think one can obtain adequate statistics from a survey by treating all observations with the same importance. However, if surveys collect respondents with unequal probabilities or in a non-random way, then ignoring the expansion factors implies the estimated statistics cannot be understood as representative of the target population (Neyman, 1934).

Suppose A and B are cities of equal size, but the interview rate of A households is only 1 in 1000 while in B it is 1 in 500. In the survey there are 30% of A people unemployed and 15% of unemployed B people. Therefore if a researcher is ignoring population weights, he will state that the unemployment rate is $30\%\frac{1}{3} + 15\%\frac{2}{3} = 20\%$. This statistic however is valid only for the set of interviewed people. In fact the real statistic for the target population is $30\%\frac{1000}{1000+2\times500} + 15\%\frac{2\times500}{1000+2\times500} = 25\%$.

A big problem is that the "target population" universe is often unknown. There is no accurate dataset listing all the Chilean households and where they live. This is true of Chile and of most countries in the world. Therefore the size of the target population and its characteristics is usually measured by using another dataset. This dataset is quite often another population survey which was obtained with a larger sample and a higher coverage, making it therefore a reliable approximation of the target population.

In the case of the United States these large population datasets are usually the US Census and the Current Population Survey (CPS). In the case of Chile there is also a Chilean Household Census - published by the Chilean Institute of National Statistics (INE) every 10 years - and the Encuesta de Caracterización Socieconómica Nacional (CASEN) survey. The CASEN survey has the advantages of measuring more variables than the Census (such as income) and of being implemented more frequently. The CASEN survey also has high response rates (90%), which limits errors from sample bias. For these reasons, the CASEN 2003 and 2006 datasets were judged to be the most suitable source of information to estimate the statistical representativity of the EFH.

In this case the CASEN data surveys a sample of 42000 Chilean households living in urban

areas and provides therefore an accurate representation of most regions and cities of Chile. This was the reason why the Center Microdatos chose the CASEN classification of urban areas as a guideline for the sample design of the first EFH 2007 wave.

3   General Methods to compute expansion factors

3.1   The inverse selection probability method

There are two main classes of methods to compute expansion factors. The first family is known as the inverse selection design probability method (Lohr, 2009). If every observation in the universe is selected with positive probability, this method is an accurate way of making the smaller sample a statistical representation of the target population. Therefore if observation i in the sample was selected with probability p_i, then the expansion factor f_i is:

1) $f_i = (p_i)^{-1}$.

A standard example of this method is a random sample where all n sample elements are chosen with equal probability from a population of size N and therefore $f_i = \dfrac{N}{n}$. If the selection design probabilities, p_i, are well specified, then the researcher can obtain consistent statistics for any variable in the survey (Neyman, 1934).

Some surveys design their sample in order to minimize the variance of estimation of a certain statistic (Lavallée and Hidiroglou, 1988, Hedlin, 2000, Rivest, 2002, Kadane, 2005, Horgan, 2006, Fabrizi and Trivisano, 2007). It is important to notice, however, that the use of the correct expansion factors still results in consistent estimates for statistics of any other variable in the survey (Neyman, 1934, Kish, 1992). Therefore a specific sample design to fit a certain statistic only raises issues of optimality and efficiency in the estimates of other statistics, but does not create problems of asymptotic consistency.

The inverse probability method is usually applied when stratified sample probabilities are carefully chosen before the survey and non-response rates are low. For several surveys, however, non-response rates are high and therefore the selection probability of each unit is unknown. This

leads to ill-defined statistics if inverse selection probabilities are used. Suppose no households in urban area X were interviewed. Then the selection probability of these households post-survey is 0 and that corresponds to a factor of infinity. Also, in some cases selection design probabilities are designed to minimize survey costs of units that are harder to reach and not to optimize the researcher's statistical work (Tillé, 2006). This implies that the sample design probabilities will give too much importance to a few observations and increase substantially the variance of the statistical methods applied to the survey sample (Kish, 1992).

For these reasons researchers often use alternative methods to estimate the expansion factors for a survey dataset (Kish, 1992). In this case the survey researcher will seek something less rigorous than finding the true sample selection probabilities and will require only that the survey weights replicate the known population totals of some auxiliary variables. The ideal alternatives seek to approximate the true inverse selection probabilities by taking into account the most important elements of household heterogeneity, while reducing the effects of sample selection variables less relevant for the outcomes (Kish and Frankel, 1974).

3.2 Post-stratification methods

The second class of methods is usually known as post-stratification methods (Kish and Frankel, 1974, Lohr, 2009) and consists in the following: 1) all the observations in both the sample and the target population are classified in several groups (or strata) according to their characteristics (it is assumed that all the strata are represented both in the sample and target population), 2) the expansion factor of each group in the sample is chosen in a way to match exactly the number of people of that group in the target population. The basic philosophy of post-stratification is the same as for the inverse selection probability method. If the strata are seen as an adequate representation of the several types of observations present in the target population, then the post-stratification gives the inverse sample selection probability of each type (Estevao and Särndal, 2000).

Assume the post-stratified method classifies observations in groups $g = 1, .., G$ and all groups are represented both in the target population and the survey sample. Then the expansion factor for each observation $i \in g(i)$ is:

2) $f_i = \dfrac{N_{g(i)}}{n_{g(i)}}$,

where $N_{g(i)}$ is the size of group $g(i)$ in the target population and $n_{g(i)}$ is the size of the group $g(i)$ in the survey sample.

If the groups in the post-stratification method include all the relevant variables that affect the selection probability of the sample units, then this method is equivalent to the inverse selection probability method (Kott, 2006a, Lohr, 2009). In this sense, one can view a good post-stratification method as one that is a reliable approximation for the sample selection probability. For this reason, the choice of groups and variables for the post-stratification procedure should be driven by an explicit theory of sample selection based on the survey design and non-response (Groves and Couper, 1995).

3.3 Differences between post-stratification methods

Several post-stratification methods exist (Deville and Sarndal, 1992, Deville, Sarndal, and Sautory, 1993, Rao, 2005, Kott, 2006a, 2006b). Most of them can, however, be seen as members of two different families: 1) full post-stratification methods (Lohr, 2009); 2) incomplete post-stratification methods and generalized calibration raking procedures.

Full post-stratification methods work by matching the population of each group (or strata) in the survey sample with their totals from other surveys, as described in equation 2). Partial post-stratification is obtained by multiplying the expansion factors that match different variables. Suppose $g_h = 1, .., G_h$ represents the strata membership of observation i in terms of variable h. Then the expansion factors of a partial post-stratification procedure using variables $h = 1, .., K$ would be:

3) $$f_i = \prod_{h=1}^{K} \frac{N_{g_h(i)}}{n_{g_h(i)}}.$$

Full post-stratification would require creating mutually exclusive groups for each combination of all the K variables, i.e., $g^* = \{g_1 \times ... \times g_K\}$, and then applying equation 2) as $f_i = \dfrac{N_{g*(i)}}{n_{g*(i)}}$. Partial post-stratification is an easier and simpler approach than full-post-stratification. However, both methods are only equivalent if all the K variables are independent of each other. If the variables are not independent of each other, then partial post-stratification does not match well any of the statistics it is adjusting for. In this case partial post-stratification can be quite imprecise and

create both bias and high variance in the expansion factors (Kozak and Verma, 2006). Therefore full post-stratification is always preferable when feasible.

Generalized raking procedures is a more complex family of post-stratification procedures than the ones described in equation 3). These procedures work by minimizing a distance function of several sample statistics with known statistics of the target population. Such procedures as the CALMAR macro are now widely used by national statistics offices in France, Canada, United Kingdom, Italy, Luxemburg, and Spain, among several other countries (Estevao, Hidiroglou, and Sarndal, 1995, Bover, 2004, Faiella and Gambacorta, 2007, Crockett, 2008).

An important difference between the two methods is that the strata population method requires that every person in the survey be classified in mutually exclusive groups. Calibration methods, however, can be used to create matches with statistics that are known for the whole population, but that are not necessarily specified for each strata in the sample. This is the case because not all surveys collect the same information. For example, Census data collect only demographic information, while Employment Surveys may collect only income and occupation information. Therefore the researcher may be unable to classify each observation in mutually exclusive groups. For this reason, several surveys first create expansion factors using full post-stratification based on Census data first and then make partial post-stratification adjustments for other statistics (Deville, Sarndal, and Sautory, 1993). In this case, it is important that the first-stage of full post-stratification results in expansion factors with a low variance, since the second-stage process of matching more statistics usually increases the variance of the expansion factors even more (Kott, 2006b).

3.4 Choice of the number of strata

Another problem is how many groups or variables to choose for post-stratification (Wu, 2002, Lohr, 2009). In general there is a bias versus variance trade-off. Creating more strata or adjusting for more variables helps the expansion factors match more statistics and reduces the sample bias. However, adjusting for more strata can make the expansion factors more volatile and therefore increases the variance of the sample estimators (Kish and Frankel, 1974).

