Essays on economic inequality, income taxes, and intergenerational mobility

by

Pablo Gutiérrez Cubillos

BA (Honours), University of Chile, 2010

MA (Honours), University of Chile, 2014

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

(Economics)

The University of British Columbia

(Vancouver)

December 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Essays on economic inequality, income taxes, and intergenerational mobility

submitted by Pablo Gutierrez Cubillos in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Economics

Examining Committee:

Prof. Kevin Milligan, Economics, University of British Columbia

co-Supervisor

Prof. Thomas Lemieux, Economics, University of British Columbia

co-Supervisor

Prof. Nicole Fortin, Economics, University of British Columbia

Supervisory Committee Member

Prof. Claudio Ferraz, Economics, University of British Columbia University

University Examiner

Prof. Sanghoon Lee, Sauder School of Business, University of British Columbia University

University Examiner

Additional Supervisory Committee Members:

Prof. Terry Moon, Economics, University of British Columbia University

Supervisory Committee Member

Abstract

The first chapter provides the first consistent estimates of intergenerational earnings mobility in Chile, based on administrative records that link a child's and their parent's earnings from the formal private labour sector. We estimate that the intergenerational earnings elasticity is between 0.288 and 0.323, whereas the rank-rank slope is between 0.254 and 0.275. We find significant non-linearities in the intergenerational mobility measures, where intergenerational mobility is very high in the bottom 80% of the parents' distribution but with extremely high intergenerational persistence in the upper part of the earnings distribution. In addition, we find remarkable heterogeneity in intergenerational mobility at the regional level, where Antofagasta, a mining region, is the most upwardly-mobile region. Finally, we estimate significant differences across municipalities in the Metropolitan Region, where our estimates suggest that the place of residence makes a significant difference in intergenerational mobility for children of upper-class families, while it is less relatively important for children of lower- and middle-class families.

The second chapter proposes a new methodology to value retained earnings as income by transforming them into accrued capital gains and develops a parametric procedure to impute corporate retained earnings to households. We use this approach to estimate income inequality for Canada using household survey data, and aggregate retained earnings information from national accounts. We show that including retained earnings by transforming it into accrued capital gains increases income inequality in Canada and changes the trend in income inequality, exhibiting more consistency with the decline in capital income after the Great Recession.

The third chapter investigates consequences of top-distribution undercoverage on the Gini coefficient. It shows that not correcting for underreporting and nonresponse at the top does not necessarily result in an underestimated Gini coefficient. In addition, this paper proposes a Gini approximation based on the Atkinson approximation to correct for underreporting at the top. Under plausible assumptions, the approximation proposed for correcting underreporting at the top is near exact.. To evaluate this methodology, this paper uses Chile and Canada as examples where we include undistributed business profits

to measure income inequality.

Lay Summary

I present three essays on economic inequality and intergenerational mobility in the Canadian and Chilean context. In the first chapter, with co-authors, we study study earnings intergenerational mobility for Chile. For this, we use a novel administrative dataset. We show that earnings intergenerational mobility is non-linear, with very high mobility for the bottom 80 percent and very high persistence for the upper tail. In the second chapter, I study the effects of corporate retained earnings on income inequality in Canada. I show that including corporate retained earnings for the measurement of income inequality can change levels and trends of income inequality. Finally, in the third chapter, I study the Gini coefficient in the context of underreporting and nonresponse in the upper tail of the income distribution. A Gini approximation is proposed and studied.

Preface

Chapter 1 constitutes joint work with Juan Díaz and Gabriel Villarroel. The research presented in Chapter 1 is covered by UBC Behavioural Research Ethics Board Certificate number H20-02952. Chapters 2 and 3 are original, unpublished and independent work by the author, Pablo Gutierrez Cubillos. In chapter 1 co-authors were equally involved in all stages of the research project, including the identification of the research question, review of the literature, preparation and analysis of the data, and writing of the paper.

Contents

List of Tables

List of Figures

Acknowledgements

Doing a PhD can feel like a Greek epic. Sometimes you think that everything is lost (even yourself!), but then, a Deus ex Machina event happens and suddenly you are finishing your thesis without even realizing it.

My Deus ex Machina, or god from the machine, was the good guidance I received during the process, so the first person that I would like to thank is my supervisor Kevin Milligan for his unflinching support, adamant confidence, and excellent advice throughout this process. Likewise, I wouldn't be at this point if it wasn't for Nicole Fortin, who always pushed me to be a better researcher and who was the person that motivated me to write the first chapter of this thesis. I also cannot forget thanking Thomas Lemieux for his good research advice and for the valuable time we spent discussing my naive research ideas, and Terry Moon for his precise guidance. Also, I would like to thank to David Green, Craig Riddell, and Erik Snowberg for giving me the opportunity to work as an RA, and Maureen for all the good advices and help in administrative decisions.

Doing a PhD is not a process one undertakes alone. As such, I would like to thank to my mother Isabel and my father Antonio for their emotional support. Even when I called them late or frustrated about the process, they were there for me. I also want to thank tía Eli and Verito for their tremendous and altruistic support during the pandemic. Their encouragement was key to finish this research. Additionally, I would like to thank Iris for her good advice and patience during the pandemic.

Also, I would like to thank to my friends for their motivation and support. First of all, I would like to thank Juanito. We started the PhD application process together and we supported each other during different stages of this process. This comradery lasted until the very end, as the first chapter of the thesis was co authored with him. Second, I would like to thank to JF "El pelao" for his partnership and honest friendship. Despite being classmates in the same PhD cohort and roommates for two years our relationship was characterized by supportiveness rather than combativeness. Even now we still support each other with the collateral damage that a PhD might bring (aka. sindrome del gato). Moreover, I would like to thank Pablito Troncoso. His encouragement, friendship and

optimism were and are valued.

Another person that I would like to thank is Javier (aka. compañere) for his genuine friendship and support. Also, I would like to thank Victor "El maestrito", he was the very first person I met in Vancouver. We later met in Paris where he provided me with accomodation for more than 10 days. He is both a remarkable researcher and even better person. In addition, I would like to thank the support of the 1st year PhD gang, aka. Don Davide and Dona Maria, Anand, Vinicius and Sev. All the good dinners and company will never be forgotten. Moreover, help and friendship from Luchito is very appreciate. I would also like to thank my classmates: Neil, Jeff, Jasmine, Hao, Hugh and Justin.

Another key part of this process is the relationship with your co-authors (mentors). In this regard, I would like to thank to Claudio for his confidence, motivation and trust on my process. The same goes with Eugenio, my former supervisor, now co-author. Thank you for your motivation and for the good advice. I would also like to extend my thanks to Ramón who introduced me to the research business and who lent me his friendship and support during the process. In addition, I would like to thank Juan Pablo Torres-Martinez, for his patience and for being a prime example of what scholarship truly means, and Pablo Tapia for his encouragement and motivation. Moreover, the relationship with my young co-authors is also important, I am really grateful to Gaëlle, Nacho, Vania and Mancu for their patience and willingness to work with me.

The PhD process is a journey. During this journey I lived in Ottawa for 18 months. Here I worked for the Canadian government where I received tremendous support from Eric, Christine, Andrew, Patrick, Jacinthe, Carlos, Danny, George, Brian, Rebekha, Jen and Wendy (for her patience and diligence in the RDC).

Even back at home there are friends and people that help you along this process. I am indebted to Tito, Mito, Dieguito, Sergio, Didi, Isabel and Pipo for their friendship and support and all the teachers that help me during my formation such as Antonio, Luis, et al.

The PhD process is also an immigration process, it is about understanding a new

culture while also missing your own one. So, I would like to thank to all my friends that I met during this process, Mario, Roni, Rolo, Karla, Mati, Seba, Carito, Chedo, Mari, Tata et al. Many thanks for making me feel like at home.

Finally, I would like to thank to Andrea for her love, her loyal and unconditional support, even in the darkest hours, and for making this journey all the more exiting and meaningful.

A mi madre, quien siempre me ha motivado a seguir adelante a pesar de las dificultades. Sin su apoyo, ejemplo de vida y visión no existiría esta tesis. Y también a mi padre por siempre confiar en mi y por tener la sabiduría de escuchar y seguir a mi madre.

Introduction

In this thesis, I present three essays on economic inequality and intergenerational mobility in the Canadian and Chilean context. The first chapter studies earnings intergenerational mobility for Chile by estimating the correlation between parent's and Children's earnings using a novel administrative dataset. The second chapter estimate income inequality in Canada adding corporate retained earnings to households. To do this, I develop a valuation and an imputation methodology. The third chapter study the Gini coefficient in the context of underreporting and nonresponse at the upper tail. We study a very simple and useful Gini approximation to solve those data issues and we provide two empirical examples, for Canada and Chile.

In the first chapter, along with co-authors, we study intergenerational mobility in Chile by building a new and unique data set after assembling three administrative data sources. We obtain information on labour earnings of children and their parents from 2002 to 2019 from the database of the Chilean government's unemployment insurance program (UIP). We link children and their parents using administrative records provided by the Civil Registry Office. We obtain the place of residence of a child when she was between 13 and 18 years old from administrative records at the Ministry of Education. To the best of our knowledge, this is the first work that uses administrative information to estimate intergenerational mobility for a non-advanced economy.

We estimate intergenerational earnings mobility at national level. We find that it is highly non-linear in Chile, and that it is extremely mobile for the bottom 80 percent of the earnings distribution, even more mobile than in advanced economies such as the US and Canada. But it is also highly persistent for the upper decile of the earnings

distribution, much more so than for any advanced economy. This result resembles what Bratsberg et al. (2008) finds for comparing the Nordic countries with the US and UK.

We also estimate intergenerational earnings mobility at the regional level. This is of particular interest for a country like Chile, where the climate and economic conditions are significantly heterogeneous across its geography. We find that the most mobile region is Antofagasta, which is a miner-intensive region located in the north of the country. This result is in line with the findings for developed economies (Australia and Canada). Meanwhile, the least mobile region is Araucanía, where about a third of the population is ethnic Mapuche (an indigenous population) - the highest proportion of any region in Chile.

Finally, we estimate intergenerational mobility across different municipalities for the Santiago Metropolitan Region. This region contains the nation's capital, Santiago, one of the cities with a better quality of life in South America. We find that Santiago is extremely heterogeneous in upward mobility, circles of poverty and circles of privilege. In particular, there is a cluster of rich municipalities where the conditional probability that child stays in the fifth quintile given that the parent was in the fifth quintile of their earnings distribution is higher than 0.7. Those rich municipalities are quite similar in terms of upward mobility.

We also make a methodological contribution; we use for the first time tools to estimate intergenerational mobility at the top of the distribution such as RIF regressions and Kernel conditional densities. In addition, we estimate the Gatsby curve for Chile and Santiago using two measures of intergenerational mobility: absolute intergenerational mobility and relative intergenerational mobility. We show that the Gatsby curve could be valid for a persistence indicator but not for an absolute mobility indicator. Meaning that inequality could be related with persistence at the top instead of mobility at the bottom.

Of course, there is a vast body of literature from economists trying to learn about social mobility from administrative records in advanced economies. For the United States, there is a series of articles that are based on a project by Raj Chetty, Nathaniel Handren and others, who use administrative tax data to estimate the intergenerational elasticity of

income.[1] For example, the work of Chetty et al. (2014) studies how social mobility varies through geographic zones called community zones in the US. For Canada, the literature on intergenerational income mobility starts with the seminal work of Corak and Heisz (1999), a pioneering paper in the use of administrative data to study intergenerational mobility of income. More recently, Corak (2019) studies intergenerational mobility in Canada utilizing census data and analyzing data at various geographic levels. Europe has also produced some interesting literature in this regard. For instance, Acciari et al. (2019) use tax data to investigate how intergenerational mobility varies geographically for Italy, as do Güell et al. (2015) for social mobility at smaller geographical units in Italy, which Heidrich (2015) also does for Switzerland. Most of these works for developed countries show that disaggregated geographical measures of intergenerational mobility provide evidence of significant heterogeneities across locations that are hidden in country-level estimates.

In the case of Chile, our work does not emerge in a vacuum. Over the last two decades, some papers have made progress in understanding social mobility by using survey data. For example, Núñez and Miranda (2010, 2011) study intergenerational income mobility by using the Two-Sample Two-Stage Least Squares (TSTSLS) methodology developed by Björklund and Jäntti (1997). Sapelli (2013) provides evidence on changes in the intergenerational mobility of education through time, using several cross-section surveys. Meanwhile, Torche (2005) analyzes the intergenerational mobility of education based on survey data, and Celhay et al. (2010) focus on the study of intergenerational mobility of income and schooling for the period 1996-2006 using longitudinal surveys. The only paper that uses administrative records to capture a specific dimension of intergenerational mobility in Chile is the work of Zimmerman (2019). Based on a regression discontinuity design, this article exemplifies the lack of upward mobility by showing that studying at an elite university has a positive effect on obtaining a managerial position with high income in the labour market, but only for those with a high-level socioeconomic background who had studied at an elite private school. In part, this study quantifies the importance of contact networks in the generation of inequality in Chile.

[1] In this paper, we make the distinction between earnings, for which the source is wages, and income, for which the sources are wages and financial asset income. Our study is developed with earnings due to the available dataset.

The second chapter studies the effect of corporate retained earnigns on income inequality. Thus, the use of retained earnings allows us to have an accrual measure of capital gains. However, the money inside the firm has a different value from that outside the firm. If an agent wants to get the money out of the company, he or she has to pay personal taxes (depending on the tax system he or she may receive a tax credit for the corporate taxes paid by the firm). The financial market may adjust for those future taxes decreasing the capital gains generated by this retained earning. That is, the tax system should be taking into account to measure the income from capital gains associated with retained earnings. In addition, if we consider that the marginal investor is a foreigner, as Boadway and Bruce (1992) states, the domestic tax system is irrelevant for the marginal investor. Thus, it is crucial to identify what is the right tax rate that should be used to value retained earnings as capital gains. The same is true if there are some transaction costs to get the money out of the firm. Thus, the first contribution of this work is to develop a conceptual framework that analyzes the effect that the ownership of the firm, the tax system and the equilibrium in capital markets have on the capital gain generated via retained earnings.