In the case of expansion factors the bias-variance trade-off is often very significant. Suppose we have a dataset where all individuals have the same expansion factor with value 1000. After adjusting the expansion factors to match statistics $x_1, ..., x_K$ their dispersion increases to values between 10 and 100000. In this case the researcher will match the chosen K statistics perfectly, but the other statistics obtained from the dataset will be very sensitive to the few observations who represent 100000 households and therefore a few observations will completely determine the estimation.

For this reason the researcher should choose to match only the most relevant statistics for his survey (Kott, 2006a). Strata should be chosen in a way that individuals inside a group are as homogeneous as possible and new strata should only be added if the individuals belonging to those strata are significant different from the other strata (Lohr, 2009). Excluding strata which are very different creates bias, but including irrelevant strata increases variance. Little and Vartivarian (2005) show that if only the relevant variables are included for the post-stratification adjustment, then it is possible to reduce both the bias and variance of the estimators.

3.5 Computing representative statistics

Mean statistics of any variable x can be estimated by using standardized expansion factors (Horvitz and Thompson, 1952):

4) $\bar{x} = \sum_i (\frac{f_i}{\sum_v f_v}) x_i.$

It is standard notation to denote the standardized expansion factor as the population weight of each observation i, $w_i = (\frac{f_i}{\sum_v f_v})$. This notation also creates an analogy between the expansion factors, population weights, and other statistical methods such as weighted moments, weighted likelihood or weighted least squares. The basic principle underneath all concepts is the same. Also, mean statistics of x can be obtained for any sub-group of individuals:

5) $\bar{x}_G = \sum_{i \in G} (\frac{f_i}{\sum_{v \in G} f_v}) x_i.$

Quantile statistics ($q_\alpha \equiv \arg\min_x \Pr(x_i \le x) \ge \alpha$) can also be computed in a similar way:

6) $\hat{q}_\alpha \equiv \arg\min_x \sum_i (\frac{f_i}{\sum_v f_v}) 1(x_i \le x) \ge \alpha.$

7) $\hat{q}_{\alpha,G} \equiv \arg\min_x \sum_{i \in G} (\frac{f_i}{\sum_{v \in G} f_v}) 1(x_i \le x) \ge \alpha.$

Standard-errors and t-statistics for estimators using expansion factors are often difficult to derive or involve long expressions (Kott, 2006a, Lohr, 2009). Some statistical software programs provide adequate methods for estimating standard-errors in surveys with expansion factors. However, if bootstrap is a method that is valid for the specified econometric estimator, then bootstrap is also guaranteed to work with expansion factors (Funaoka, Saigo, Sitter, and Toida, 2006). Therefore bootstrap is a safe option for the researcher in most applications (Rao and Wu, 1988).

Also, as explained before, in most cases the expansion factors of each observation are the result of statistical methods and not a factor that is known with certainty. The researcher, however, can account for model and sample uncertainty of the expansion factors when estimating his model. This is easily achieved by replicating different expansion factors for each bootstrap sample and obtaining model estimates for each bootstrap draw of the expansion factors (Brownstone and Chu, 1994, Brownstone, 1997).

3.6 Consistency of regresson models

Unbiasedness of the weighted estimates of a model is not feasible, since weighted estimators involve a multiple of two random variables (the random observations and the weights). However, weighted estimates are approximately unbiased, presenting a bias of order $N - 1$, negligible in large samples (Wolter, 1985).

It is also possible to estimate regressions models without expansion factors and to get consistent estimates of the coefficients. However, this result is only valid if the regression model is taken as the real model and the sample selection design is ignorable, which is rarely a realistic assumption. Also, several statistical studies (Little, 1981, Nathan and Smith, 1989, Kott, 1991, Särndal, Swensson, and Wretman, 1992) show that weighted estimates are more robust to model mis-specification, omitted variable problems, and to the heteroskedasticity that normally characterizes sample survey data. Wooldridge (1999) proves that, when the selection probabilities are based on exogenous

variables, traditional M estimators and their variances are consistent even without expansion factors. However, the asymptotic variance of the unweighted M estimator is only optimal if a generalized information matrix condition is met. Basically, this would mean the unweighted M estimator would only be optimal under correctly specified MLE models and homecedastic error terms. Since in most realistic applications errors are heterocedastic, this would be an argument for using the M-estimator with population weights.

In most cases the researcher understands the regression model as a first-order approximation of the true model and not as an absolute truth. In this case omitting variables which are related to the expansion factors may worsen the problems of mis-specified models. Reiter, Raghunathan, and Kinney (2006) show that in the case of multiple imputation problems, ignoring expansion factors can have a strong effect.

Even if the regression model is taken to be the real world model, the researcher will only be getting valid estimates for $E(Y \mid X)$. If the researcher wants to obtain inferences for the whole population, then expansion factors are needed to get valid estimates of $\Pr(X)$ and therefore estimate $E(Y) = \int E(Y \mid X) \Pr(X) \partial X$. This problem is a big concern for any welfare analysis. Therefore the use of expansion factors is highly advised to estimate models, especially if the researcher intends to make statements about the values for the whole target population.

It is also important to note that Maximum Likelihood standard-errors without expansion factors are invalid, because the sample selection design makes the estimator inefficient. Therefore Maximum Likelihood model variances should be obtained using the Huber-White robust variance matrix.

4   Classification of urban areas according to population wealth

I start by explaining how the EFH attempts to survey individuals of higher wealth strata. In general, it is not feasible to observe the wealth level of households prior to the interviews (Kalton, 2009), therefore the EFH builds its survey by sampling urban areas with a greater concentration of wealthier households. This led to the use of 2 distinct samples that were jointly used to form the EFH 2007 dataset: 1) a sample of 691 households of high income whose addresses were given by the Chilean Servicio de Impuestos Internos (SII, which is the Chilean equivalent of the Income Revenue Service authority); 2) a sample of 3330 households selected from several urban areas classified in the

Chilean Censo de Población y Viviendas. This selection process is similar to the ones implemented in the American Survey of Consumer Finances (Kennickell, 1997) and the Encuesta Financiera de Familias of the Bank of Spain (Bover, 2004). Microdatos and the Central Bank of Chile denote the first 691 households as the SII sample and the other 3330 as the Censo sample as a way to summarize their different origin.

The 691 households of the SII sample were selected with basis on their reported 2006 taxable income. The distribution of the SII sample was implemented with a stratification based on the first 9 deciles and the top 10 percentiles of taxable income. Although representing a small part of the EFH this sample guarantees the representation of a significant number of top income households and is therefore highly important to insure the adequate representativity of the survey across the richest households. For the SII sample Microdatos built expansion factors by using partial post-stratification (implemented as described before in equation 3) in relation to two variables, regions (grouped in 4 groups denoted as "Zonas") and 3 wealth strata (deciles 1 to 5, 6 to 8, and 9 to 10) from the Encuesta de Protección Social de 2004 (Behrman, Bravo, Mitchell, and Todd, 2006).

The 3330 Censo households, which compose the main part of the EFH 2007 sample, were selected from a sample with a high representation of rich urban areas. I present here a summary of the classification process of urban areas in different wealth types, but the reader is referred to the Informe Final Encuesta Financiera de Hogares (2008) for more details.

Urban areas in Chile are classified as cities (known by the Spanish term, "comunas") and smaller sub-areas inside each city, denominated "segmentos". The largest cities (comunas) in Chile were immediately chosen as part of the population from which the EFH household addresses were to be sampled. Then a large set of smaller comunas were randomly sampled to be representative of the rest of Chile. The random sampling of smaller cities was necessary, since with a 4000 household sample it is not possible for the EFH survey to efficiently cover all the 346 comunas of Chile.

Microdatos used the CASEN 2003 dataset to obtain the national deciles of household income per capita. Then each segmento is classified among 3 wealth types: 1) areas having at least 75% of households of deciles 9 and 10 are classified as type 3; other areas are then classified as type 2 if at least 75% of their households are of deciles 6 and above; finally, the remaining urban areas are classified as type 1.

Using this system, Microdatos classified 3764 distinct urban areas of the CASEN 2003 dataset according to 3 wealth types[1]. By choosing a higher number of urban areas from types 2 and 3, the EFH was able to over-sample higher income households. Microdatos sampled a large number of segmentos to insure a wider coverage of Chile. A small number of segmentos could create a sample with observations highly correlated, therefore a large number of segmentos insures a greater degree of independence between observations. In the final sample of households effectively interviewed by the EFH there are 694 segmentos: 259 of type 3, 194 of type 2, 225 of type 1, and 16 new segmentos with no wealth classification. The 16 segmentos without a wealth classification are meant to represent new urban areas which were created between 2003 and 2007.

After selecting the comunas and segmentos, a given number of households was selected in each segmento. The sampling probability of each household i in segmento j of comuna m can be summarized as the multiple of the probability of the comuna selection ($f_{1,m(i)}$), segmento selection ($f_{2,j(i)}$), and the selection of the household in its segmento ($f_{3,i}$).

8) $f_{total,i} = f_{3,i} \times f_{2,j(i)} \times f_{1,m(i)}$ [2].

After the survey the sample design expansion factors were then adjusted for sample non-response across different segmentos. This was the inverse selection probability method used to compute the first version of the EFH expansion factors (Centro de Microdatos, 2008). Finally, the SII and Censo samples were jointly combined by multiplying the expansion factors with their respective sample

---

[1]It is important to note that the CASEN 2003 was only used to randomly select segmentos and comunas. There was no sampling of the households interviewed in the CASEN 2003 survey, therefore it is not possible to link the households in either sample.

[2]In the Informe Final of Microdatos these probabilities are explained in greater detail. Let $h(m)$ be the wealth type of each comuna. Let $M(k)$, $M(m)$, and $M(j)$ be the number of households in each region k, comuna, and segmento, respectively. Let $M(h)$ be the total number of households of type $h = 1, 2, 3$ collected by the EFH survey. $c(h, k)$ denotes the number of towns of type $h = 1, 2, 3$ selected in each region k. The number $n(j, m)$ denotes the number of urban areas inside comuna m selected for the survey. Finally, $g(i, j)$ denotes the number of households that will be selected for survey inside each urban area j. Note that that some comunas are auto-selected (meaning selected with 100% probability), while other comunas are randomly sampled among a set of other similar towns. Among the randomly sampled towns in each region, each town m of type $h = 1, 2, 3$ is selected with probability $f_{1,m} = c(h(m), k) \dfrac{M(m)}{M(k(m))}$. Then segmentos inside each comuna are selected with probability $f_{2,j} = n(j, m(j)) \dfrac{M(j)}{M(m(j))}$. Finally, each household i in segmento j is selected with probability $f_{3,i} = \dfrac{g(i, j(i))}{M(j(i))}$.

proportions for each wealth strata (types 1,2,3). I will now show that substantial improvements can be made in relation to these expansion factors by using post-stratification procedures.

5   Computing expansion factors for the EFH sample

5.1   Why not use a pure inverse probability weight?

Sampling theory usually advises that selecting a number of households proportional to the population of each area is optimal (Lohr, 2009). However, it was too expensive for Microdatos to interview a small number of households by segmento, since this would demand interviewers to cover a wide geographical area. Therefore 93% of the Censo sample is represented by segmentos with 10 to 32 households independently of the segmento size. Unfortunately, this has large implications for the variance of the expansion factors based on segmentos. Suppose two different segmentos have similar households, but their population sizes are different. One segmento has 100 households, while the other segmento has 3000 households. If due to non-response, the final sample ends up with 20 households in the small segmento and 1 household in the large segmento, then the households of each segmento are assigned respectively expansion factors of 5 and 3000. Obviously, this increases a lot the variance of any estimator using inverse probability expansion factors (Kish, 1992).

Also, all measures of survey population - including the ones provided by Census data - have measurement error. Since the population of smaller areas such as segmentos has a much larger measurement error than the estimates for bigger areas such as cities, this implies that expansion factors based on segmentos have substantial measurement errors. Considerations of excessive variance and measurement error imply therefore that an inverse selection probability expansion factor is unlikely to be optimal for the EFH survey.

5.2   Linking the CASEN and EFH households in different types

Typically, small surveys do not cover all the areas of a country. However, the goal of the researcher is to make an inference from the small survey to the whole target population nationwide. For

this reason there must be a correspondence rule linking all the non-represented of Chile to units effectively represented in the survey. Deville and Lavallée (2006) provides the best description of the importance of this problem. To make the EFH representative of Chile we must classify all its households in a similar set of strata as other larger surveys of Chile.

The CASEN 2003 and 2006 datasets are the most ideal ones to build expansion factors for the EFH, since they were used in its sample design and have the same measures of household income. However, these surveys represent the distribution of urban households of Chile in the past. For this reason I update the expansion factors of these surveys to reflect demographic growth between those years and the EFH sample years.

Also, the EFH households must be classified in terms of their income in a similar way as the CASEN 2003 and 2006 samples. However, the CASEN 2003 and 2006 income percentiles reflect only the income distribution of those years. These percentile values fail to take into account the income growth or income inequalities over the more recent years[3]. Therefore I used the national percentiles of the income distribution of the urban CASEN 2006 data and then I updated each percentile for the nominal income growth of each decile between 2006-07 using the Encuesta Suplementaria de Ingresos (ESI) of the Chilean INE[4]. The EFH households were then classified according this updated measure of the income percentiles in 2007.

5.2.1 Matching the CASEN surveys with the INE population projections

Since the CASEN surveys include fewer comunas than the total of 346 comunas of Chile, I take into account the non-represented comunas by assigning their population to the comunas of the same

---

[3]This ability to update the previous Chilean surveys to reflect the income distribution of the EFH sample years is one of the reasons why I choose to build strata based on household income. It is possible to use other strata classifications, such as asset wealth. However, there are no measures of how much household asset wealth has grown in Chile since 2004. Therefore using strata defined by assets incurs the risk of seriously mis-measuring the current situation of Chile.

For this reason I chose not to use the asset wealth deciles of the EPS 2004 to classify EFH households, as was done in the original expansion factors of Microdatos.

[4]Due to its size, the ESI data is ideal to measure the updated income distribution of Chile, since it covers 33000 Chilean households in the last quarter of every year.

region in the same proportion as their population size[5]. Then I compute the ratio of population of each comuna in the CASEN surveys with their projected population by the INE for the EFH sample years, $ratio\_INE\_CASEN(i) = \frac{population\ INE(i)}{population\ CASEN(i)}$. Then I multiply the population of each comuna in the CASEN survey by this ratio to get a CASEN dataset that more accurately reflects the current Chilean population.

5.2.2 Classifying the CASEN and EFH in similar strata

Since the EFH does not include all the urban areas of the CASEN we must take into account each urban area is representative of a wider population than just itself. I do this by assigning the population of the non-represented urban areas in the CASEN survey to areas of similar income represented in the EFH.

Therefore I classified all the urban households in 3 income types of the CASEN 2003 according to their national income decile (deciles 1 to 5, deciles 6 to 8, deciles 9 to 10)[6]. Then all the segmentos and comunas of the CASEN 2003 are classified as types 1, 2, and 3, according to their population of each household type in the same way as described in section 4. Segmentos and comunas in the EFH are classified with the same types 1, 2, and 3, that were obtained from the CASEN 2003. For the 16 segmentos of the EFH not present in the CASEN 2003 I use their own sample observations to assign them a type 3, 2, or 1, classification. The expansion factor for these 16 segmentos will be the same one as for the segmentos of the same type in the same comuna.

In this way all households in the CASEN and EFH are classified in the groups based on the households' income type (1,2,3) and their segmento. Let $i = 1,..,K$ denote each population group (denoted by segmento number and the household strata). Then I obtain a list of all the common

---

[5]I update the population of each represented comuna as: $population\_1(i) = population(i) +$ $\dfrac{population(i)}{\sum\limits_{k=1,k\in S,k\in G(i)}^{K} population(k)} \times \sum\limits_{k=1,k\notin S,k\in G(i)}^{K} population(k)$. , where S is the set of comunas represented in the CASEN survey and $G(i)$ is the set of comunas in the same Chilean region as comuna i.

[6]This procedure is the same one as described in section 4 with a small difference. Here I use deciles of total household income and not household income per capita. I prefer the use of total income, because it is more stable than income per capita (with a per capita income criterion a household could change income strata if a new member is born or an older one dies). However, this distinction has a small impact on the results.

groups in the EFH and CASEN. The population of each group i represented in the EFH is updated as:

$$9)\ population\_1(i) = population(i) + \frac{population(i)}{\sum_{k=1,k\in S,k\in G(i)}^{K} population(k)} \times \sum_{k=1,k\notin S,k\in G(i)}^{K} population(k),$$

where S is the set of groups represented in the EFH survey. $G(i)$ is the set of groups in the same region of Chile that have the same type of segmento (i.e., 12,3) and household strata (i.e., 1,2,3) as group i[7]. This procedure therefore makes the EFH urban areas representative of Chile by matching them with the CASEN data (which was also updated to match the INE projections for the population of Chile by comuna in 2007).

5.3 The Censo sample

The EFH is the first household survey in Chile that seeks to measure a large number of variables (as many as 200 relevant financial-economic characteristics of households), but with only a small number of households (1200 or 4000). Especially since it is designed to become a useful tool for policy makers and financial institutions it is quite important to build expansion factors that represent Chile in an adequate way, but with minimum variance. The survey design of the Censo sample worries only about the comunas and segmentos of residence of the household. To take into account the stratified sample design the researcher must forcibly deal with the different areas of residence of each household and with the larger non-response rates of higher income households (Kennickell, 1997, 1999, Bover, 2004, 2008).

However, it is not clear how many strata we should adjust for. The ideal is to build expansion factors that take into account relevant household heterogeneity in terms of income and wealth, but that ignore redundant strata that only increase variance. The original expansion factors took into account heterogeneity by segmento. However, unless we believe the finances of households in

---

[7]If some types of strata defined by region, segmento type and household wealth type, are not represented in the EFH, then a similar adjustment is made based on a strata classification with fewer categories (i.e., just region and segmento type).

segmentos of different sizes are extremely heterogeneous, then this option implies a large variance in the expansion factors. In order to reduce the variance of the expansion factors one consider to build expansion factors adjusting only for comunas and types of segmento (1,2,3). However, since there are 80 comunas in the EFH it is still possible that even adjusting only for comunas creates unnecessary variance. One could therefore consider adjusting for groups of comunas (according, for example to their wealth types 1,2,3) to reduce the number of strata.