We do not have access to administrative data. To overcome this limitation, we propose a parametric methodology to impute corporate retained earnings to families using household survey microdata and aggregate national account information. This procedure is based on the exponential-Pareto model established by Dragulescu and Yakovenko (2000) and Silva and Yakovenko (2004).[2] In addition, this method follows the spirit of Jenkins (2017) and Hundenborn et al. (2018). It uses survey data for a fraction p of the population and aggregate national account data as an additional source for the remaining $1-p$.[3] To evaluate the pertinence of this parametric imputation methodology, we compare it with a non-parametric imputation procedure a capitalization approach similar to that applied

[2]Others studies that use an exponential distribution for the bottom part of the income distribution are Banerjee et al. (2006) and Jagielski and Kutner (2013).

[3]The use of survey data is justified because in some countries it is not mandatory for individuals making certain income levels to file a tax declaration; thus, data that are just generated only from tax declarations may have a bias in the lower tail of the income distribution. Also, some developing countries do not have another reliable data source than a household survey, or it is politically difficult to get access to administrative data.

by Saez and Zucman (2016).[4] This method contributes to an extensive literature that estimate income and wealth inequality using parametric methods such as Kleiber and Kotz (2003), Chotikapanich, Griffiths and Rao (2007), Clementi and Gallegati (2016), among many others.

The valuation and imputation methodology is tested empirically for Canada, where we impute corporate retained earnings and then use the generated data to compute income inequality measures. To do so, we use the Survey of Consumer Finances (SCF) for 1984 and the Survey of Financial Security(SFS) for 1999, 2005, 2012 and 2016. The justification for using these surveys is that they can be harmonized, allowing us to make a comparison with the capitalization method used by Saez and Zucman (2016) because those surveys have rich information on assets in addition to income.[5] The inequality measures estimated here are not as precise as those estimated by Saez and Veall (2005), Veall (2012) and Wolfson et al. (2016). Despite this, it has the value of correcting a household survey for a form of under-reporting such as Burkhauser et al. (2011), Bourguignion (2018), Blanchet, Flores and Morgan (2018) among many others.

Empirically, we find that the inclusion of corporate retained earnings and its measure as accrued capital gains increase the estimated measure of income inequality and that this also affects the trend in income inequality. Indeed, for 2005, the share of the top 1% increases by 4.5 percentage points (from 7.8% to 12.3%), and the Gini coefficient increases by 4.4 points (from 47.5 to 51.9) which implies higher income inequality than in 2012 and 2016; this was not the case before accrued capital gains were considered. Those results are robust to the method used to impute corporate retained earnings. In this context, this work contributes to a broader literature that study Canadian income inequality such as Saez and Veall (2005), Fortin et al. (2012), Lemieux and Riddell (2015), Milligan and Smart (2015), Wolfson et al (2016), Green, Riddell and St-Hilaire (2016), among many others.

[4]One reason to establish another imputation procedure is that, as Kopczuk (2016) states, "In an environment with a low rate of return, a small bias in the estimated rate of return has large consequences on the estimations of wealth inequality."

[5]Dividends could also be used to impute corporate retained earnings. However, as Alstæaeter et al. (2017) shows, this is not a good procedure to impute retained earnings because (i) retained earnings and dividends move in different directions, (ii) mechanically, the imputations results are not adequate in periods in which the aggregated retained earnings are negative.

Chapter 3 studies the effects of two types of undercoverages at the top of the income distribution: underreporting (i.e., missing income) and nonresponse (i.e., missing people). Underreporting occurs when individuals in a population report less income or wealth than they earn (e.g., tax evasion, top coding, information omissions in household surveys). On the other hand, nonresponse occurs when individuals in a population are unrecorded in the data source (i.e., truncated data; e.g., people not submitting their household surveys or not declaring taxes). Bourguignon (2018), Lustig (2018), Blanchet, Flores and Morgan (2018) recently studied these two missing-information types. Their works discuss how adjustments for these biases affect income-inequality measures. In particular, Lustig (2018) develops a taxonomy to differentiate the different types of undercoverage at the upper tail. Bourguignon (2018) shows, in a didactic manner, how different adjustments in the upper tail affect the income distribution. In particular, he argue that the adjustments of the original data relies on three key parameters: i) How much is to be allocated to the top of the distribution; ii) how broad should the top b; iii) what share of the population should be added to the top. Finally, Blanchet, Flores and Morgan develops a novel methodology to find the point where tax data describes better the income distribution than survey data. Their method can be used to correct for underreporting and nonresponse at the top.

In this chapter we depart from Bourginon (2018) and Blanchet, Flores and Morgan, instead of studying the whole income distribution, we only study the effects underreporting and nonresponse in the Gini coefficient. Our first contribution is that we demonstrate that not correcting for underreporting and nonresponse at the top does not necessarily result in an underestimated Gini coefficient.

To correct the Gini coefficient for undercoverage at the top, Atkinson (2007) proposes a simple and pragmatic approximation. He uses household-survey information and tax data, and he approximates the Gini index as $G = G_{1-p}(1 - S_p) + S_p$, where G_{1-p} is the Gini coefficient computed from a household survey representative of a population's poorest $1 - p$ percent, and where S_p is the income share owned by the population's top p percent (e.g., the share of the top 1%) and computed from income-tax data. Alvaredo (2011) further develops this procedure and analytically derives and extends his formula,

proposing an exact Gini decomposition to be used when a p proportion of the population is not well measured in a data source but is better measured in another source. However, Alvaredo's decomposition requires additional information: it depends on (i) the value of p and (ii) the income distribution within the $1-p$ population.[6] Some scenarios lack information on either of these elements (e.g., measuring either income inequality adding undistributed profits or tax-haven wealth[7]). A modified version of the Atkinson approximation can be used under such information scarcity and thereby can correct the Gini coefficient.

Thus, this chapter's second contribution is that it proposes a simple approximation of the Gini coefficient in the case of underreporting at the top which is a slightly modified version of the traditional Atkinson approximation without the necessity of knowing the size of the top p. In addition, the approximation's analytical bias is computed. In addition, we show that the bias is higher when the traditional Atkinson approximation is used for solving nonresponse and underreporting instead of the new adjusted formula to correct underreporting. It shows, numerically, that the proposed approximation is near exact when used to correct the Gini coefficient for underreporting but may be heavily upward biased for correcting the Gini coefficient for nonresponse. That is, in order to use the underreporting methodology we need to first correct for missing people at the top.

Thus, this paper's third contribution is to propose and apply a methodology for estimating the missing proportion at the upper tail. Thus, we can estimate an underreporting and nonresponse corrected Gini coefficient by estimating the proportion of nonrespondants and then apply the underreporting correction. It applies this methodology to two countries: Chile and Canada, and corrects the income Gini coefficient by adding undistributed business profits, a source of capital income underreported in household surveys, and administrative-tax-declaration data. Indeed, as Smith, Yagan, Zidar and Zwick (2019) argue "a primary source of top income is private "pass-through" business profit, which can include entrepreneurial labour income for tax reasons", thus some

[6] As was discussed by Cowell and Flachaire (2015) and Higgins, Lustig and Vigorito (2018) estimating this p is a major challenge.

[7] Issues that are tremendously relevant for inequality measurement, see for instance, Alstadsæter et al. (2017, 2018).

part of labour income is transformed into capital income and left inside the firm. Indeed, some tax reforms induces to keep business income inside the firm whereas others generate incentives to take out profits as dividends (see for instance the 2003 US dividend tax reform). Thus, not accounting for undistributed business income when we measure income inequality could lead to artificial changes that bias levels and trends of income inequality estimates. Thus, the methodology developed here could be used to estimate level and trends of income inequality that are robust to tax changes and tax avoidance behaviour.

Chapter 1

Intergenerational mobility in Chile

1.1 Introduction

This paper asks whether the association between parents' and their child's earnings in Chile varies with parental earnings level and children's place of residence. Chile is an interesting case study not only due to having made significant progress in its economic development in the last three decades (reaching a GDP per capita of US$ 16,143 in 2018, IMF, 2018) but also because it is one of the countries with the most unequal income distribution in the world. It has a Gini index of 0.477 points (World Bank, 2017), and the fraction of the country's total income received by the richest 10% of the population is extremely high (37.1%) when compared to the OECD average of 24.7% (OECD, 2018). Moreover, conservative estimates suggest that the share of total income that the richest 1% take is 15%, while less conservative estimates establish it at 22-26% (Fairfield and Jorrat, 2016; Flores et al. 2019).

Under what conditions an unequal society can be tolerated is a subject of long-standing debate, especially in Chile. Supporters of meritocracy argue that economic inequality can be legitimated in a society if income differences stem from differences in reward for talent, hard work and skill, but not due to luck or transmission of advantages. According to this view, income inequality should not be tolerated in a society with less social mobility and greater transmission of privileges or disadvantages from parent to child,

where children born in poverty (richness) remain in poverty (richness) in their adulthood, regardless of their skills or efforts. In part, the de-legitimization of income inequality is one of the main causes behind the social outbreak that occurred in Chile in October 2019, when the perception of unfairness in the distribution of income and privileges provoked lower and middle classes to take to the streets to express their indignation with the current situation. In this context, understanding social mobility in Chile is crucial to disentangle the origins of its current levels of economic inequality.

In this paper, we study intergenerational mobility in Chile by building a new and unique data set after assembling three administrative data sources. We obtain information on labour earnings of children and their parents from 2002 to 2019 from the database of the Chilean government's unemployment insurance program (UIP). We link children and their parents using administrative records provided by the Civil Registry Office. We obtain the place of residence of a child when they were between 13 and 18 years old from administrative records at the Ministry of Education. To the best of our knowledge, this is the first work that uses administrative information to estimate intergenerational mobility for a non-advanced economy.

We estimate intergenerational earnings mobility at the national level. We find that it is highly non-linear in Chile, and intergenerational mobility is very high for the bottom 80 percent of the earnings distribution, and exceed the rate of intergenerational mobility in advanced countries such as the US and Canada. But earnings are also highly persistent for the upper decile of the earnings distribution, much more so than for any advanced economy.[1] We can summarize this finding as: intergenerational churning, and socio-economic uncertainty, for the masses contrasts with secure inherited privilege for the elite.

We also estimate intergenerational earnings mobility at the regional level. This is of particular interest for a country like Chile, where the climate and economic conditions are significantly heterogeneous across its geography. We find that the most mobile region is Antofagasta, which is a minering intensive region located in the north of the country.

[1]This result resembles what Bratsberg et al. (2008) find when comparing the Nordic countries with the US and UK.

This result is in line with the findings for developed economies (Australia and Canada). Meanwhile, the least mobile region is Araucanía, where about a third of the population is ethnic Mapuche (an indigenous population) - the highest proportion of any region in Chile.

Finally, we estimate intergenerational mobility across different municipalities for the Santiago Metropolitan Region. This region contains the nation's capital, Santiago, one of the cities with a better quality of life in South America. We find that Santiago is extremely heterogeneous in upward mobility, circles of poverty and circles of privilege. In particular, there is a cluster of rich municipalities where the conditional probability that child stays in the fifth quintile given that the parent was in the fifth quintile of their earnings distribution is higher than 0.7. Those rich municipalities are quite similar in terms of upward mobility.

We also make a methodological contribution, we use for the first time tools to estimate intergenerational mobility at the top of the distribution such as RIF regressions and Kernel conditional densities. In addition, we estimate the Gatsby curve for Chile and Santiago using two measures of intergenerational mobility: absolute intergenerational mobility and relative intergenerational mobility. We show that the Gatsby curve could be valid for a persistence indicator but not for an absolute mobility indicator. Meaning that inequality could be related with persistence at the top instead of mobility at the bottom.

Of course, there is a vast body of literature from economists trying to learn about social mobility from administrative records in advanced economies. For the United States, there is a series of articles that are based on a project by Raj Chetty, Nathaniel Handren and others, who use administrative tax data to estimate the intergenerational elasticity of income.[2] For example, the work of Chetty et al. (2014) studies how social mobility varies through geographic zones called community zones in the US. For Canada, the literature on intergenerational income mobility starts with the seminal work of Corak and Heisz (1999), a pioneering paper in the use of administrative data to study intergenerational mobility of income. More recently, Corak (2019) studies intergenerational mobility in Canada

[2]In this paper, we make the distinction between earnings, for which the source is wages, and income, for which the sources are wages and financial asset income. Our study is developed with earnings due to the available dataset.

utilizing census data and analyzing data at various geographic levels. Europe has also produced some interesting literature in this regard. For instance, Acciari et al. (2019) use tax data to investigate how intergenerational mobility varies geographically for Italy, as do Güell et al. (2015) for social mobility at smaller geographical units in Italy, which Heidrich (2015) also does for Switzerland. Most of these works for developed countries show that disaggregated geographical measures of intergenerational mobility provide evidence of significant heterogeneities across locations that are hidden in country-level estimates.

In the case of Chile, our work does not emerge in a vacuum. Over the last two decades, some papers have made progress in understanding social mobility by using survey data. For example, Nuñez and Miranda (2010, 2011) study intergenerational income mobility by using the Two-Sample Two-Stage Least Squares (TSTSLS) methodology developed by Björklund and Jäntti (1997). Sapelli (2013) provides evidence on changes in the intergenerational mobility of education through time, using several cross-section surveys. Meanwhile, Torche (2005) analyzes the intergenerational mobility of education based on survey data, and Celhay et al. (2010) focus on the study of intergenerational mobility of income and schooling for the period 1996-2006 using longitudinal surveys. The only paper that uses administrative records to capture a specific dimension of intergenerational mobility in Chile is the work of Zimmerman (2019). Based on a regression discontinuity design, this article illustrates the lack of upward mobility by showing that studying at an elite university has a positive effect on obtaining a managerial position with high income in the labour market, but only for those with a high-level socioeconomic background who had studied at an elite private school. This study, in part, quantifies the importance of contact networks in the generation of inequality in Chile.

1.2 Development of the income intergenerational mobility literature

Are the children of the poor doomed to stay poor? Are the children of the rich destined to stay rich? How difficult is it for someone who was born poor to belong to the middle class

during her adulthood? These questions have been addressed at the international level, where there is vast literature on intergenerational income mobility. Jäntti and Jenkins (2015) and Corak (2013) summarize the historical results in this literature. Corak and Heisz (1999) were the first to use high-frequency administrative data on the income of parents and children in adulthood in their seminal study on intergenerational mobility in Canada.[3] This study was so innovative and ahead of its time that it took 15 years for literature to replicate this study for other countries. In fact, thanks to the development of computer science and generalization in the use of administrative data, the literature of intergenerational mobility has been given a new lease of life. The works of Chetty et al. (2014), Chetty et al. (2017), and Chetty et al. (2018a, 2018b) have extensively studied intergenerational mobility in the United States using the same type of data.

Undoubtedly, the novelty of these studies is in the data used, which mainly correspond to confidential high-frequency administrative data that cover a sufficiently long period and link the income of the parents with the adult income of their children. The advantage of administrative data is that they do not have the traditional problems present in household surveys. In fact, traditional household surveys in general are not longitudinal but cross-sectional, which makes it difficult to obtain information on the income of the parent and child in adulthood. In addition, household surveys have problems such as sampling, self-reporting and non-response, and it is known that non-response rises as the respondent's income increases (Bollinger et al., 2018).

Understanding the intergenerational mobility of income in the United States has been tremendously important in understanding the generation of inequality. There is a series of articles that are based on a project by Raj Chetty, Nathaniel Handren and others, who use administrative tax data to estimate the intergenerational elasticity of income. The work of Chetty et al. (2014) studies geographic zones called community zones. The abovementioned investigation by Chetty and others differentiate between absolute and relative intergenerational mobility, which has been of interest to both politicians and researchers. The Canadian literature on intergenerational income mobility starts with the

[3]Others important studies on intergenerational mobility for Canada are Fortin and Lefebvre (1998), and Simard-Duplain and St-Denis (2020)

seminal work of Corak and Heisz (1999), pioneering in the use of administrative data to study intergenerational mobility of income. More recently, Corak (2019) studied intergenerational mobility in Canada, using census data and analyzing intergenerational mobility within Canada at a geographic level. Acciari, Polo and Violante (2019) investigate intergenerational mobility for Italy by taking tax data, also analyzing what happens geographically. Finally, this literature has also progressed in Europe, mostly based in the Nordic countries. Jäntti (2006) illustrates very well the use of these data. Also, there are the studies for Switzerland by Heidrich (2015) and Güell et al. (2015) for Italy. Both studies are at the provincial and inter-country levels.

1.2.1 Intergenerational mobility of income, the case of developing countries

Research on intergenerational mobility of income in developing countries faces additional complications. Having longitudinal data that gather parents and children is very difficult (Daude and Robano, 2015, Neidhöfer, 2019, Neidhöfer et al., 2018) due to the limitation of household surveys and/or the difficulty of accessing administrative data.

One way to address the limitations of the data is to restrict the analysis to children and parents living in the same household or to impute an income for the parents based on multiple waves of a household survey. For example, Lambert et al. (2014) studies intergenerational mobility in Senegal and Torche (2014) summarizes intergenerational mobility in Latin America from studies that have used surveys as a primary source of information.

Recently, progress has been made to investigate intergenerational mobility using census data from 26 African countries (Alesina et al., 2019) and for the regions of India, Asher et al. (2018). In this context, our research project will be pioneering in Latin America because it uses administrative data, which is the way in which the frontier literature is studying intergenerational mobility.

1.2.2 Intergenerational mobility of income, regional differences

Recent literature has concentrated on studying the regional differences that exist within countries.[4] They find that regional intergenerational income mobility behaves differently among countries. Chetty and Corak find differences among regions, where there are certain territories that have less intergenerational mobility than other parts. However, for Switzerland, Heidrich (2015) does not find many differences. In the Chilean case, Núñez and Miranda (2011) find that the intergenerational mobility of income is higher in Santiago compared to the Chilean average. Inequality has been studied at the regional level in Chile. However, how regional intergenerational mobility varies in Chile has not been studied.

1.3 Data

1.3.1 Information on labour earnings

We obtain the information on labour earnings of children and their parents from the database of the UIP in Chile. The UIP is a benefit that covers all employees in the private sector over 18 years old and with a formal contract, whether fixed-term or permanent. Participation in the scheme is mandatory for all contracts started after September 2002 and voluntary for contracts started before that date. This means that these administrative records contain the monthly labour earnings of all employed workers over the age of 18 who initiated a work-under-contract relationship in the private sector from October 2002 to December 2019. This data set also includes the workers with labour contracts established prior to October 2002 who voluntarily joined the UIP. It is worth mentioning that this data set excludes workers with training contracts, workers under the age of 18, domestic workers, pensioners, self-employed or own-account workers, and public sector employees.

Table 1.1 provides information on the proportion of workers covered by the UIP over several years. As can be seen, due to the voluntary retroactive nature of the UIP

[4]See Chetty et al. (2014), Chetty et al. (2018a, 2018b), Corak (2019), Güell et al. (2015), Heidrich (2015) Connolly et al. (2018).

policy, the coverage rate for formal contract workers was below 50% in 2003 and 2004. In the following years, this coverage rate significantly increased, attaining 65% in average in 2005-2007 and 80% in 2012. Part of the 20% of formal contract workers still not covered by the UIP in 2012 are public sector employees, who are covered under a similar but separate scheme. Table 1.1 also shows information on workers covered by the UIP as a proportion of the total labour force. Initially, the labour force coverage rate was 42% in average for the years 2003-2007, which rapidly converged to 65% in 2012. The 35% not covered by the UIP in 2012 is explained by public sector employees, the unemployed, and informal workers.[5]

Table 1.1: Representativity of the unemployment insurance program dataset.

Year	Total UIPD	W ENE	Coverage W	LF ENE	Coverage LF
2003	1349.5	3672.7	36.7%	5119.1	26.3%
2004	1849.5	3806.3	48.6%	5286.1	34.9%
2005	2337.8	3987.4	58.6%	5438.7	43.0%
2006	2701.3	4166.4	64.8%	5442.2	49.6%
2007	3103.1	4360.3	71.1%	5555.5	55.8%
2008	3309.2	4583.5	72.2%	5762.4	57.4%
2009	3419.8	4500.1	76.0%	5839.9	58.6%
2010	3742.4	4908.1	76.2%	6210.1	60.2%
2011	4050.4	5146.7	78.7%	6448.8	62.8%
2012	4286.4	5360.2	80.0%	6520.0	65.7%

This dataset is compared with the information of the ENE (Encuesta Nacional de Empleo) questionnaire administred by the goverment statistics agency in Chile (INE-Instituto Nacional de Estadisticas). W ENE refers to the total number of formal employees recorded by ENE and LF ENE is the total labour force (formal and informal) recorded by ENE. The information regarding ENE numbers is from Sehnbruch and Carranza (2015). Units are measured on thousands

[5]As we can see, this dataset converges to a coverage rate of 80% of the formal workers but only to 65% for the total labour force. This is in part because this dataset has limited coverage for the (cont'd) unemployed. Sehnbruch (2006) and Ruiz-Tagle and Sehnbruch (2010) argue that this is because a large proportion of unemployed register by ENE previously worked in the informal sector.

We must acknowledge that the low formal contract workers' coverage rate during the first years of the data (56% in average in 2003-2007) is a concern for our analysis because —as explained below— it impacts how we model permanent parental earnings for our baseline sample. To assess the plausibility of our findings, we perform a robustness exercise. We frame our analysis using data for years with a higher formal contract workers' coverage rate to construct the permanent parental earnings.

1.3.2 Information on child-parent linkage

We link children and their parents using administrative records provided by the Civil Registry Office (CRO). In Chile, the CRO registers all births, deaths, and marriages. It is a legal requirement in Chile that all births must be registered in the CRO, each of which is backed by a birth certificate. This birth certificate contains the information on the child and the parents given at the time of registration. We use the information provided for all the birth certificates in Chile to build the pairs of children and parents included in the UIP database.[6] In our baseline analysis, the sample of children is composed of individuals that were 28-33 years old in 2018, while the sample of parents are individuals that were 42-87 years old in 2018.

1.3.3 Measurement of earnings

Our administrative records have information on labour earnings in the formal private sector, excluding any form of capital income for the workers covered by the UIP. In our baseline sample, we measure parental earnings as the 5-year average of monthly earnings for months worked in the formal private sector between 2003 and 2007. For example, if a parent records 30 months worked within a 5-year period, the measure of earnings used is the total income in those 5 years divided by 30. In our baseline sample, we only consider parents that worked at least 6 months in the formal private sector during 2003-2007. If both parents worked more than 6 months in the period, we consider the average parental

[6]Families are ever changing, so the parenting person or persons at any point in time may not be the birth parents.

earnings as the sum of parental earnings divided by two, in line with Chetty et al. (2014) and Corak (2019).

Our measure of parental earnings excludes the zeros because a zero in our data set does not mean that the individual has no earnings, since he/she could be earning as a public employee, in the informal sector, or in the formal private sector but not covered by the UIP, especially in its earlier years.

As with the parents, we measure child earnings in our baseline sample as the five-year average of monthly earnings for worked months in the formal private sector between 2014 and 2018. In our baseline analysis, we consider children that worked at least six months in the formal private sector in 2014-2018. This measure of child earnings not only excludes the zeros for the same reasons as for their parental earnings, but also because children may start participating in the private formal labour market in their late 20s, giving a series of months with earnings preceded by a series of zeros corresponding to not being in the labour market.

To minimize the noise provoked by low earners due to the uncertainty surrounding the low earnings registered with the UIP, we only consider children and parents who on average earn more than half the minimum wage.[7] In our baseline sample, we have 505,524 parent-child links.

1.3.4 Comparison between unemployment insurance program dataset and ENE survey

In Chile, 29.6 percent of the population works the informal sector. One potential issue for our dataset is that only contains information on private formal earnings. To see how different are the percentiles including all workers, we compare the earnings percentiles generated by our dataset and the Encuesta Nacional de Empleo (ENE).

[7]Half the minimum wage for children is $133,000 in 2019 Chilean pesos (measured from 2014 to 2018) and $103,000 in 2019 Chilean pesos for parents (from 2003 to 2007). Using CASEN 2017 information, 14.1 percent of the population were under the minimum wage.

Table 1.2: Comparison of earnings between our dataset and ENE for individuals between 28-33 years old.

Percentile	UIP	ENE
1%	152,889	170,613.6
5%	218,433	231,840
10%	263,508	250,902
25%	343,076	330,000
50%	490,707	451,624
75%	767,851	700,000
90%	1,173,052	1,003,609
95%	1,544,161	1,304,692
99%	237,1979	2,500,000

This dataset is compared with the information of the ENE (Encuesta Nacional de Empleo) questionnaire administred by the goverment statistics agency in Chile (INE-Instituto Nacional de Estadisticas). W ENE refers to the earnings percentiles for all workers – formal, informal and self employed. Units are in 2018 Chilean pesos.

Table 1.2 compares our dataset earnings percentiles with ENE dataset percentiles for 2018. We can see that percentiles similar using the whole population and types of sector and the formal private sector.

1.3.5 Information on child residential address

We link the pairs of child and parental earnings with the residential address of the child while attending 12th grade in school. We obtain this information from administrative records provided by the Ministry of Education of Chile. If the child's residential address while attending 12th grade is not available, we use the most recently-available residential address while she was enrolled from 7th to 11th grade in school (when the child is 13-18

years old).[8] We end up with 93.95% of the children's sample linked to their residential address.

1.4 Intergenerational mobility for Chile

We begin our empirical analysis by characterizing the relationship between parental and child earnings at the national level. We present a set of baseline estimates of relative intergenerational mobility and then evaluate the robustness of our estimates to alternative samples.

1.4.1 Traditional indicators of intergenerational mobility

Intergenerational earnings mobility

One of the most commonly used measures of intergenerational mobility is the intergenerational earnings elasticity, i.e., the effect that a 1 percent increase in the parental earnings has over their child's earnings. In our work, we estimate the intergenerational elasticity of earnings rather than of income because our dataset only contains information on wages and not on financial asset income. We measure this elasticity by estimating the following equation:

$$y_i^c = \alpha + \beta y_i^p + \epsilon_i, \tag{1.1}$$

where y_i^c is the earnings of child i in logarithms, y_i^p is the earnings of that child's parents in logarithms, and β is the intergenerational earnings elasticity. This parameter is equal to

$$\beta = \frac{cov(y_i^p, y_i^c)}{var(y_i^p)} = \rho \cdot \frac{sd(y_i^c)}{sd(y_i^p)}, \tag{1.2}$$

where ρ is the intergenerational earnings correlation, and $sd(y_i^c)$ and $sd(y_i^p)$ are the standard deviation of child and parental log earnings, respectively. To prevent any attenuation bias, we measure child and parental earnings as the 5-year average of earnings.

[8] We also estimate our results by making the geographic link from 5th to 12th grade. The results are similar.

Table 1.3: OLS estimates of the intergenerational earnings elasticity for our baseline linkage

	(1)	(2)	(3)	(4)
y_p	0.288***	0.297***	0.311***	0.323***
	(0.001)	(0.001)	(0.002)	(0.002)
Constant	9.506***	9.426***	9.298***	9.193***
	(0.016)	(0.018)	(0.021)	(0.027)
Observations	505,524	416,818	282,979	173,683
R-squared	0.091	0.098	0.108	0.117

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Earnings are measured as average earnings over the months where a children-parents pair report positive earnings over the studied 5-year period. We keep individuals that appear at least 6 times with positive earnings in the dataset with average earnings greater than half of the corresponding minimum wage. Columns (1) to (4) report results for male and female children. (1) considers individuals with at least 6 months of positive earnings, (2) considers individuals with at least 12 months of positive earnings, (3) considers individuals with at least 24 months of positive earnings and (4) considers individuals with at least 36 months of positive earnings.

Table 1.3 summarizes our estimates for intergenerational earnings elasticity (IGE), i.e., the OLS estimates of the regression slope of the log child earnings on log parental earnings. Columns (1) to (4) report results for male and female children: (1) considers individuals with at least 6 months of positive earnings (our baseline sample); (2) considers individuals with at least 12 months of positive earnings; (3) considers individuals with at least 24 months of positive earnings; and, (4) considers individuals with at least 36 months of positive earnings.

Our baseline estimation for IGE equals 0.288. With our most restrictive sample —individuals with at least 36 months of positive earnings—, this estimate equals 0.323. This means that an increase of 10 percent in parental earnings implies, on average, an increase of between 2.88 and 3.23 percent in their child's earnings.[9]

Tables 1.4 and 1.5 estimate the IGE for female and male children respectively.

[9]This estimate is lower compared with previous estimates in the Chilean literature. Nunez and Miranda (2010,2011), and Celhay et al. (2010) estimate an elasticity between 0.5 and 0.6. Our differences can be explained by the kind of data used and the method implemented to estimate IGE. Appendix C discusses this point in detail.