Since it is not obvious how many strata to use in the EFH I created a set of different expansion factors and tested which ones made a better fit to the income distribution of Chile. The set of expansion factors is detailed as follows. Each expansion factor is labelled with 4 digits, "expr****".

The first digit classifies the "smoothness" of the factor estimates: it has value 0 if the expansion factor uses the true strata population or 1 for a "smoothed" estimate. The "smoothed" expansion factors "expr1***" are obtained by using the mean predicted values of a regression of "expr0***". I tried two regression options. The first one was a kernel regression[8] using the median and interquartile range for the household total income of the comuna and segmento of each observation. The second option was a linear regression using the mean years of education, mean number of persons per household, and the 25%, 50% and 75% household income quantiles of the comuna and segmento of each observation. Both options gave similar results, therefore I only report the results using the second option.

The second digit indicates the level of aggregation of the geographical areas of residence: 6 or "Zonas 2" differentiates for two aggregate groups of Chilean regions (Región Metropolitana, Otras Regíones de Chile); 5 or "Zona" differentiates for four groups of Chilean regions (Regíon Norte includes Chilean regions 1 to 4, Regíon Centro includes regions 5 to 8, Regíon Sur includes regions 9 to 12, and Regíon Metropolitana includes region 13); 4 or Región differentiates for the 13 Chilean regions that composed the whole of Chile until 2007; 3 or "Provincia" differentiates for each province of Chile; 2 or "Comuna" differentiates for each town of Chile; and, finally, 1 or "Segmento" differentiates for each segmento of Chile.

---

[8]Here I apply the Nadaraya-Watson kernel smoothing method, using a Epanechnikiv density function and the Silverman's rule of thumb bandwidth (Manski, 1990).

Da Silva and Opsomer (2009) applied a similar method to compute alternative expansion factors for the National Health and Nutrition Examination Survey dataset. Their research found that local polynomial regressions were quite effective in reducing the bias and variance of non-response adjustments to the expansion factors.

The third digit indicates whether the strata consider different types of comunas and segmentos: 0 - does not differentiate for comunas and segmentos of different wealth types; 1 - differentiates for the 3 different types of segmento wealth; 2 - differentiates for the 3 different types of comunas; and 3 - differentiates for the 9 different types of wealth of both the comuna (3 types) and segmento (3 types) of residence of the household.

Finally, the fourth digit indicates whether the strata differentiate for different types of household income: 0 - indicates no differentiation across households in the same area; 1 - indicates a differentiation for households according to 3 income types (1 - percentiles 1 to 50%, 2 - percentiles 51-80%, 3 - percentiles 81-100%, measured at the national level); and 2 indicates a differentiation for households of the national income percentiles 1-35, 36-50, 51-60, 61-68, 69-75, 76-80, 81-85, 86-88, 89-90, 91-92, 93-94, 95, 96, 97, 98, 99, 99.5, and 100.

This classification method results in a "phonebook list" of expansion factors. So for instance, an expansion factor based only on segmentos would be expr0100, while an expansion factor based on comunas and types of segmentos would be expr0210. A factor based on aggregate zonas and types of comunas and segmentos would be expr0530. Even inside each comuna the EFH sample design had a large oversample of richer neighborhoods. Therefore all the options I analyze consider either household wealth type or segmento type as a control variable for the strata of each household.

Note that the expansion factor expr0100 use the information on the population of all the urban sampling units in the EFH and how they were selected from the CASEN urban areas. Therefore the expansion factor expr0100 is equivalent to the inverse probability of selection of the unit (IPSU).

5.4   The SII sample

The SII sample is highly biased towards high income households and requires therefore a different statistical treatment from the rest of the sample. Due to the small sample size I decided to use a stratification based on only two aggregate regions ("Zonas 2": Región Metropolitana, Otras Regíones de Chile), but with several income strata (national income percentiles 1-35, 36-50, 51-60, 61-68, 69-75, 76-80, 81-85, 86-88, 89-90, 91-92, 93-94, 95, 96, 97, 98, 99, 99.5, and 100). The large number of income strata was required, because the SII sample was indeed highly focused on the top percentiles of income and therefore those households demanded differential treatment from the

others. The SII sample was then matched to the population values of these strata in the CASEN 2006 data.

In terms of the notation above the SII factor method would correspond to an expansion factor of the type expr0602. Adding more types of income (such as strata for the percentiles 1-20 and 21-35) does not change the results significantly.

5.5 Combining the Censo and SII samples

The Censo and SII samples were then combined in the EFH 2007 to form a joint sample of 4021 households. The final expansion factor was then updated to reflect the proportion each sample has on the final dataset:

$$10) \ \text{factor}_h(k) = \text{factor}_h(k) \times \frac{n_h(k)}{n(k)},$$

where $h \in \{SII, Censo\}$ denotes the sample origin of the observation and $k \in \{$Regíon Metropolitana, Otras Regíones de Chile$\}$ denotes the geographical area of the observation. $n_h(k)$ represents the number of households that sample h has in area k, while $n(k) = n_{SII}(k) + n_{Censo}(k)$ denotes the total number of households in the EFH present in area k[9]. The final factor is then labelled "expr****" in the same way as the factor for the Censo sample[10].

6 Evaluating the expansion factors

6.1 Do the original Microdatos expansion factors provide efficient estimates?

Center Microdatos provided a set of provisory expansion factors for the EFH survey when information from the CASEN 2003 and 2006 was not yet publicly available. The expansion factor of Centro

---

[9]This adjustment is necessary in order that the average household of both samples represents the same number of Chilean households. The adjustment requires different proportions for the Regíon Metropolitana and the Otras Regíones de Chile in order to keep the representativity both at the national level and at the level of the Regíon Metropolitana. This is important, because the EFH waves in 2008 and 2009 were only implemented in the Regíon Metropolitana. Therefore this adjustment allows for comparability across different EFH waves.

[10]Since the SII factor is kept the same for all different options, there is no confusion in keeping the factor Censo notation.

Microdatos uses the population of each comuna and the type of segmento in each comuna to create an expansion factor for the Censo sample. This procedures is based on a partial post-stratification method that considers comunas and types of segmentos as independent strata. Also, this procedure did not take into account how comunas in the EFH were selected and that therefore not all comunas in the EFH are equally representative of the other comunas in Chile. Microdatos also used a partial post-stratification procedure for the SII sample based on 4 aggregate geographical regions of Chile and 3 wealth strata (1-50, 51-80, 81-100 percentile groups) from the Encuesta de Proteccion Social in 2004 (EPS). Afterwards, Microdatos combined the two samples using the same procedure described in equation 10).

It is important to note that the provisory Microdatos procedure suffers from three flaws. One, in the Censo sample it is not taken into acccount that the richest comunas in Chile were chosen to be part of the EFH and therefore that the Chile outside the EFH sample design is poorer than the one that was included. Second, the SII sample is based on only three wealth strata, when the sample selected was highly based on individiuals of the highest income percentiles in Chile (95% and above). A third failure is that Microdatos uses partial post-stratification on both the Censo and SII samples. This procedure does not take into account that strata are not independent, since some Chilean comunas and regions are much richer than others. One can easily perceive that all these flaws create a bias in the Microdatos sample towards over-counting the number of wealthy households in Chile.

I will now show in the next section that the original EFH weights created by Microdatos provide highly biased estimates of the education and income distribution in Chile. In particular, Microdatos weights estimated the percentiles of income in Chile with a mean absolute bias of 28.3%.

6.2 Comparing factor options with income strata

To evaluate the performance of the expansion factors I first computed the minimum, maximum, mean, and standard-deviation of each expansion factor option. These dispersion statistics allow me to measure how variance reduction there is in choosing each option. I also computed their correlation coefficient in relation to the factor expr0602. The factor expr0602 is the one that takes into account the largest heterogeneity in income strata. Therefore the correlation coefficient gives a

rough measure of whether the dispersion in each expansion factor is indeed significant for explaining the wealth type of each household.

In table 1 I show that indeed there is substantially more variance in choosing expansion factors based on segmentos, comunas, and even provincias. Also, factors based on segmentos imply that some households have a statistical representation for just 10 or 15 households, while other ones represent 15000 or 20000 households, which would be equivalent to some of the largest comunas in Chile. Choosing expansion factors based on segmentos and comunas results in standard-deviations larger than the mean. Creating expansion factors based on aggregate regions of types 5 and 6 instead of segmentos successfully reduces this standard-deviation from above 1300 to less than 900 and reduces the minimum-maximum range from 15-15000 to around 50-5000. Note also that these large standard-errors are all caused by the Censo sample and not by the SII sample. Since the sample design of this sample could only distinguish people living in 3 types of neighborhoods, there is no reason why some households are worth one thousand times less than other similar households.