Table 1.4: OLS estimates of the intergenerational earnings elasticity for female children

	(1)	(2)	(3)	(4)
y_p	0.300***	0.307***	0.315***	0.326***
	(0.002)	(0.002)	(0.003)	(0.003)
Constant	9.253***	9.209***	9.169***	9.086***
	(0.024)	(0.026)	(0.032)	(0.042)
Observations	222,397	178,916	116,182	68,644
R-squared	0.103	0.111	0.119	0.128

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Earnings are measured as average earnings over the months where a children-parents pair report positive earnings over the studied 5-year period. We keep individuals that appear at least 6 times with positive earnings in the dataset with average earnings greater than half of the corresponding minimum wage. Columns (1) to (4) report results for male and female children. (1) considers individuals with at least 6 months of positive earnings, (2) considers individuals with at least 12 months of positive earnings, (3) considers individuals with at least 24 months of positive earnings and (4) considers individuals with at least 36 months of positive earnings.

Table 1.5: OLS estimates of the intergenerational earnings elasticity for male children

	(1)	(2)	(3)	(4)
y_p	0.282***	0.294***	0.314***	0.329***
	(0.002)	(0.002)	(0.002)	(0.003)
Constant	9.655***	9.529***	9.313***	9.175***
	(0.022)	(0.024)	(0.028)	(0.036)
Observations	283,127	237,902	166,797	105,039
R-squared	0.087	0.094	0.107	0.117

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Earnings are measured as average earnings over the months where a children-parents pair report positive earnings over the studied 5-year period. We keep individuals that appear at least 6 times with positive earnings in the dataset with average earnings greater than half of the corresponding minimum wage. Columns (1) to (4) report results for male and female children. (1) considers individuals with at least 6 months of positive earnings, (2) considers individuals with at least 12 months of positive earnings, (3) considers individuals with at least 24 months of positive earnings and (4) considers individuals with at least 36 months of positive earnings.

Our results suggest that female children are slightly less intergenerationally mobile than male children.

Rank-rank correlation

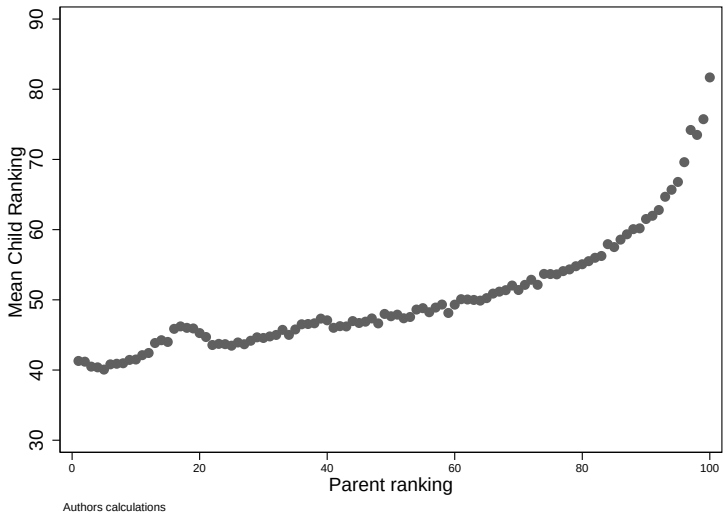
Another measure of intergenerational mobility that has become extremely popular is rank-rank correlation. This correlation measures the effect that an increase of a percentile in the parental earnings distribution has over the child earnings distribution. One of the arguments to use rank-rank correlation is that the rankings on the earnings distribution are determined at earlier ages and are difficult to change throughout the age distribution. We measure this correlation by estimating the following equation by OLS:

$$r_i^c = \alpha^r + \beta^r r_i^p + \epsilon_i, \tag{1.3}$$

where r_i^c is the ranking of i-th child in the national distribution of child earnings by cohorts, r_i^p is the ranking of i-th child's parent on the national distribution of parental earnings, and β^r is the rank-rank correlation.[10] This correlation is an indicator of relative mobility that compares the maximum influence of parental ranking on expected child ranking. In addition, α^r is a measure of absolute mobility because it states the expected ranking that a child would have if her parents belong to the bottom of the parental earnings distribution.

[10]Note that we compute the ranking of the whole cohort of children and parents, regardless of whether they are linked.

Figure 1.1: Expected child ranking conditional on parental ranking



We estimate the expected child ranking non parametrically using a simple average. Rankings were computed over the national distribution. For children we compute the cohort ranking, and for parents we compute the ranking of people 42-87 years old (in 2018).

Figure A.3 presents a binned scatter plot of the mean percentile rank of children versus their parents' percentile rank. This graph illustrates a nonparametric estimation of the conditional expectation of a child's rank given her parents' rank ($E[r_i^c|r_i^p = p]$). As we can see, the relationship between parental ranking and child ranking is close to a linear function until the 80th parental percentile, while for parental percentiles higher than 80 it is highly non-linear with an increasing gradient as the parental ranking increases.

Table 1.6 presents our estimates for the rank-rank slope. To measure the percentile rank of the children, we consider their rankings in the distribution of child earnings within their birth cohorts. In the same way, we compute the percentile rank of the parents from their positions in the distribution of parental earnings in the baseline sample. Based on the child and parental percentile ranks, the rank-rank slope estimate is the OLS estimate of the regression slope of the percentile rank of a child on the percentile rank of her parents. As before, columns (1)-(4) in Table 1.6 present the results for 6 (baseline sample), 12, 24, and 36 months of positive earnings. The rank-rank correlation is between 0.254 and 0.275, that is, the maximum expected difference in child earnings rankings that

Table 1.6: OLS estimates of the rank-rank correlation for our baseline linkage

	(1)	(2)	(3)	(4)
r^p	0.254***	0.261***	0.270***	0.275***
	(0.001)	(0.001)	(0.002)	(0.002)
Constant	37.397***	38.668***	40.859***	43.368***
	(0.080)	(0.089)	(0.110)	(0.141)
Observations	505,524	416,818	282,979	173,683
R-squared	0.064	0.068	0.073	0.078

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Earnings are measured as the average earnings over the months in which an individual reports positive earnings over 5 years. We keep children-parents linkages that appear at least 6 times with positive earnings in the dataset with average earnings greater than half of the corresponding minimum wage. Columns (1) to (4) report results for male and female children. (1) considers individuals with at least 6 months of positive earnings, (2) considers individuals with at least 12 months of positive earnings, (3) considers individuals with at least 24 months of positive earnings and (4) considers individuals with at least 36 months of positive earnings.

depends on parental ranking is between the 25th and 28th child earnings percentiles.

Table 1.7 show the rank-rank correlation estimates only for female children, and Table 1.8 show rank-rank correlation estimates for male children. Comparing female and male, results show that the rank-rank correlation is higher for female children. This indicates that for females, parental ranking is more persistent than for males. In addition, absolute mobility, measured as the constant of each regression, is higher for males than for females, which means that male children of poor parents are expected to locate in a higher ranking than female children of poor parents.

Quintiles transition matrices

These child and parental earnings rankings also allow us to estimate the quintile transition probabilities. These probabilities are defined by the conditional probability that a child is in quintile m (with $m = 1, 2, 3, 4, 5$) of the child earnings distribution given that her parent is in quintile n (with $n = 1, 2, 3, 4, 5$) of the parental earnings distribution.

In the intergenerational mobility literature, there are three probabilities that are

Table 1.7: OLS estimates of the rank-rank correlation for our female children

	(1)	(2)	(3)	(4)
r^p	0.278***	0.285***	0.293***	0.300***
	(0.002)	(0.002)	(0.003)	(0.004)
Constant	31.669***	33.234***	35.762***	38.362***
	(0.121)	(0.138)	(0.176)	(0.233)
Observations	222,397	178,916	116,182	68,644
R-squared	0.075	0.079	0.083	0.088

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Earnings are measured as the average earnings over the months in which an individual reports positive earnings over 5 years.We keep children-parents linkages that appear at least 6 times with positive earnings in the dataset with average earnings greater than half of the corresponding minimum wage. Columns (1) to (4) report results for male and female children. (1) considers individuals with at least 6 months of positive earnings, (2) considers individuals with at least 12 months of positive earnings, (3) considers individuals with at least 24 months of positive earnings and (4) considers individuals with at least 36 months of positive earnings.

Table 1.8: OLS estimates of the rank-rank correlation for our female children

	(1)	(2)	(3)	(4)
r^p	0.239***	0.247***	0.258***	0.264***
	(0.002)	(0.002)	(0.002)	(0.003)
Constant	41.682***	42.504***	44.118***	46.312***
	(0.103)	(0.114)	(0.139)	(0.175)
Observations	283,127	237,902	166,797	105,039
R-squared	0.060	0.064	0.070	0.076

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Earnings are measured as the average earnings over the months in which an individual reports positive earnings over 5 years. We keep children-parents linkages that appear at least 6 times with positive earnings in the dataset with average earnings greater than half of the corresponding minimum wage. Columns (1) to (4) report results for male and female children. (1) considers individuals with at least 6 months of positive earnings, (2) considers individuals with at least 12 months of positive earnings, (3) considers individuals with at least 24 months of positive earnings and (4) considers individuals with at least 36 months of positive earnings.

broadly studied: i) the circle of poverty, defined by the probability that, given parents who belong to the bottom quintile, the child will also belong to the bottom quintile. We denote this probability as p_{11}; ii) the circle of privilege, defined by the probability that, given parents who belong to the top quintile, the child will belong to the top quintile. We denote this probability as p_{55}; and, iii) the rags to riches, defined by the probability that, given parents who belong to bottom quintile, the child will belong to the top quintile. We call this probability p_{15}. Notice that p_{11} and p_{55} are measures of intergenerational persistence that provide evidence on transmission of disadvantages and advantages, respectively; while p_{15} is a measure of upward intergenerational mobility.

Table 1.9: Transition matrix of parental earnings quintiles to child earnings quintiles

		Child quintile				
		1	2	3	4	5
Parental quintile	1	0.271	0.235	0.204	0.170	0.120
	2	0.236	0.235	0.213	0.186	0.130
	3	0.206	0.223	0.220	0.200	0.150
	4	0.171	0.193	0.215	0.223	0.198
	5	0.112	0.125	0.161	0.226	0.376

Quintiles are measured using earnings and the baseline dataset. Rows refer to parental quintile and columns to child quintiles.

Table 1.9 shows the matrix of quintile transition probabilities using our baseline sample. As can be seen in Table 1.9, p_{11} is equal to 0.271 meaning that a child whose parents belong to the bottom quintile has an observed probability of 27.1 percent of remaining in the bottom quintile; p_{55} is equal to 0.376, which means that a child whose parents belong to the top quintile has a probability equal to 37.6 percent of remaining in the top earnings quintile; and p_{15} is equal to 0.120 which means that the probability that a child whose parents belong to the bottom quintile will herself belong to the top quintile is 12 percent.

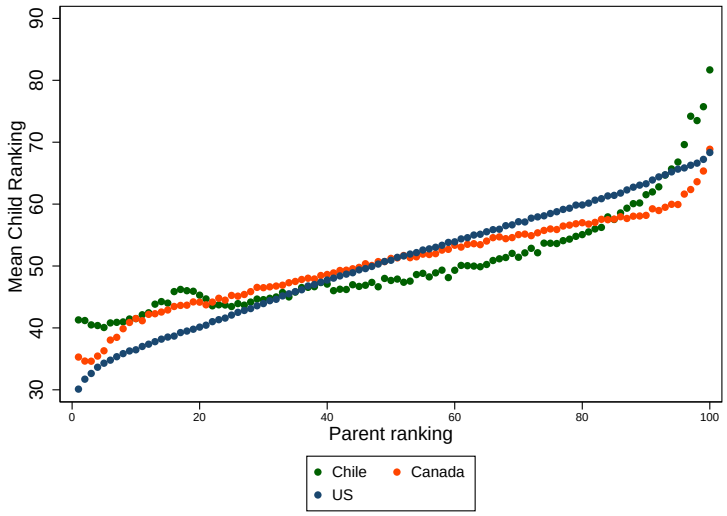
Our results suggest that there is some persistence of parental earnings because

p_{55} and p_{11} are higher than 0.2, which is the value of a transition probability, assuming that parental-child transitions are random. We also find that $p_{55} > p_{11}$ meaning that persistence is higher at the top of the distribution than at the bottom. Notice that the transition probabilities of the first 4 quintiles are relatively similar and close to random transitions; however, our results reveal that the main departure from randomness occurs at the top quintile where there is a notorious intergenerational earnings persistence.

International comparison with the US and Canada

To put our analysis in perspective, we can compare Figure A.3 with findings for the US and Canada. As reference, we use the results in Chetty et al. (2014) for the US, and the findings in Corak (2019) for Canada. Notice that, whereas for Chile we use earnings information, the works of Corak (2019) and Chetty et al. (2014) use income information.[11]

Figure 1.2: International comparison of expected child earnings ranking conditional to the parental earnings ranking



We estimate the expected child ranking non parametrically using a simple average. Rankings were computed over the national distribution. We compute the cohort ranking for children and for parents we compute the ranking of people between 42 and 87 years old (in 2018). Information for Canada is from Corak (2019) and for the US is from Chetty et al. (2014).

[11]Studies show that income is more persistent than earnings, especially at the bottom of the distribution. Thus, our results for Chile can be interpreted as a lower bound for persistence.

Figure 1.2 shows that Chile has a flatter gradient until the 80 percent in parental income/earnings. This evidence suggests that Chile is more mobile than Canada and the US in parental income/earnings until the 80th percentile. Remarkably, after the 80th parental percentile, Figure 1.2 also shows that the relationship between parental and child earnings in Chile becomes much steeper than those in the US and Canada. This graphical analysis suggests that intergenerational earnings mobility for Chile is much more non linear than the results found by the US and Canada.

1.4.2 More on non-linearities

The previous graphical analysis suggests that the relationship between parental and child earnings in Chile is highly non-linear, even more so than in the US and Canada, with the particularity of displaying significant intergenerational mobility until the 80th parental earnings quintile but a notorious degree of persistence of privileges (transmission of advantages from parent to child) at the top of the earnings distribution.

To better understand this finding, we perform three empirical exercises. First, we show the estimates of the transition probabilities for the top decile and percentiles. Second, we estimate the conditional distribution of child earnings given a parental decile (percentile), for different parental deciles (percentiles). Finally, we again estimate the IGE equation but, instead of using OLS, we use conditional and unconditional quantile regressions.

Decile and percentile intergenerational transition matrices

We now present decile transition probabilities. These estimates allow us to gain deeper understanding on how the child earnings distribution behaves within quintiles —especially for children with parents in the top quintile. Table 1.10 shows the matrix of decile transition probabilities.

As can be seen in Table 1.10, the transition matrix —excluding the row with the 10% richest parents— shows a somewhat intergenerationally-mobile context, with all the transition probabilities roughly close to 10%, as we would expect under random

Table 1.10: Decile Transition Matrix

		Child deciles									
		1	2	3	4	5	6	7	8	9	10
Parental deciles	1	0.158	0.136	0.125	0.114	0.104	0.098	0.086	0.073	0.065	0.042
	2	0.125	0.123	0.118	0.114	0.107	0.099	0.094	0.085	0.075	0.059
	3	0.122	0.126	0.124	0.114	0.111	0.102	0.095	0.085	0.072	0.050
	4	0.109	0.115	0.118	0.114	0.110	0.104	0.101	0.091	0.079	0.060
	5	0.106	0.107	0.112	0.117	0.112	0.108	0.102	0.093	0.083	0.061
	6	0.096	0.104	0.108	0.109	0.110	0.111	0.106	0.099	0.091	0.066
	7	0.086	0.093	0.100	0.104	0.109	0.112	0.110	0.107	0.100	0.080
	8	0.078	0.084	0.087	0.095	0.102	0.108	0.114	0.115	0.117	0.100
	9	0.067	0.069	0.073	0.080	0.091	0.099	0.110	0.127	0.140	0.143
	10	0.044	0.042	0.044	0.051	0.059	0.071	0.092	0.122	0.173	0.301

transition from parent to child. However, given the parental earnings top decile, we notice that the dynamic of the transition probabilities is significantly different. For instance, the probability of persistence in privilege p_{1010} is equal to 0.3. In contrast, the probability of persistence in poverty p_{11} is close to a half of p_{1010}, suggesting that the transmission of advantages (circle of privilege) is twice as persistant as the transmission of disadvantages (circle of poverty).

We now study p_{1010} in depth by showing the probabilities associated with transitions from parental percentiles to child percentiles, for percentiles from 91 to 100. Table 1.11 summarizes this information.

As can be seen in Table 1.11, the transition probabilities for children whose parents belong to the 91st to 95th percentiles of the parental earnings distribution are relatively similar, while the probability of persistence at the top percentile, $p_{100,100}$, is significantly higher compared to the rest of transition probabilities presented in Table 1.11. This means that the top percentile is even more persistent than the rest of the 10th decile. In sum, this analysis provides evidence supporting a high persistence at the top, which increases as long as parental earnings increase.

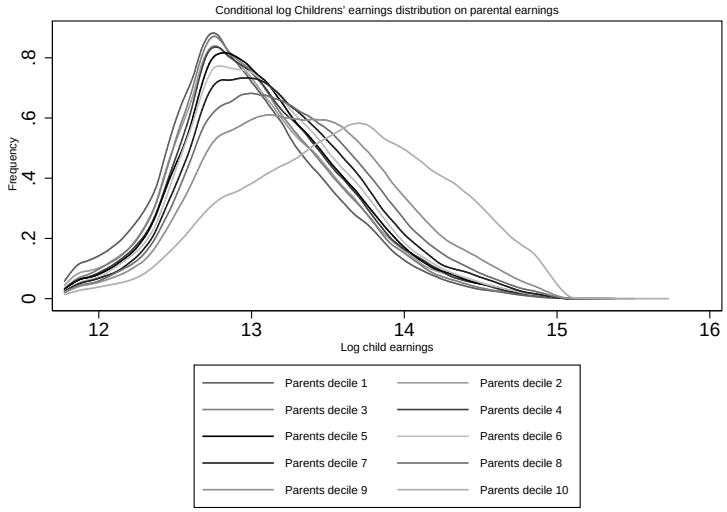
Conditional distribution of child earnings, given parental deciles

Another way to understand the association between child and parental earnings is by estimating the conditional distribution of child earnings, given parental earnings $f(y^c|y^p)$. Thus, instead of just observing a change in the mean, we can study variations in the entire distribution. To do this, we perform kernel estimations of the conditional distribution of child earnings, given parental deciles.

Table 1.11: 91st to 100th parental percentile to 91st to 100th child percentile transition matrix

		Child percentiles									
		91	92	93	94	95	96	97	98	99	100
Parental percentile	91	0.019	0.018	0.017	0.019	0.020	0.019	0.017	0.019	0.019	0.023
	92	0.020	0.017	0.019	0.020	0.019	0.020	0.018	0.018	0.022	0.021
	93	0.019	0.021	0.020	0.019	0.024	0.019	0.025	0.023	0.021	0.023
	94	0.021	0.022	0.021	0.022	0.028	0.025	0.022	0.021	0.024	0.025
	95	0.024	0.022	0.018	0.027	0.025	0.026	0.029	0.031	0.028	0.027
	96	0.026	0.025	0.026	0.026	0.028	0.026	0.032	0.034	0.033	0.039
	97	0.024	0.025	0.032	0.033	0.035	0.038	0.042	0.043	0.050	0.056
	98	0.024	0.023	0.026	0.029	0.035	0.038	0.042	0.042	0.045	0.057
	99	0.026	0.029	0.030	0.029	0.032	0.043	0.039	0.047	0.055	0.066
	100	0.027	0.035	0.035	0.029	0.045	0.051	0.056	0.060	0.068	0.105

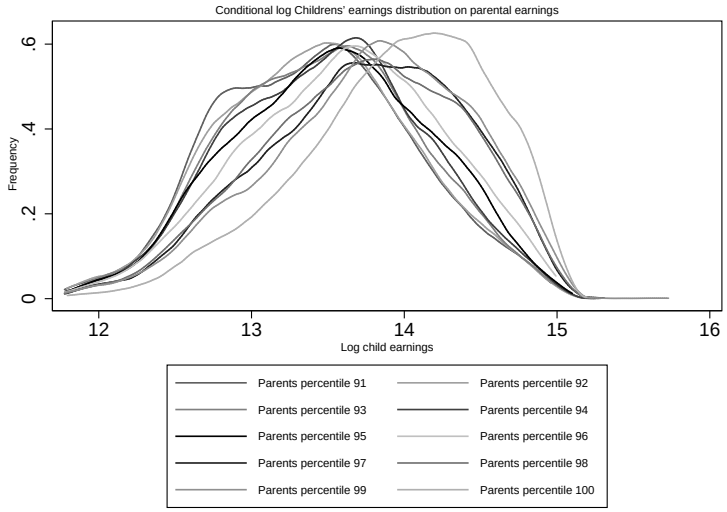
Figure 1.3: conditional (on parental deciles) child earnings distribution



This figure estimates conditional (on parental deciles) child earnings distribution, using kernel to estimate child earnings distribution. We use the Epanechnikov method to estimate optimal bandwith.

Figure 1.3 shows the conditional distribution of the logarithm of child earnings given that parents belong to a particular earnings decile, for earnings deciles from 1 to 10. As can be seen in Figure 1.3, roughly speaking, the conditional distributions of child earnings are unchanged between parental decile 1 and 7. After decile 8, it tends to move. Indeed, conditional on parents belonging to the top decile, the conditional distribution of log child earnings is significantly shifted to the right. This evidence is consistent with our previous findings of transmission of privileges, since it suggests that it is more likely for children whose parents belong to the top earnings decile to obtain higher earnings. As can be also seen in Figure 1.3, the conditional distribution of log child earnings for top parental earnings has a higher variance than conditional on lower parental earnings. In sum, this analysis supports the idea that for children in the bottom and middle part of the earnings distribution, parental earnings do not affect their own distribution of earnings; however, child earnings located at the top of their distribution are dramatically affected by parental earnings.

Figure 1.4: onditional (on parental percentiles in the top decile) child earnings distribution



This figure estimates conditional (on parental percentiles in the top decile) child earnings distribution, using kernel to estimate child earnings distribution. We use the Epanechnikov method to estimate optimal bandwith.

Figure 1.4 presents the estimation of the conditional distribution of log child earnings, given parents that belong to a specific percentile, for percentiles from 91 to 100. As can be seen in Figure 1.4, while the conditional distribution of child earnings is quite similar for those with parents in percentiles 91 to 99, it is starkly different when we condition by parents belonging to the top 1 percent. This evidence supports our finding that the relationship between parental and child earnings is highly non-linear, even at the top parental distribution where this relationship becomes significantly more positive.[12]

Unconditional and conditional quantile regressions

The two previous empirical analyses provide evidence on non-linearities in the relationship between parental and child earnings when conditional on parental earnings. In other

[12]This result is in line with Zimmerman's (2019) findings. Zimmerman (2019) shows that studying in an elite college only increases the probability of belonging to the top managerial positions (obtaining higher earnings) if the student attends a top private high school, and he also shows that it is more likely that parents that belong to the top 1 percent can afford the tuition costs of private schools. Thus, Zimmerman's findings are one component of this persistence at the top where the transmission of privileges from parent to child would be through paying the tuition costs for attending a top private high school.

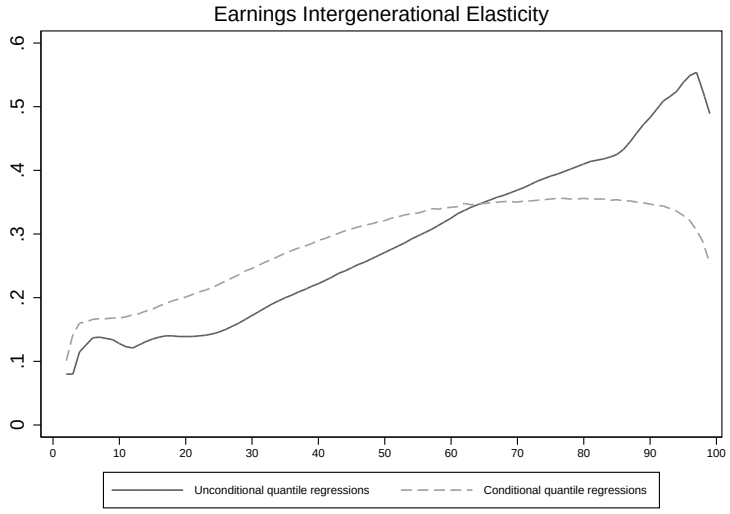
words, as long as parental earnings increase, there is a higher persistence of child earnings, especially at the top of the parental earnings distribution. However, we can also study the non-linearities in this relationship, conditional on child earnings. Specifically, given a child earnings percentile, is the effect of an increase in parental earnings stronger? We answer this question by using quantile regressions.

The IGE is estimated using OLS as the expected percent change in average child earnings, given an increase of 1 percent on the average parental earnings. Additional information regarding the relationship between parental and child earnings can be obtained by estimating the effect of a change in parental earnings on any other distributional moments of child earnings other than the mean.

We can estimate, for instance, the effect of an increase in parental earnings on the median, the 75th percentile, or the bottom 5 percentiles of the child earnings distribution. The magnitude of those effects would allow us to understand more in depth where in the child earnings distribution an increment of the parental earnings can improve their outcome. We obtain these effects by fitting quantile regressions. Following the works of Firpo, Fortin and Lemieux (2009) and Baltagi and Ghosh (2017), there are two effects of an increase in parental earnings over the quantile distribution of child earnings: a "between effect" and a "within effect". The between effect is defined by the increase in expected child earnings due to an increase in parental earnings, while the within effect is given by the change in child earnings variance associated with a change in parental earnings. Relying on these works, a conditional quantile regression allows us to estimate the between effect, and an unconditional quantile regression is useful to estimate both effects. Thus, the analysis of both methodological instruments would allow us to understand more about the observed non-linearities in the association between parental and child earnings estimated so far.[13] Figure 1.5 presents the estimates of the unconditional quantile and conditional quantile regressions in our applications.

[13] Appendix B explains with more detail the relationship between conditional and unconditional quantile regressions.

Figure 1.5: Unconditional quantile and conditional quantile estimates of the regression slope of log child earnings vs log parental earnings



Unconditional quantile and conditional quantile estimates of the regression slope of log child earnings vs log parental earnings for the 1st to the 99th child earnings percentiles. Unconditional quantile regressions are estimated using the RIF methodology developed by Firpo, Fortin and Lemieux (2009).

As can be seen in Figure 1.5, the unconditional quantile effect is lower than the conditional quantile effect until the 65th child percentile. This suggests that for the first 65 percentiles, the between effect is mitigated by the effect that an increase in parental earnings has over the child earnings variance. Meanwhile, for higher percentiles, the between effect is reinforced by the within effect, since an increment of parental earnings increases the variance of child earnings. These findings can be interpreted as that, at some point of the child earnings distribution, there are some higher-reward job opportunities that may be available that increase expected child earnings and the child earnings variance for those who can access these work positions.[14] To sum up, this analysis reveals that intergenerational earnings mobility in Chile is high and stable for the bottom 65% of the children earnings distribution; however, for the rest of the population economic status is highly persistent.

[14]This interpretation can be related with the results found by Zimmerman (2019) because the children who can access these higher-reward jobs are those with parents at the top of the earnings distribution.

1.4.3 Robustness checks

We now evaluate the robustness of our estimates of intergenerational mobility to alternative subsamples and specifications. We begin by evaluating three potential sources of bias: coverage of the dataset in initial years of the UIP, lifecycle bias, and attenuation bias.

Dataset coverage

As can be seen in Table 1.1, coverage of the unemployment insurance dataset in its first two years is less than 50% of total formal workers. To see whether this low coverage rate affects our baseline mobility estimates, we perform new estimates by considering different windows of years to measure permanent parental earnings.

Table 1.12: Estimations of IGE and rank-rank slope for different years where parental earnings were measured.