However, perhaps the biggest problem for the factor types based on segmentos, comunas and provincias is that their correlation with the income strata population (given by expr0602) is low. By using aggregate regions (such as type 5 or 6) instead of segmentos one can increase the correlation with the income strata from 40% to above 60%. It is important to note that the original expansion factors (labelled CMD for Centro de Microdatos) only have a correlation factor of 23% with the income strata, substantially below the new options.

Using the expansion factor expr0602 it is possible to get a reliable estimate of the Chilean household income distribution using the EFH 2007. To compare how well each expansion factor estimates the Chilean income distribution, I compute the average absolute deviations from the estimates of the income distribution of each factor type in relation to the income distribution obtained with factor 0602:

11) $abs\_dev_b(expr^{****}) = \frac{1}{K(b)} \sum_{i=1}^{K(b)} \left| \log(stat(i))_{expr^{****}} - \log(stat(i))_{expr^{0602}} \right|,$

where $b \in \{percentiles, deciles, quintiles\}$, $stat(i) \in \{percentile, decile, quintile\}$, and $K(b) \in \{99, 9, 4\}$. The top percentile is excluded from the mean absolute deviation statistic, since income is potentially unbounded and therefore there is no upper limit for the top percentile of household income.

In table 2 I show the mean values of $abs\_dev_b(expr^{****})$. Again, it is quite clear that expansion factors based on more aggregate areas (such as 0431, 531, and 0631) are more efficient in estimating the income distribution of Chile. Therefore the new work improves estimates greatly at the same time that it reduces factor variances. Tables 2 also shows that accounting for segmento type is also not enough to get a good representation of the income distribution. It is therefore important to account for both types of segmento and household.

Overall, the expansion factor 0531 appears to provide a good balance between reducing factor variance and providing a good fit of the income distribution of Chile. Therefore the study of this paper brings substantial improvements to the original expansion factors estimated by Microdatos. Overall, it is easy to conclude that the original expansion factors over-estimated the income of Chilean households by around 20%.

6.3   Monte Carlo performance of the expansion factors

The results from table 2 have some criticisms. First, they are the result of a single sample and therefore do not really take into account estimation variance. Second, the results are based on a counterfactual estimation of the income distribution of Chile using the CASEN 2006 and the ESI 2007. If this counterfactual estimation is not correct, then the results are misleading.

However, there are 3301 households from the Censo sample whose segmentos are all present in the CASEN 2003. This allows us to make the following Monte Carlo experiment. Let us get 2500 bootstrap samples of 3301 households from the CASEN 2003, using the same segmentos as the EFH. Then we apply the factor CMD and the other factor types I constructed before, obtaining statistics for the absolute deviations of the percentiles, deciles and quartiles of each factor in relation to the national urban CASEN 2003 income distribution. This exercise does not depend on counterfactual estimations of the income distribution, since the original population is known and remains fixed. Therefore it allows us to make an evaluation of how different factor types perform in different samples[11].

---

[11]Since there are few households per segmento in the CASEN 2003, I assign randomly households of segmentos from the same income types (1,2,3) to these segmentos. I do this because the short availability of households per segmento will tend to decrease the actual variance of the factors based on segmentos in relation to the truth.

In table 3.1 I show that the mean bias of the Microdatos sample is large, being around 10.8%. It is the largest bias next to the factors that only condition on the comuna and not on types of segmentos. This experiment clearly shows that it is not enough to condition on comunas to get consistent statistics in the EFH.

In table 3.2 I show the mean values of $abs\_dev_b(expr^{****})$ across all the 2500 Monte Carlo samples. The qualitative results are similar to tables 1 and 2. Again, expansion factors based on aggregate regions (such as types 5 and 6) work fairly well. In particular, the option expr0531 appears to be quite robust across all Monte Carlo samples and it is laso one of the alternatives with lowest estimated standard-error in all the 2500 Monte Carlo samples.

I also show the dispersion statistics of the performance of each factor type in tables 4.1-4.7. It is easy to verify that the expansion factor expr0531 performs quite well and that its top percentile of mean absolute error is still quite below the lowest percentile of mean absolute error for the original Microdatos factors. Therefore even in the worst 1% scenario, the expansion factor expr0531 would still estimate the income percentiles of Chile with a mean absolute error of only 5.2%.

6.4 Demographic representation of the EFH

Finally, in table 5 I compare the education statistics of the EFH 2007 with the CASEN 2006 and 2009, using both the original expansion factors by Microdatos and the expansion factor 0531. It is quite clear from this table that again the original expansion factors were portraying a richer Chile than the true one. The original expansion factors over-estimate the number of people with a post-graduate degree at the national level by 40%[12]. It also over-estimates the number of people with college degrees by 17%.

The new expansion factor 0531 is also accurate in representing the age distribution in Chile. It is relevant to note that the Center Microdatos expansion factors were also accurate in portraying the age distribution. This can be explained by the small inequality in the age distribution across Chile. Age is much more similar across different cities than education or income, because the mortality of

---

However, there is little qualitative impact in adding households or not to the selected segmentos of the EFH. Therefore I do not report the statistics obtained with the restricted EFH segmentos option.

[12] This work uses the CASEN as a reference for the true education distribution in Chile. However, the Census 2002, CASEN 2003, and EPS 2004 datasets give similar conclusions as the CASEN 2009 data.

poor households is not dramatically higher than for richer households. Therefore estimates of the age distribution have less variance and are less sensitive to the use of different expansion factors.

In table 7 I show that the unemployment rates estimated for the EFH using the original expansion factors are too low relative to official INE statistics[13]. Again, the new expansion factors make a much better job of providing a good fit of Chile, while the original ones seem to portray a richer country than Chile actually is (in terms of income, education, and unemployment rates).

A similar comparison is valid if we use the distribution of education and age for the Chilean household heads instead of all persons. The comparisons in tables 5, 6, and 7, are also similar if we use other expansion factors besides the 0531. These results are available from the author upon request.

The conclusion, therefore, is that the new expansion factors are a substantial improvement in making the EFH portray Chile effectively and represent a more reliable tool for both research and policy work.

7   Conclusions

This work re-estimates the expansion factors for the EFH 2007 dataset, using several full post-stratification options based on the geographical area and the income type of each household. I conclude that factors based on aggregate regions of Chile (such as groups of 4 or 2 regions only) have much less variance than the expansion factors based on more detailed geographical areas, such as comunas or segmentos. I also find that the factor types of the aggregate regions actually have less bias in measuring the income distribution than the expansion factors based on more detailed geographical strata. This result could be caused by measurement error in the population sizes of small urban units such as segmentos.

In particular, the option expr0531 appears to be quite robust in measuring the income distribution of Chile. Since the option 0531 appears to have lower bias and standard-error in a Monte Carlo exercise with 2500 Monte Carlo simulations, I recommend that this expansion factor should be

---

[13]There is another unemployment survey for the Santiago area implemented by the Universidad de Chile, the Encuesta de Ocupacíon y Disocupacíon (EOD). The EOD unemployment rates for the final quarter of 2007 were around 8.5%, which gives an even worse picture than the INE statistics.

included in the official EFH 2007 dataset. In addition, the new expansion factors estimated for the EFH dataset seem to provide a much better fit of the education of Chilean households than the original Microdatos factors.

This research brings substantial improvements to the estimates of Chilean household finances. In particular, computing statistics for education of the households, income, asset wealth, and debt levels, it is easy to conclude that the original Microdatos expansion factors over-estimated the income of Chilean households by around 20%. The new expansion factors make a much better job of providing a good fit of Chile, while the original ones seem to portray a richer country than Chile actually is (in terms of income, education, and unemployment rates).

The conclusion, therefore, is that the new expansion factors are a substantial improvement in making the EFH portray Chile effectively and represent a more reliable tool for both research and policy work.