Parental year used	IGE	Rank-rank slope	N
2003-2007	0.288	0.254	504,990
2004-2008	0.288	0.256	550,668
2005-2009	0.287	0.260	584,770
2006-2010	0.284	0.263	607,545
2007-2011	0.283	0.268	622,339
2008-2012	0.281	0.270	632,820
2009-2013	0.280	0.272	636,640
2010-2014	0.278	0.272	638,481
2011-2015	0.280	0.275	637,808

Table 1.12 presents IGE and rank-rank slope estimates for different windows of years to build our measure of permanent parental earnings. We can see that IGE and rank-rank slope estimates do not depend on the choice of the window of years. Specifically, IGE estimates ranges between 0.278 and 0.288, whereas the rank-rank slope is between 0.254 and 0.275.

Lifecycle bias

Prior research has shown that measuring children's income at early ages can understate intergenerational persistence in lifetime income because children with high lifetime incomes have steeper earnings profiles when they are young (Haider and Solon, 2006, Grawe, 2006, Solon 1999). To evaluate whether our baseline estimates suffer from such lifecycle bias, we can estimate the intergenerational earnings elasticity by single child cohorts. To do this, we study the effects of parental earnings on child earnings when children are 23 to 33 years old. To be consistent with the literature (Chetty et al., 2014; Corak, 2019), we measure the effect of parental earnings when their children were teenagers.

Table 1.13: Estimates of IGE and rank-rank slope for different child ages.

Child age	IGE	Rank-rank	N
23	0.042	0.053	72,863
24	0.095	0.102	81,765
25	0.151	0.153	86,767
26	0.193	0.185	90,241
27	0.220	0.215	93,866
28	0.245	0.230	96,693
29	0.259	0.241	94,492
30	0.285	0.256	89,286
31	0.305	0.269	81,261
32	0.321	0.275	75,010
33	0.333	0.276	68,231

Table 1.13 shows the estimates of IGE and rank-rank slope by single child cohorts. We can see that intergenerational persistence rises as child age increases. This is consistent with Chetty et al. (2014). In particular, IGE is more affected by child cohorts than the rank-rank correlation, a fact that has been discussed previously in the intergenerational mobility literature.[15]

[15]Indeed, Becker and Mincer noted that if individuals can freely choose among occupations with dif-

Attenuation bias

Earnings in a single year is a noisy measure of lifetime earnings, which attenuates estimates of intergenerational persistence (Solon, 1992). To evaluate whether our baseline estimates suffer from such attenuation bias, we provide the estimates of the rank-rank slope, varying the number of years used to build our measure of permanent parental earnings.

Table 1.14: Estimates of IGE and rank-rank slope using different years to average parental earnings.

Parental years used	IGE	Rank-rank slope	N
1	0.258	0.220	156,760
2	0.272	0.235	273,673
3	0.277	0.241	363,805
4	0.284	0.248	438,302
5	0.288	0.254	505,524
6	0.291	0.258	559,666
7	0.293	0.263	603,481
8	0.293	0.267	642,176
9	0.294	0.272	676,494
10	0.294	0.275	708,541

Table 1.14 presents the estimates of the IGE and rank-rank correlations by using different numbers of years to create the permanent parental earnings. As can be seen in Table 1.14, IGE remains somewhat stable after averaging 4 years, whereas the rank-rank slope varies slightly between 0.254 and 0.275 over 4 years.

fering age/earnings profiles, an equilibrium with equality across occupations in the net present value of lifetime earnings is consistent with (indeed predicts) inequality in annual earnings (or 5 year averages of monthly earnings), both within age cohorts and overall. In this human capital equilibrium of equality in net present value, age/earnings profiles cross in the early thirties. Hence, annual incomes (or 5 year averages thereof) are plausible indicators of inequality of lifetime income for the 30-35 age cohort, but are heavily influenced by the fanning out of age/earnings profiles at later ages.

1.5 Geographic variation in intergenerational mobility: the case of Chilean regions

The previous sections suggest that the relationship between parental and child earnings varies non-linearly with parental earnings, especially with parents at the top of the earnings distribution.

Another source of variation of the relationship between parental and child earnings that has been studied in the recent literature is geographical location. The literature has found remarkable differences in intergenerational mobility across geographies within a country. For example, Connolly et al. (2018) find that commodity booms may be important drivers of intergenerational upward mobility.[16]. In addition, Deutscher and Mazumder (2020) finds the same result for Australia. Thus, a boom of the copper price can impact directly wages and the labour market in geographies that are intensive in copper production. This finding is important for Chile because it is the main copper producer in the world by a large margin, with approximately 28% of the total world production in 2018.

1.5.1 Chilean regional context

Chile is divided into 16 regions, the first-level administrative division of the country. Each region is designated by a number —from 1 to 16— and a name. Each region is divided into provinces, the second-level administrative division. In total, there are 56 provinces, each one divided into municipalities, the third and lowest-level administrative division.[17]

In Table A.10, we present current information of each region. Among the 16 regions, the Metropolitan Region (the 13th region) stands out as the most populated region in the country (in number and density), with a population of over 7.5 million in 2017

[16]Connolly et al. (2018) finds for Canada that commodity-producing provinces such as Alberta and Saskatchewan, and mid-west US states, present the highest upward mobility indicators.

[17]Until 2007, there were only 13 regions geographically located from north to south of the country with their numbers in geographically sequential order, except for the Metropolitan Region, also known as the 13th region, which is located roughly in the middle of the country, between the 5th and 6th regions. In the period 2007-2017, the 14th, 15th, and 16th regions were created after dividing into two areas the 10th, 1st, and 8th regions, respectively.

(41% of Chile's population) according to the National Institute of Statistics of Chile (INE). Significantly, this region contains the capital of Chile, the city of Santiago, which has been recognized as one of the cities with the best quality of life in South America. Based on estimates of the Central Bank of Chile (BCCh) for 2018, the Metropolitan Region produces 46% of Chile's GDP, with manufacturing, services, retail, and financial services as principal economic activities. According to official estimates by the Government of Chile for 2017, 5.4% of the population of this region lived in poverty in 2017 and this region has a GDP per capita of 3%, with a Gini coefficient of 0.43.

The Antofagasta Region (the 2nd region), in the northern area of the country, stands out with a production of 10% of Chile's GDP, with the mining industry —led by copper— as its principal economic activity. In fact, according to estimates of the BCCh for 2018, mining output represents 54% of regional production. This region had a population of 623,851 inhabitants in 2017 (3% of Chile's population) according to INE. This region has the highest GDP per capita in the country —over USD 25,000—, 5.1% of its population live in poverty in 2017, and its Gini coefficient is 0.41.

On the other end of the income scale in Chile, we have the AraucanÃa Region (the 9th region), in the southern part of Chile, which is the country's poorest region in terms of GDP per capita, with USD 6,000 per inhabitant, on average. This region contributes with 3% of Chile's GDP, with 17.2% of its population living in poverty —the highest regional poverty rate in the country. It's worth noting that a third of the region's population of 994,888 (6% of Chile's population) is of indigenous Mapuche ethnicity, which represents the highest concentration of this community (or, indeed, of any other national indigenous peoples) of any Chilean region.

1.5.2 Intergenerational earnings mobility at the regional level

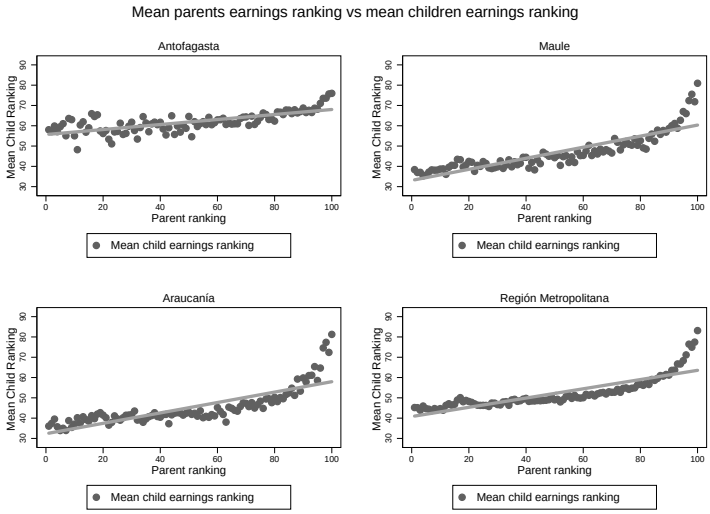
To characterize the variation in intergenerational mobility across geographic areas within Chile, we permanently assign each child to a single region. We use the child's residential address while attending 12th grade in school. We obtain this information from administrative records provided by the Chilean Ministry of Education. If the residential address

of the child when attending 12th grade is not available, we instead use the child's most recent residence while she was enrolled during 7th-11th grade in school.[18]

Measures of relative and absolute mobility

We measure mobility at the regional level using the baseline sample and the definitions of parental and child earnings described in Section 2. We continue to rank both children and parents on the basis of their positions in the national earnings distribution (rather than the distribution within their regions).

Figure 1.6: Expected child ranking conditional on parental national ranking for 4 different regions.



This figure plots the expected child ranking conditional on parental national ranking for 4 different regions. We estimate the expected child ranking non-parametrically using a simple average. Rankings were computed over the national distribution. For children we compute the cohort ranking, and for parents we compute the ranking of people between 42 and 87 years old (in 2018).

Figure 1.6 presents a binned scatter plot of the mean child rank versus parent rank for children who grew up in the second region (Antofagasta), the seventh region (Maule), the ninth region (Araucanía), and the Metropolitan region. As can be seen in Figure 1.6,

[18]The region where a child grew up does not necessarily correspond to the region she lives in as an adult at age 28-33 in 2018.

in each region there is a linear relationship between the parental and child ranks for the bottom part of the parental earnings distribution. The higher levels of persistence at the top of the parental earnings distribution are a common characteristic of the four regions displayed in Figure 1.6. Despite this non-linearity at the top of the distribution, we rely on Chetty et al. (2014) and Acciarri et al. (2020) to characterize the relationship between child rank given the parents' rank in each region using a simple linear regression. More formally, we regress child rank on parental rank by region to calculate absolute upward mobility and relative mobility by region. We define absolute upward mobility as

$$r_{abs_r} = \alpha_r + \beta_r E(r_p | r_p < 50), \tag{1.4}$$

where α_r and β_r are the intercept and the rank-rank regression slope estimated for region r, respectively. That is, the conditional expected child's position on the national earnings distribution given that her parental earnings are below the median of the national distribution. We approximate this value as $r_{abs_r} = \alpha_r + \beta_r \cdot 25$.[19] In addition, we define persistence as the conditional expectation of a child's percentile on the national earnings distribution given her parent belonging to the 10th decile. We measure this expression as $r_{per_r} = \alpha_r + \beta_r E(r_p | r_p > 90)$ and approximate it as $r_{per_r} = \alpha_r + \beta_r \cdot 95$.[20] We complement this analysis studying the three transition probabilities described in section 3. Specifically, we show transition probabilities p_{11} (circle of poverty), p_{15} (rags to riches), and p_{55} (circle of privilege).

[19] We also estimate the absolute upward mobility coefficient using a nonparametric estimation of $E(r_p | r_p < 50)$. Results remain unchanged.

[20] We also estimate $E(r_p | r_p > 90)$ nonparametrically. Results remain almost unchanged.

Table 1.15: Intergenerational mobility indicators for different Chilean regions.

Region	N	β_r	α_r	r_{abs}	r_{per}	p_{15}	p_{11}	p_{55}
1	6584	0.145	47.324	50.942	61.072	0.243	0.190	0.360
2	16911	0.146	54.351	58.012	68.265	0.321	0.126	0.451
3	9851	0.146	49.064	52.726	62.978	0.206	0.165	0.368
4	19962	0.169	42.522	46.746	58.575	0.166	0.257	0.337
5	48015	0.199	37.554	42.527	56.450	0.116	0.282	0.313
6	28806	0.244	36.332	42.441	59.548	0.112	0.310	0.368
7	28874	0.228	35.024	40.731	56.710	0.088	0.317	0.321
8	42993	0.197	38.060	42.997	56.820	0.111	0.275	0.307
9	20891	0.202	34.385	39.439	53.589	0.082	0.311	0.308
10	21105	0.197	36.002	40.932	54.735	0.091	0.255	0.305
11	4400	0.142	38.055	41.600	51.528	0.095	0.260	0.243
12	5462	0.183	39.700	44.278	57.094	0.120	0.194	0.291
13	196004	0.256	39.103	45.509	63.447	0.135	0.222	0.398
14	6631	0.178	38.998	43.454	55.931	0.094	0.250	0.349
15	4228	0.128	46.601	49.799	58.755	0.167	0.212	0.345
16	10510	0.189	37.608	42.341	55.595	0.108	0.289	0.306

As can be seen in Table 1.15, there is substantial heterogeneity across regions. For instance, the region with the highest absolute mobility is Antofagasta, where a child whose parents earn below the median national earnings level has an expected national ranking of 54.4; whereas, for Araucanía, the same child can expect to place in the 34.3(th) percentile of the child earnings distribution. In the same way, for probability p_{11} we estimate 0.126 for Antofagasta and 0.311 for Araucanía. In addition, we can notice something similar for the rags to riches probability. For Antofagasta, p_{15} is equal to 0.321 and for Araucanía is equal to 0.082, thus a child who grew up in Antofagasta with a parent that belongs to the bottom quintile is almost 4 times more likely to arrive to the top quintile than the same child who grew up in Araucanía. Finally, persistence is also higher in Antofagasta than in Araucanía: children with parents in the top earnings quintile are more likely to

remain in the top quintile in Antofagasta than in Araucanía.

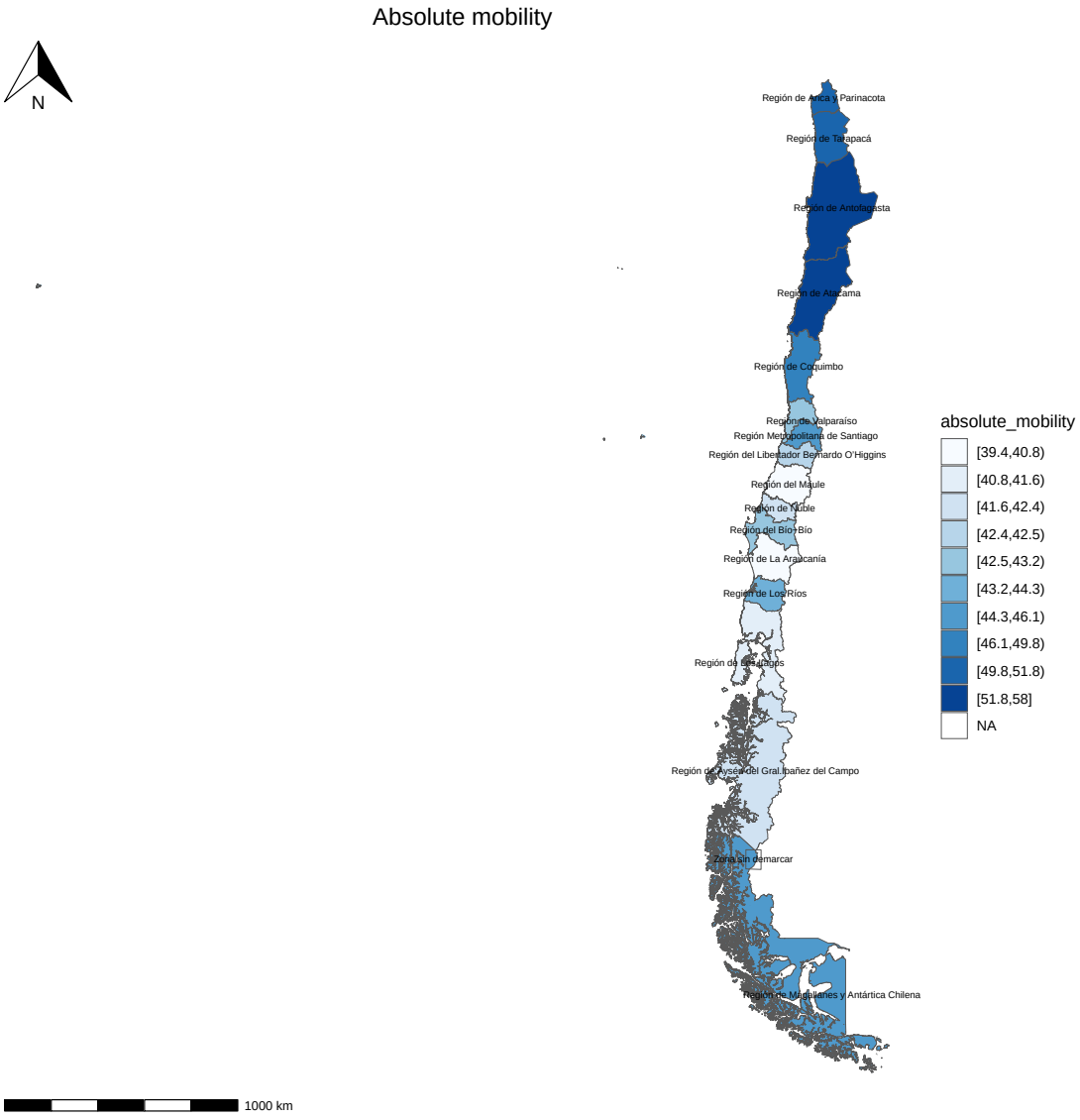
Figure 1.7 and 1.8 present heat maps of absolute upward mobility and relative mobility for Chilean regions. We can see that the most upwardly-mobile regions are those located at the north of the country. In particular, Antofagasta is the most upwardly-mobile region. Regarding relative mobility, the least mobile region is the Metropolitan region.

Figures 1.9 and 1.10 present heat maps of circle of poverty p_{11} and circle of privilege p_{55} transition probabilities for Chilean regions. We can see that the regions most persistent in poverty are those located in the upper south area of the country, particularly El Maule and Araucanía regions. In contrast, the most persistent regions in privileges are those located in the north and the Metropolitan region. Thus, we corroborate Conolly et al. (2018) results by providing evidence that Antofagasta, a commodity-intensive region, presents the highest upward mobility indicators.

Is there a Gatsby curve in Chile?

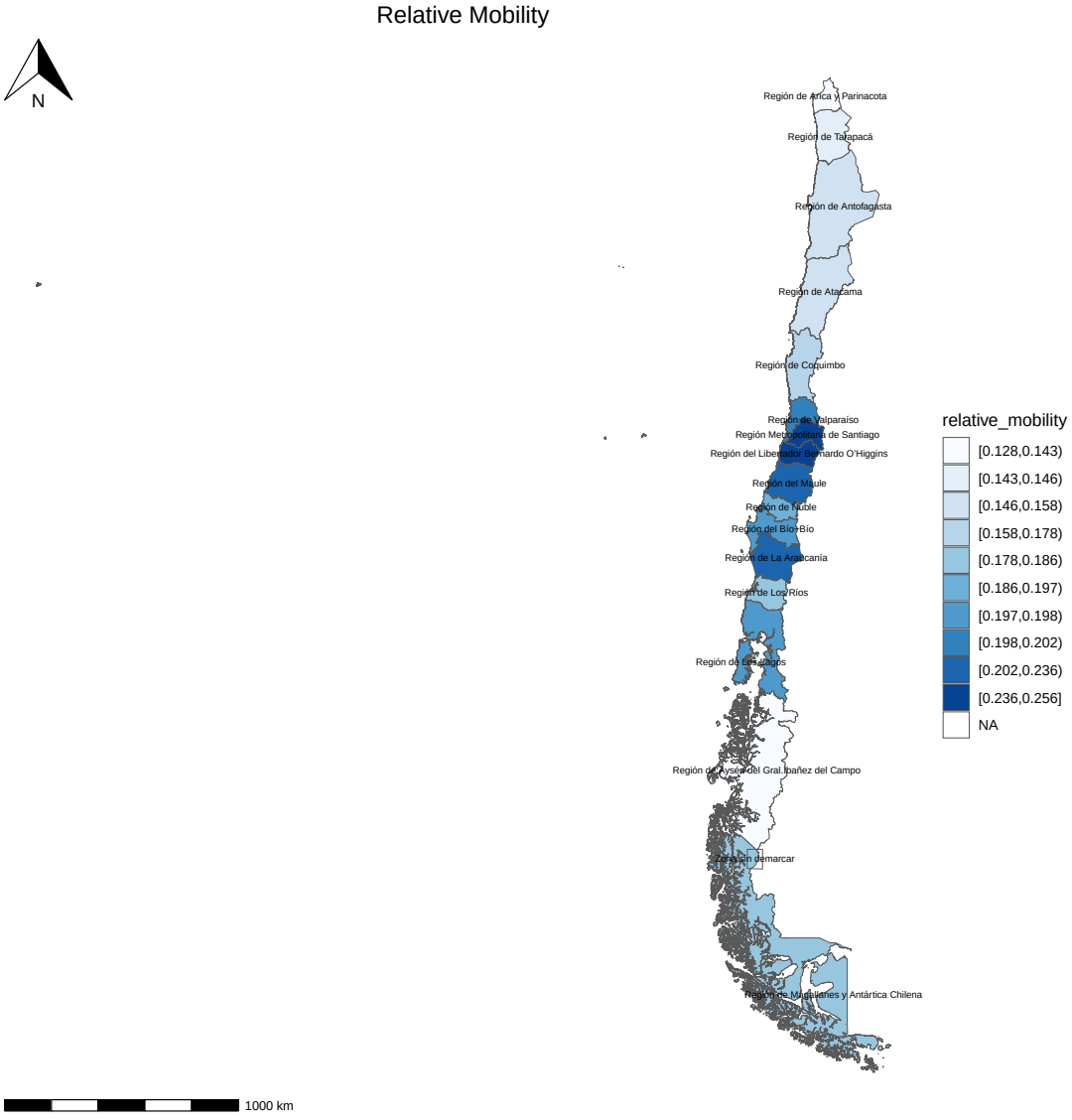
The Gatsby curve refers to the negative relationship between income inequality and intergenerational mobility. This relationship has been extensively explored by the literature (see for instance Corak, 2013). We use the geographical variation across regions in Chile to study the Gatsby curve.

Figure 1.7: Heat maps for absolute upward mobility in Chilean regions



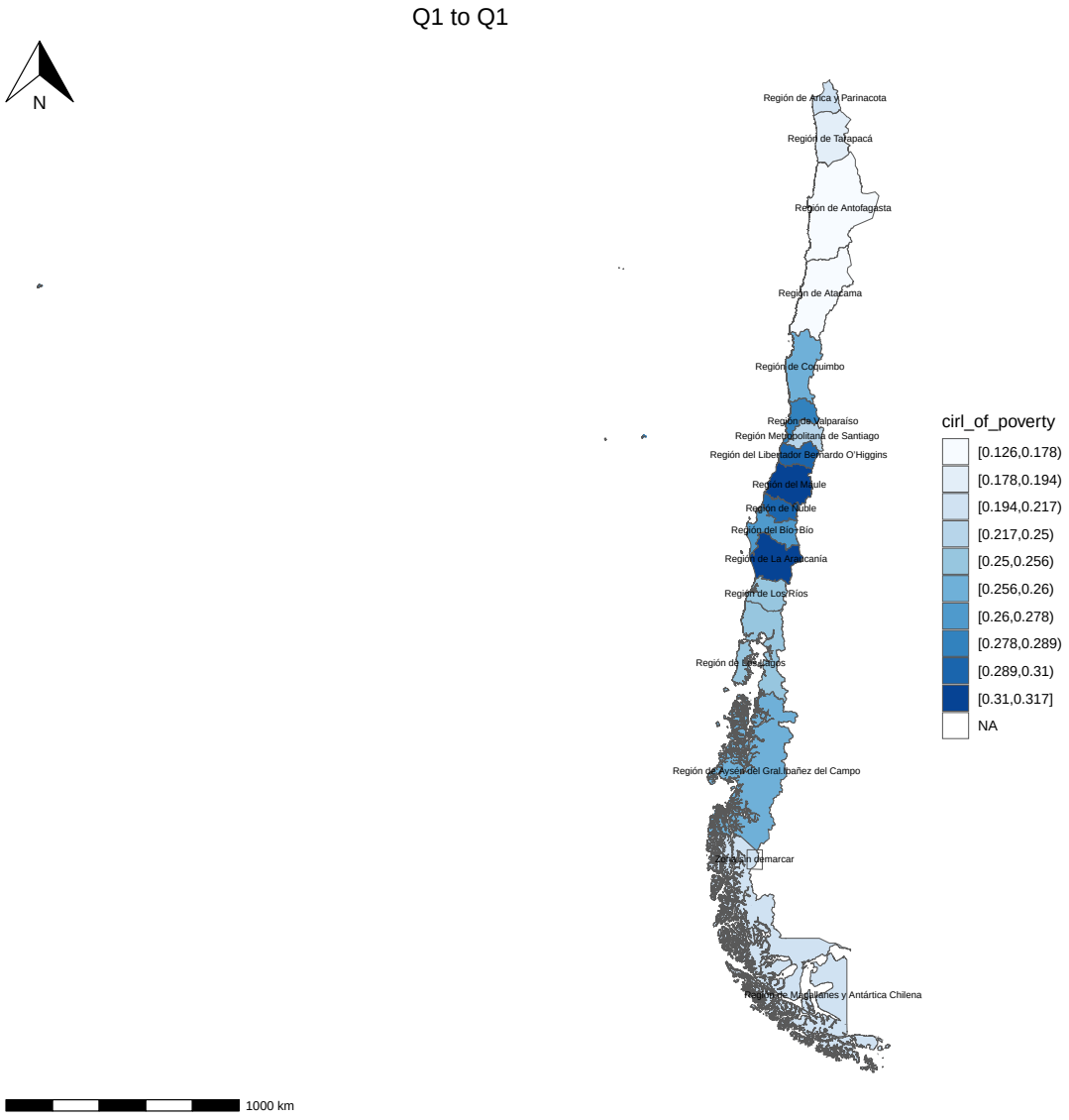
A darker blue means a higher value for the indicator

Figure 1.8: Heat maps for relative mobility in Chilean regions



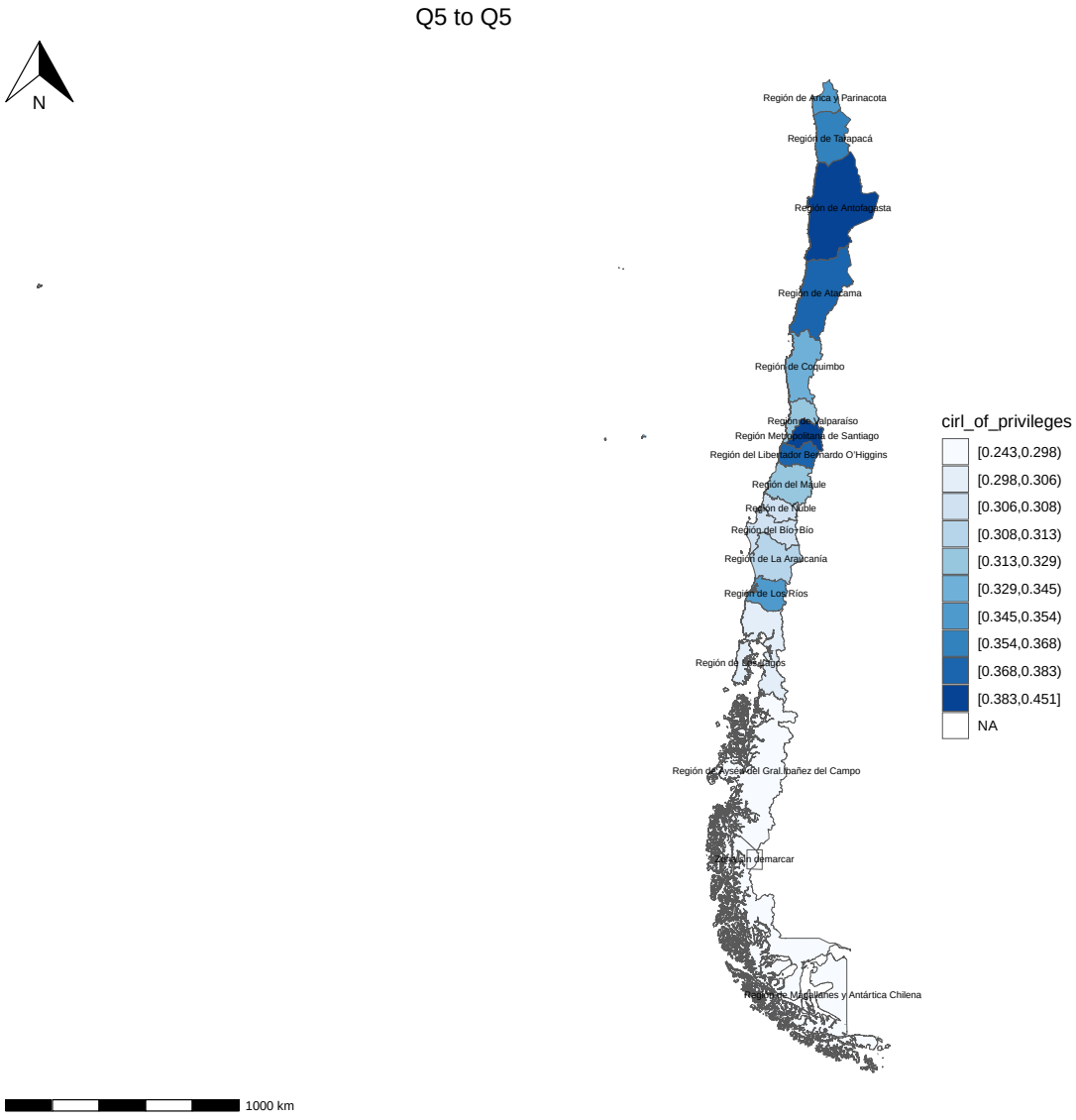
A darker blue means a higher value for the indicator

Figure 1.9: Heat maps for circle of poverty p_{11} transition probability for Chilean regions.