References

[1] Alfaro, R, N. Gallardo and R. Stein (2010), "The Determinants of Household Debt Default", Working Paper 574, Banco Central de Chile

[2] Banco Central de Chile (2009), "Encuesta Financiera de Hogares: Metodología y Principales Resultados EFH 2007", Banco Central de Chile

[3] Banco Central de Chile (2010), "Informe de Estabilidad Financiera", Banco Central de Chile

[4] Behrman, J., D. Bravo, O. Mitchell and P. Todd (2006), "Encuesta de Protección Social 2004: Presentación General y Principales Resultados", Subsecretaria de Previsíon Social, Chile

[5] Bover, O. (2004), "The Spanish survey of household finances (EFF): description and methods of the 2002 wave", Occasional Paper No. 0409, Banco de España

[6] Bover, O. (2008), "The Spanish survey of household finances (EFF): description and methods of the 2005 wave", Occasional Paper No. 0803, Banco de España

[7] Brownstone, D. and X. Chu (1994), "Multiply Imputed Sampling Weights for Consistent Inference with Panel Attrition", Transportation Center Working Paper 590, University California Irvine

[8] Brownstone, D. (1997), "Multiple imputation methodology for missing data, non-random response and panel attrition", Institute of Transportation Studies Working Paper 97-4, University California Irvine

[9] Centro de Microdatos (2008), "Informe Final Encuesta Financiera de Hogares", Universidad de Chile

[10] Crockett, A. (2008), "Weighting the Social Surveys", Economic and Social Data Service (ESDS) Government, UK

[11] da Silva, D. and J. Opsomer (2009), "Nonparametric propensity weighting for survey nonresponse through local polynomial regression", *Survey Methodology*, 35 (2), 165-176

[12] Deville, J.C. and C.E. Sarndal (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, 87 (418), 376-382

[13] Deville, J.C., C.E. Sarndal, and O. Sautory (1993), "Generalized Raking Procedures in Survey Sampling", *Journal of the American Statistical Association*, 88 (423), 1013-1020

[14] Deville, J.C. and P. Lavallée (2006), "Indirect Sampling: The Foundations of the Generalized Weight Share Method", *Survey Methodology*, 32 (2), 165-176

[15] Estevao, V., M. Hidiroglou, and C.E. Sarndal (1995), "Methodological Principles for a Generalized Estimation System at Statistics Canada," *Journal of Official Statistics*, 11, 181-204

[16] Estevao, V. and C.E. Särndal (2000), "A Functional Form Approach to Calibration", *Journal of Official Statistics*, 16, 379-399

[17] Fabrizi, E. and C. Trivisano (2007), "Efficient stratification based on nonparametric regression methods", *Journal of Official Statistics,* 23, 35-50

[18] Faiella, I. and R. Gambacorta (2007), "The Weighting Process in the SHIW," Economic working papers 636, Bank of Italy

[19] Fuenzalida, M. and J. Ruiz-Tagle (2009), "Riesgo Financiero de los Hogares", *Economía Chilena*, 12(2): 35-53

[20] Funaoka, F., H. Saigo, R.. Sitter, and T. Toida (2006), "Bernoulli Bootstrap for Stratified Multistage Sampling", *Survey Methodology*, 32 (2), 151-156

[21] Groves, R. and M. Couper (1995), "Theoretical motivation for post-survey non-response adjustment in household surveys", *Journal of Official Statistics*, 11 (1), 93-106

[22] Hedlin, D. (2000), "A procedure for stratification by an extended Ekman rule", *Journal of Official Statistics*, 15, 15-29

[23] Horgan, J. (2006), "Stratification of Skewed Populations: A review", *International Statistical Review*, 74, 67-76

[24] Horvitz, D. and D. Thompson (1952), "A Generalization of Sampling Without Replacement From a Finite Universe", *Journal of the American Statistical Association*, 47, 663-685

[25] Instituto Nacional de Estadísticas (2010), "Programa de proyecciones de la poblacíon", INE Chile

[26] Instituto Nacional de Estadísticas (2010), "Encuesta Suplementaria de Ingresos, 1990-2008", INE Chile

[27] Kalton, G. (2009), "Methods for oversampling rare subpopulations in social surveys", *Survey Methodology*, 35 (2), 125-141

[28] Kadane, J. (2005), "Optimal Dynamic Sample Allocation Among Strata", *Journal of Official Statistics*, 21, 531-541

[29] Kennickell, A. and R. Woodburn (1997), "Consistent Weight Design for the 1989, 1992 and 1995 SCFs, and the Distribution of Wealth", Federal Reserve Board, mimeo

[30] Kennickell, A. (1999), "Revisions to the SCF Weighting Methodology: Accounting for Race/Ethnicity and Homeownership", Federal Reserve Board, mimeo

[31] Kish, L. and M. Frankel (1974), "Inference from Complex Samples", *Journal of the Royal Statistical Society*, Series B (Methodological), 36 (1), 1-37

[32] Kish, L. (1992), "Weighting for unequal Pi", *Journal of Official Statistics*, 8, 183-200.

[33] Kott, P. (1991): "A Model-Based Look at Linear Regression with Survey Data," *The American Statistician*, 45, 107–112.

[34] Kott, P. (2006a), "Sample survey theory and methods: A correspondence course", National Agricultural Statistics Service (USDA), mimeo.

[35] Kott, P. (2006b), "Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors", *Survey Methodology*, 32 (2), 133-142

[36] Kozak, M. and M. Verma (2006), "Geometric Versus Optimization Approach to Stratification: A Comparison of Efficiency", *Survey Methodology*, 32 (2), 157-163

[37] Lavallée, P., and Hidiroglou, M. A. (1988), "On the stratification of skewed populations," *Survey Methodology*, 14, 33-43.

[38] Little, R. (1981): "Robust model-based inference for a finite population mean from unequally weighted samples," in *Proceedings of the Survey Research Methods Section American Statistical Association*.

[39] Little, R. and S. Vartivarian (2005), "Does Weighting for Nonresponse Increase the Variance of Survey Means?", *Survey Methodology*, 31 (2), 161-168

[40] Lohr, S. (2009), "Sampling: Design and analysis", Duxbury Press, USA.

[41] Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80 (2), 319-323

[42] MIDEPLAN (2005), "Encuesta de Caracterización Socieconómica Nacional (Casen 2003): Marco metodológico", Gobierno de Chile.

[43] MIDEPLAN (2007), "Encuesta de Caracterización Socieconómica Nacional (Casen 2006): Documento metodológico", Gobierno de Chile.

[44] Nathan, G., and T. Smith (1989), "The Effect of Selection in Regression Analysis," in Analysis of Complex Surveys," Wiley.

[45] Neyman, J. (1934), "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection", *Journal of the Royal Statistical Society*, 97, 558-606

[46] Rao, J. and C. Wu (1988), "Resampling Inference With Complex Survey Data", *Journal of the American Statistical Association*, 83, 231-241

[47] Rao, J. (2005), "Interplay Between Sample Survey Theory and Practice: An Appraisal", *Survey Methodology*, 31 (2), 117-138

[48] Reiter, J., T. Raghunathan, and S. Kinney (2006), "The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data", *Survey Methodology*, 32 (2), 143-149

[49] Rivest, L. (2002), "A generalization of the Lavallée-Hidiroglou algorithm for stratification in business surveys", *Survey Methodology*, 28, 191-198

[50] Ruiz-Tagle, J. and F. Vella (2010), "Borrowing Constraints and Credit Demand", Working Paper 578, Banco Central de Chile

[51] Sagner, A. (2009), "Determinantes del Precio de Viviendas en Chile", Working Paper 549, Banco Central de Chile

[52] Särndal, C., B. Swensson, and J. Wretman (1992), "Model Assisted Survey Sampling," Springer-Verlag.

[53] Sautory, O. (2003), "CALMAR 2: A new version of the CALMAR calibration adjustment program", *Statistics Canada International Symposium Series: Proceedings*, Statistics Canada

[54] Tillé, Y. (2006), "Sampling Algorithms", Springer.

[55] Wooldridge, J. M. (1999), "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples," *Econometrica*, 67, 1385–1406.

[56] Wu, C. (2002), "Optimal calibration estimators in survey sampling", *Statistics Canada International Symposium Series: Proceedings*, Statistics Canada

8   Appendix

Table 1: Dispersion of the expansion factors and correlation with income strata

Aggregation area	Factor	Mean	Standard deviation	Min	Max	Correlation w/ 0602	Correlation w/ IPSU
6	**0602**	968	619	70	2216	100.0%	43.2%
1	**CMD**	982	1315	17	22054	23.6%	46.8%
1	**0100/ IPSU**	957	1251	14	15316	43.2%	100.0%
2	**0200**	957	952	27	5337	41.7%	69.6%
2	**0201**	957	1068	22	8137	56.3%	71.0%
2	**1200**	957	776	13	5337	50.2%	57.9%
2	**1201**	957	776	13	5337	50.2%	57.9%
2	**0210**	957	1208	26	8279	46.4%	87.2%
2	**0211**	957	1313	10	10000	47.0%	78.4%
3	**0310**	957	999	27	7358	54.2%	76.5%
3	**0311**	957	1078	10	9145	57.0%	69.9%
4	**0410**	957	901	27	5337	59.9%	73.6%
4	**0411**	957	968	10	5366	62.6%	69.0%
4	**0430**	957	903	27	5337	60.0%	73.4%
4	**0431**	957	970	10	5366	62.5%	68.8%
5	**0510**	957	844	27	5337	62.5%	69.1%
5	**0511**	957	910	27	5337	66.4%	65.0%
5	**0520**	957	691	27	5337	53.6%	55.4%
5	**0521**	957	771	27	5337	77.5%	58.3%
5	**0530**	957	846	27	5337	62.7%	68.9%
5	**0531**	957	912	27	5337	66.3%	64.8%
6	**0610**	957	835	27	5337	62.9%	68.4%
6	**0611**	957	900	27	5337	67.1%	64.3%
6	**0630**	957	837	27	5337	63.1%	68.2%
6	**0631**	957	902	27	5337	67.0%	64.1%

Table 2: Absolute deviations of the income distribution of each factor type in the EFH07 relative to factor 0602

Aggregation area	Factor	percentiles	dciles	quintiles
1	**CMD**	28.3%	27.2%	26.5%
1	**0100**	9.1%	7.9%	8.2%
2	**0200**	22.3%	21.5%	21.4%
2	**0201**	3.5%	1.4%	1.1%
2	**1200**	23.6%	22.4%	22.2%
2	**1201**	23.6%	22.4%	22.2%
2	**0210**	6.7%	5.5%	5.7%
2	**0211**	4.0%	2.3%	1.7%
3	**0310**	7.9%	7.0%	7.3%
3	**0311**	4.2%	2.8%	2.3%
4	**0410**	8.4%	7.6%	7.9%
4	**0411**	3.8%	2.5%	1.8%
4	**0430**	8.1%	7.3%	7.8%
4	**0431**	3.7%	2.4%	1.8%
5	**0510**	10.1%	8.8%	9.4%
5	**0511**	4.0%	2.0%	1.3%
5	**0520**	26.1%	24.6%	24.0%
5	**0521**	4.0%	1.6%	0.6%
5	**0530**	9.2%	8.3%	8.8%
5	**0531**	3.9%	1.8%	1.2%
6	**0610**	10.0%	9.1%	9.8%
6	**0611**	4.0%	2.0%	1.3%
6	**0630**	9.5%	8.6%	9.4%
6	**0631**	3.9%	1.8%	1.3%

Table 3.1: Mean Bias of the income distribution of each factor type (2500 MC simulations)

Aggregation area	Factor	percentiles	dciles	quintiles
1	**CMD**	10.8%	11.1%	11.3%
1	**0100**	-6.7%	-6.9%	-7.4%
2	**0200**	21.5%	22.2%	23.0%
2	**0201**	-1.1%	-1.3%	-1.9%
2	**1200**	22.2%	23.0%	23.7%
2	**1201**	22.3%	23.1%	23.8%
2	**0210**	-8.9%	-9.0%	-9.5%
2	**0211**	-5.4%	-5.5%	-6.0%
3	**0310**	-6.8%	-6.9%	-7.4%
3	**0311**	-3.5%	-3.4%	-4.0%
4	**0410**	-6.1%	-6.2%	-6.7%
4	**0411**	-2.4%	-2.3%	-2.7%
4	**0430**	-6.4%	-6.5%	-7.0%
4	**0431**	-2.6%	-2.5%	-2.8%
5	**0510**	-6.0%	-6.1%	-6.7%
5	**0511**	-2.2%	-2.0%	-2.5%
5	**0520**	24.5%	25.1%	25.8%
5	**0521**	0.2%	-0.1%	-1.4%
5	**0530**	-6.3%	-6.4%	-7.0%
5	**0531**	-2.3%	-2.1%	-2.5%
6	**0610**	-5.8%	-5.9%	-6.5%
6	**0611**	-2.1%	-1.9%	-2.3%
6	**0630**	-6.1%	-6.3%	-6.8%
6	**0631**	-2.2%	-2.0%	-2.4%

Table 3.2: Mean Absolute deviations of the income distribution of each factor type (2500 MC simulations)

Aggregation area	Factor	percentiles	dciles	quintiles
1	**CMD**	11.6%	11.4%	11.4%
1	**0100**	7.6%	7.2%	7.5%
2	**0200**	21.6%	22.2%	23.0%
2	**0201**	3.4%	2.9%	2.3%
2	**1200**	22.3%	23.0%	23.7%
2	**1201**	22.4%	23.1%	23.8%
2	**0210**	9.1%	9.0%	9.5%
2	**0211**	5.8%	5.6%	6.0%
3	**0310**	7.2%	7.1%	7.5%
3	**0311**	4.1%	3.9%	4.0%
4	**0410**	6.5%	6.4%	6.8%
4	**0411**	3.3%	2.9%	2.8%
4	**0430**	6.7%	6.7%	7.1%
4	**0431**	3.4%	2.9%	2.9%
5	**0510**	6.4%	6.4%	6.7%
5	**0511**	3.3%	2.8%	2.6%
5	**0520**	24.5%	25.1%	25.8%
5	**0521**	3.9%	3.2%	1.9%
5	**0530**	6.6%	6.6%	7.0%
5	**0531**	3.2%	2.8%	2.6%
6	**0610**	6.3%	6.2%	6.5%
6	**0611**	3.2%	2.7%	2.5%
6	**0630**	6.5%	6.4%	6.8%
6	**0631**	3.2%	2.7%	2.5%

Table 3.3: Standard-error of the Absolute deviations

of the income distribution of each factor type

(2500 MC simulations from the EFH segmentos in CASEN 2003)

Aggregation area	Factor	percentiles	dciles	quintiles
1	CMD	1.7%	1.8%	2.1%
1	0101	1.5%	1.4%	1.6%
2	0200	1.5%	1.5%	1.5%
1	1101	1.5%	1.4%	1.5%
2	1200	1.3%	1.3%	1.4%
2	1201	1.3%	1.3%	1.3%
2	0210	2.0%	2.0%	2.1%
2	0211	1.4%	1.3%	1.4%
3	0310	1.7%	1.6%	1.8%
3	0311	1.1%	1.0%	1.2%
4	0410	1.5%	1.5%	1.7%
4	0411	0.9%	0.9%	1.0%
4	0430	1.5%	1.5%	1.7%
4	0431	0.9%	0.9%	1.0%
5	0510	1.4%	1.4%	1.6%
5	0511	0.9%	0.8%	0.9%
5	0520	1.2%	1.2%	1.2%
5	0521	0.8%	0.7%	0.8%
5	0530	1.5%	1.5%	1.6%
5	0531	0.9%	0.8%	0.9%
6	0610	1.4%	1.4%	1.6%
6	0611	0.9%	0.8%	0.9%
6	0630	1.5%	1.5%	1.6%
6	0631	0.9%	0.8%	0.9%

Table 4.1: Q-1% Absolute deviations of the income distribution of each factor type (2500 MC simulations)

Aggregation area	Factor	percentiles	dciles	quintiles
1	CMD	8.0%	7.7%	7.3%
1	0100	3.4%	2.9%	2.5%
2	0200	18.2%	18.8%	19.3%
2	0201	1.6%	1.1%	0.5%
2	1200	19.4%	20.0%	20.5%
2	1201	19.5%	20.1%	20.5%
2	0210	4.8%	4.6%	4.6%
2	0211	2.9%	2.8%	2.8%
3	0310	3.7%	3.5%	3.4%
3	0311	1.9%	1.6%	1.5%
4	0410	3.2%	3.1%	3.0%
4	0411	1.4%	1.0%	0.7%
4	0430	3.3%	3.2%	3.1%
4	0431	1.4%	1.0%	0.8%
5	0510	3.4%	3.2%	3.1%
5	0511	1.4%	1.0%	0.6%
5	0520	21.8%	22.3%	23.1%
5	0521	2.3%	1.7%	0.4%
5	0530	3.5%	3.4%	3.3%
5	0531	1.4%	1.0%	0.6%
6	0610	3.2%	3.1%	3.0%
6	0611	1.4%	1.0%	0.6%
6	0630	3.3%	3.2%	3.2%
6	0631	1.4%	1.0%	0.6%

Table 4.2: Q-10% Absolute deviations of the income distribution of each factor type (2500 MC simulations)

Aggregation area	Factor	percentiles	dciles	quintiles
1	CMD	9.5%	9.1%	8.8%
1	0100	5.1%	4.8%	4.7%
2	0200	19.7%	20.3%	21.1%
2	0201	2.2%	1.7%	1.1%
2	1200	20.6%	21.3%	22.0%
2	1201	20.7%	21.4%	22.0%
2	0210	6.5%	6.5%	6.7%
2	0211	4.1%	3.9%	4.3%
3	0310	5.1%	5.0%	5.1%
3	0311	2.7%	2.5%	2.6%
4	0410	4.6%	4.5%	4.6%
4	0411	2.2%	1.8%	1.6%
4	0430	4.7%	4.7%	4.9%
4	0431	2.2%	1.9%	1.7%
5	0510	4.6%	4.5%	4.6%
5	0511	2.1%	1.8%	1.5%
5	0520	22.9%	23.5%	24.3%
5	0521	2.9%	2.2%	0.9%
5	0530	4.8%	4.7%	4.9%
5	0531	2.1%	1.7%	1.5%
6	0610	4.5%	4.4%	4.4%
6	0611	2.1%	1.7%	1.4%
6	0630	4.6%	4.6%	4.7%
6	0631	2.1%	1.7%	1.4%

Table 4.3: Q-25% Absolute deviations of the income distribution of each factor type (2500 MC simulations)

Aggregation area	Factor	percentiles	dciles	quintiles
1	**CMD**	10.5%	10.2%	9.9%
1	**0100**	6.2%	5.9%	6.0%
2	**0200**	20.6%	21.2%	22.0%
2	**0201**	2.7%	2.2%	1.6%
2	**1200**	21.4%	22.2%	22.8%
2	**1201**	21.5%	22.2%	22.9%
2	**0210**	7.7%	7.7%	8.0%
2	**0211**	4.9%	4.7%	5.1%
3	**0310**	6.0%	6.0%	6.2%
3	**0311**	3.3%	3.1%	3.2%
4	**0410**	5.4%	5.4%	5.7%
4	**0411**	2.7%	2.3%	2.2%
4	**0430**	5.6%	5.6%	5.9%
4	**0431**	2.7%	2.3%	2.3%
5	**0510**	5.4%	5.3%	5.6%
5	**0511**	2.6%	2.2%	2.0%
5	**0520**	23.7%	24.3%	25.0%
5	**0521**	3.3%	2.6%	1.4%
5	**0530**	5.6%	5.6%	5.9%
5	**0531**	2.6%	2.2%	2.0%
6	**0610**	5.3%	5.2%	5.5%
6	**0611**	2.6%	2.2%	1.9%
6	**0630**	5.5%	5.5%	5.8%
6	**0631**	2.6%	2.2%	1.9%

Table 4.4: Q-50% Absolute deviations of the income distribution of each factor type (2500 MC simulations)

Aggregation area	Factor	percentiles	dciles	quintiles
1	**CMD**	11.5%	11.3%	11.2%
1	**0100**	7.6%	7.2%	7.5%
2	**0200**	21.7%	22.3%	23.0%
2	**0201**	3.4%	2.8%	2.3%
2	**1200**	22.3%	23.0%	23.8%
2	**1201**	22.4%	23.1%	23.8%
2	**0210**	9.1%	9.1%	9.5%
2	**0211**	5.7%	5.5%	6.0%
3	**0310**	7.2%	7.1%	7.4%
3	**0311**	4.1%	3.8%	4.0%
4	**0410**	6.5%	6.4%	6.8%
4	**0411**	3.3%	2.9%	2.8%
4	**0430**	6.7%	6.7%	7.1%
4	**0431**	3.3%	2.9%	2.9%
5	**0510**	6.4%	6.3%	6.7%
5	**0511**	3.2%	2.8%	2.6%
5	**0520**	24.5%	25.0%	25.7%
5	**0521**	3.9%	3.1%	1.9%
5	**0530**	6.6%	6.6%	7.0%
5	**0531**	3.2%	2.8%	2.6%
6	**0610**	6.3%	6.2%	6.5%
6	**0611**	3.1%	2.7%	2.5%
6	**0630**	6.5%	6.4%	6.8%
6	**0631**	3.1%	2.7%	2.5%

Table 4.5: Q-75% Absolute deviations of the income distribution of each factor type (2500 MC simulations)

Aggregation area	Factor	percentiles	dciles	quintiles
1	**CMD**	12.7%	12.5%	12.7%
1	**0100**	8.9%	8.5%	8.9%
2	**0200**	22.7%	23.2%	24.0%
2	**0201**	4.1%	3.5%	2.9%
2	**1200**	23.2%	23.9%	24.6%
2	**1201**	23.3%	23.9%	24.7%
2	**0210**	10.4%	10.4%	10.9%
2	**0211**	6.7%	6.4%	7.0%
3	**0310**	8.3%	8.2%	8.6%
3	**0311**	4.8%	4.5%	4.8%
4	**0410**	7.5%	7.4%	7.9%
4	**0411**	4.0%	3.5%	3.5%
4	**0430**	7.7%	7.7%	8.1%
4	**0431**	4.0%	3.5%	3.6%
5	**0510**	7.4%	7.4%	7.8%
5	**0511**	3.8%	3.3%	3.1%
5	**0520**	25.4%	25.9%	26.5%
5	**0521**	4.4%	3.7%	2.5%
5	**0530**	7.6%	7.6%	8.1%
5	**0531**	3.8%	3.3%	3.2%
6	**0610**	7.2%	7.2%	7.6%
6	**0611**	3.8%	3.3%	3.0%
6	**0630**	7.4%	7.5%	7.9%
6	**0631**	3.7%	3.3%	3.1%

Table 4.6: Q-90% Absolute deviations of the income distribution of each factor type (2500 MC simulations)

Aggregation area	Factor	percentiles	dciles	quintiles
1	CMD	13.9%	13.8%	14.2%
1	0100	10.1%	9.7%	10.2%
2	0200	23.5%	24.1%	24.8%
2	0201	4.8%	4.2%	3.6%
2	1200	24.1%	24.7%	25.4%
2	1201	24.1%	24.7%	25.5%
2	0210	11.8%	11.6%	12.2%
2	0211	7.6%	7.3%	7.9%
3	0310	9.4%	9.2%	9.8%
3	0311	5.6%	5.2%	5.5%
4	0410	8.5%	8.3%	8.9%
4	0411	4.6%	4.0%	4.1%
4	0430	8.7%	8.6%	9.2%
4	0431	4.6%	4.1%	4.2%
5	0510	8.3%	8.2%	8.8%
5	0511	4.4%	3.9%	3.7%
5	0520	26.2%	26.6%	27.3%
5	0521	5.0%	4.2%	3.0%
5	0530	8.6%	8.5%	9.1%
5	0531	4.4%	3.8%	3.7%
6	0610	8.2%	8.0%	8.6%
6	0611	4.3%	3.8%	3.6%
6	0630	8.4%	8.3%	8.9%
6	0631	4.3%	3.8%	3.6%

Table 4.7: Q-99% Absolute deviations of the income distribution of each factor type (2500 MC simulations)

Aggregation area	Factor	percentiles	dciles	quintiles
1	**CMD**	16.1%	16.2%	16.9%
1	**0100**	12.2%	11.8%	12.2%
2	**0200**	25.2%	25.7%	26.4%
2	**0201**	6.1%	5.3%	4.7%
2	**1200**	25.3%	25.9%	26.7%
2	**1201**	25.4%	25.9%	26.7%
2	**0210**	14.0%	13.9%	14.3%
2	**0211**	9.2%	8.7%	9.4%
3	**0310**	11.5%	11.0%	11.7%
3	**0311**	7.0%	6.4%	6.8%
4	**0410**	10.2%	9.9%	10.7%
4	**0411**	5.8%	5.1%	5.1%
4	**0430**	10.4%	10.2%	10.9%
4	**0431**	5.8%	5.1%	5.1%
5	**0510**	9.9%	9.7%	10.4%
5	**0511**	5.4%	4.8%	4.6%
5	**0520**	27.4%	27.9%	28.6%
5	**0521**	5.9%	5.1%	3.9%
5	**0530**	10.1%	10.1%	10.7%
5	**0531**	5.4%	4.8%	4.7%
6	**0610**	9.8%	9.6%	10.2%
6	**0611**	5.3%	4.8%	4.5%
6	**0630**	10.0%	9.9%	10.5%
6	**0631**	5.3%	4.8%	4.5%

Table 5 - Distribution of Education

National - Persons

Education	CASEN06	EFH07 - CMD	EFH07 - 0531	CASEN09
Below high school	37.4%	32.3%	37.6%	37.9%
Some high school	39.2%	38.1%	39.2%	40.7%
Some College	16.1%	19.1%	15.5%	14.3%
Post-graduate	7.3%	10.5%	7.7%	7.1%
Total	100.0%	100.0%	100.0%	100.0%

Table 6 - Distribution of Age

National - Persons

Age	CASEN06	EFH07 - CMD	EFH07 - new	CASEN09
0/17	27.2%	26.0%	26.5%	27.0%
18/24	12.5%	13.0%	12.2%	13.1%
25/34	13.6%	14.0%	12.7%	13.2%
35/44	14.5%	13.8%	13.2%	13.5%
45/54	12.7%	13.7%	14.3%	13.5%
55/64	9.1%	9.3%	9.8%	9.1%
65/+	10.4%	10.1%	11.2%	10.5%
Total	100.0%	100.0%	100.0%	100.0%

Table 7: Unemployment rates

2007 Q4

	Chile	RM
INE	7.2%	7.1%
EFH CMD	6.5%	5.8%
EFH 0531	7.3%	6.7%

Documentos de Trabajo
Banco Central de Chile

Working Papers
Central Bank of Chile

NÚMEROS ANTERIORES   PAST ISSUES

DTBC – 631   Mayo 2011
Aplicaciones del Modelo Binominal para el Análisis de Riesgo
Rodrigo Alfaro, Andrés Sagner y Carmen Silva.

DTBC – 630   Mayo 2011
Jaque Mate a las Proyecciones de Consenso
Pablo Pincheira y Nicolás Fernández

DTBC – 629   Mayo 2011
Risk Premium and Expectations in Higher Education
Gonzalo Castex

DTBC – 628   Mayo 2011
Fiscal Multipliers and Policy Coordination
Gauti B. Eggertsson

DTBC – 627   Mayo 2011
Chile's Fiscal Rule as a Social Insurance
Eduardo Engel, Chistopher Neilson y Rodrigo Valdés

DTBC – 626   Mayo 2011
Short – term GDP forecasting using bridge models: a case for Chile
Marcus Cobb, Gonzalo Echavarría, Pablo Filippi, Macarena García, Carolina Godoy, Wildo González, Carlos Medel y Marcela Urrutia