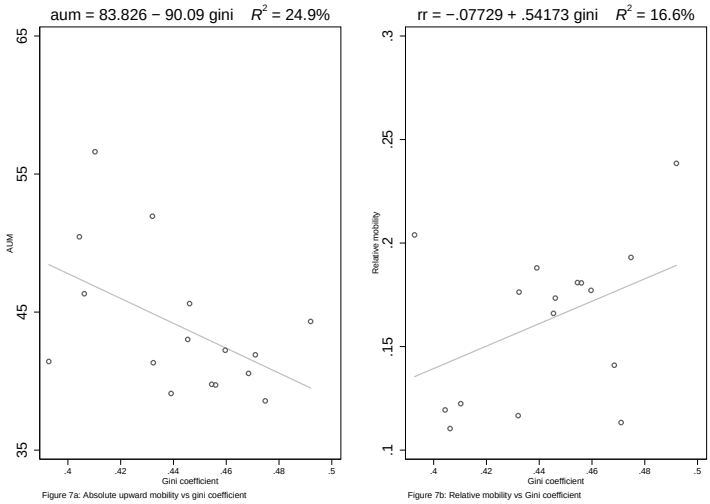
A darker blue means a higher value for the indicator.

Figure 1.10: Heat maps for circle of privilege p_{55} transition probability for Chilean regions.



A darker blue means a higher value for the indicator.

Figure 1.11: Gatsby curve Chilean regions



This figure plots the relationship between upward mobility and the Gini coefficient at the regional level. We measure the Gini coefficient using the 2017 CASEN survey, considering the total income before transfer and tax variant. Those results remain unchanged when we use other income definitions to measure the Gini coefficient.

Figure 1.11 left reports the relationship between absolute upward mobility and the Gini coefficient, while Figure 1.11 right reports the relationship between relative mobility and the Gini coefficient. As can be seen in these figures, there is evidence of a Gatsby curve, where more unequal regions experience less intergenerational earnings mobility. This evidence suggests the existence of a vicious circle between intergenerational mobility and inequality.

1.6 Geographical variation in intergenerational mobility within the Metropolitan region

We now study the intergenerational mobility across municipalities, which are the least aggregated geographic units in Chile. We do this analysis inside the Metropolitan Region of Santiago —the finance and government center of Chile. It contributes with 40% percent

of Chile's GDP, contains the capital of Santiago (the largest city in the country), and is the most densely populated region in the country, with close to 40 percent of the total population. This allows us to estimate intergenerational mobility at municipality level.

1.6.1 The Metropolitan Region

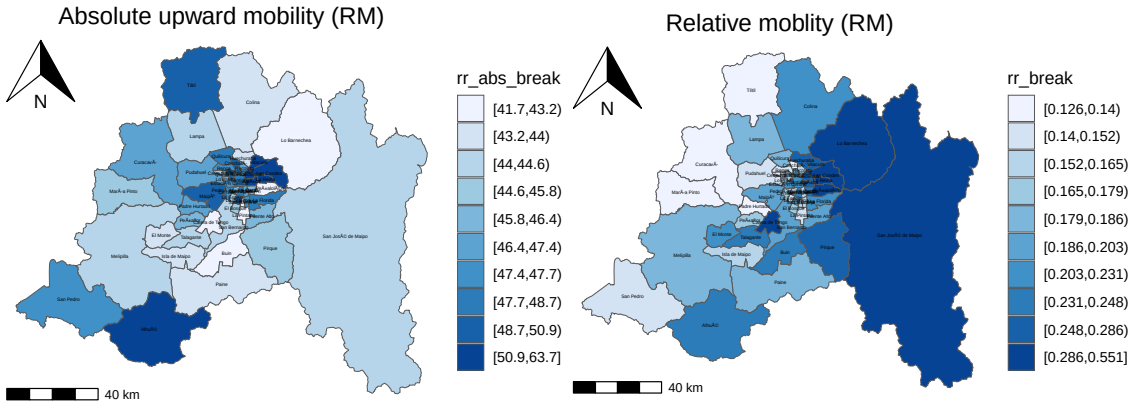
Although the Metropolitan Region of Santiago shows obvious signs of modernization, especially in the city of Santiago —which exhibits modern buildings and highways, a subway system, malls, and an extensive telecommunications network—, there are also elements that make it a residentially-segregated region, reflecting the economic inequality that characterizes the Chilean economy. Residential segregation in Santiago has its origin in several urban planning policies dating from the 1950s that tended to create residential areas for the lower classes (social housing) on the urban periphery of the city. This residential segregation intensified because of the implementation of slum eradication policies under the military dictatorship during the 1980s, where inhabitants of slum neighborhoods were relocated to social housing constructed on the periphery of the city. This policy of building social housing on the periphery continued after the return to democracy, as the proportion of social housing units in peripheral municipalities was continuously increasing and no new social housing was constructed in the upper-class municipalities.

1.6.2 Estimates of intergenerational mobility

We estimate the same measures of intergenerational mobility as for the regional case. Table A.11 in Appendix A summarizes these mobility measures by municipality.

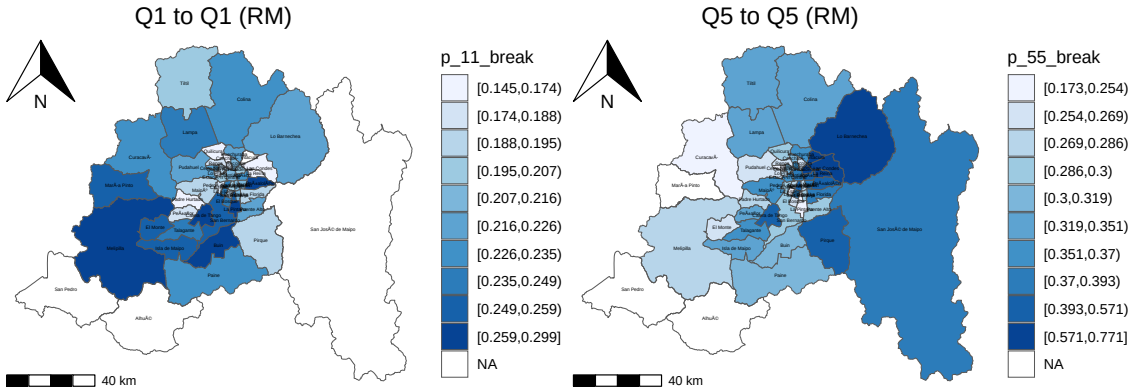
Figures 1.12 and 1.13 present color maps for intergenerational earnings mobility on the metropolitan region There is a remarkable heterogeneity across municipalities. For poor municipalities such as Cerro Navia, La Pintana and San Ramón, absolute upward mobility is not lower than 42, which means that chilndren whose parents belong to the bottom 50 percent of the earnings distribution, are expected to locate at least in the 42th percentile of the children earnings distribution. In addition, persistence at the bottom and at the top probabilities are not to far from 0.2 which means that there are not markedly

Figure 1.12: Heat maps for absolute upward mobility and relative mobility indicators for Metropolitan region municipalities.



A darker blue means a higher value for the indicator.

Figure 1.13: Heat maps for circle of poverty p_{11} and circle of privilege p_{55} transition probabilities for Metropolitan region municipalities.



A darker blue means a higher value for the indicator.

persistence. However, the rags to riches probability is lower than 0.1.

On the other hand, almost all the rich municipalities in the northeast of the city, such as Las Condes, Vitacura, and Lo Barnechea, are the most persistent municipalities at the top, with probabilities of persistence of privileges, the conditional probability that a child is in the fifth quintile given that his parent is in the fifth quintile. Lo Barnechea (0.720), Las Condes (0.672) and Vitacura (0.723) have the highest circle of privilege probability of the Metropolitan region by far, the mean of which is 0.337. Thus, for a child with a parent that belongs to the highest quintile, it is highly likely that that child will also be in the upper quintile.

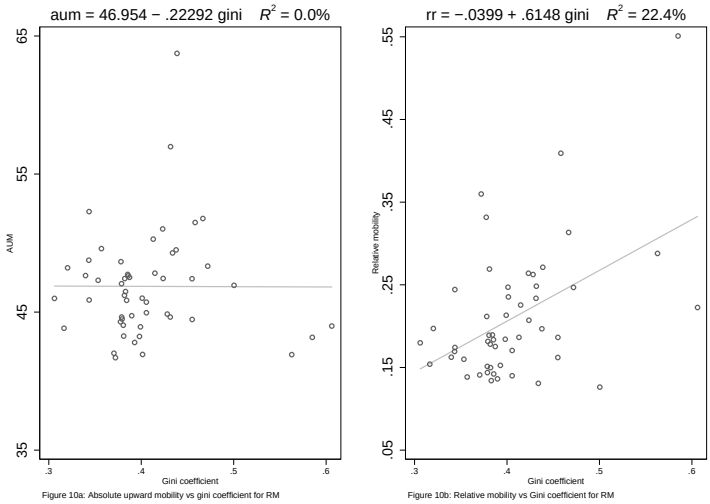
But the differences in absolute upward mobility with more middle-class municipalities such as Ñuñoa, Santiago[21] or Maipu are relatively small. For instance, absolute upward mobility in Las Condes (51.48), is very close to Ñuñoa (51.02) and Maipu (50.28). Different is the case of Lo Barnechea, where upward mobility is very low compare to the other rich municipalities and is closer to La Pintana, which is a poor municipality. The major differences on persistence of privileges found between the rich municipalities and the rest indicates that the place of residence is an important factor to explain the high persistence at the top of the earnings distribution. One possible explanation for this finding is that social connection may play an important role on persistence of privileges.

The Gastby curve in the Metropolitan region

We can study the relationship between intergenerational mobility and inequality inside the metropolitan region.

[21]Santiago is the name of the city and also the name of a municipality —the latter is the statistic presented in this table. The municipality of Santiago is what inhabitants refer to as "downtown" and contains the presidential building La Moneda.

Figure 1.14: Gatsby curve Metropolitan Region municipalities



This figure plots the relationship between upward mobility and the Gini coefficient at municipality level for the Metropolitan region. We measure the Gini coefficient using the 2017 CASEN survey, considering the "total income before transfer and tax" variant. Those results remain unchanged when we use other income definitions to measure the Gini coefficient.

Figure 1.14 shows the Gastby curve for the Metropolitan region. Comparing with Figure 1.13 can see that the intergenerational mobility and inequality relationship is more steeper for persistence than for upward mobility compared with regions. In particular, upward mobility does not strongly relate with inequality in the Metropolitan region. However persistence does strongly relate with inequality. This relationship is stronger than the regional relationship.

1.6.3 Geographic correlations and mobility across the Metropolitan region

The goal of this section is to take a first step toward understanding what local characteristics can account for the divergence in upward mobility across Chilean municipalities in the Metropolitan region that we documented in Section 5. We do not claim that the correlations we uncover should be interpreted as causal relations, but they certainly serve to

guide future research on the deeper determinants of intergenerational mobility. A similar analysis has been recently performed by Chetty et al. (2014) for the U.S. and by GÃ$\frac{1}{4}$ell et al. (2018) for Italy.

To study the relationship between mobility and municipal socioeconomic characteristics, we start from a large set of correlates based on the literature. We use the Gini coefficient and the share of the top 1 percent as i) measures of inequality. We correlate the proportion of immigrants, monoparental households and the proportion of people of indigenous ethnicity as ii) demographic characteristics. We also include municipal per capita expenditure and per capita square meters of green areas as iii) municipal amenities. We include proportion of students in publicly-funded schools and proportions of students in voucher schools, proportion of people with public health plans, proportion of overcrowded households, and poverty as iv) socio economic characteristics.

Table 1.16: Correlation between mobility measures and socio-economic characteristics

Indicator	β_r	p_{abs_r}	p_{per_r}	p_{11}	p_{55}	p_{15}
β_r	1‘					
p_{abs_m}	-0.1055	1				
p_{per_m}	0.8762	0.3869	1.000			
p_{11}	0.1027	-0.6958	-0.242	1.000		
p_{55}	0.9055	0.2206	0.947	-0.128	1.000	
p_{15}	0.2505	0.8202	0.630	-0.468	0.515	1.000
Gini	0.5033	-0.1066	0.415	-0.045	0.436	-0.022
Share top 1 percent	0.373	-0.0289	0.332	0.018	0.356	0.055
% Immigrants	0.1609	0.3693	0.328	-0.324	0.249	0.361
% People with more than 18 years of schooling	0.7221	0.4134	0.870	-0.142	0.825	0.579
% Monoparental households	-0.2878	-0.4585	-0.489	0.100	-0.423	-0.538
% Public health plan	-0.3299	-0.7967	-0.692	0.554	-0.529	-0.771
Per capita expenditures	0.7209	0.3973	0.861	-0.152	0.819	0.533
Per capita square meters of green areas	0.4769	-0.0206	0.4322	-0.0827	0.3754	0.0761
% Indigenous	-0.6007	-0.3156	-0.71	0.1812	-0.7648	-0.5119
% Students in publicly-funded schools	-0.2817	0.0297	-0.2469	0.0989	-0.1079	-0.1222
% Students in voucher schools	-0.6876	-0.0576	-0.6655	-0.0804	-0.769	-0.240
Poverty	-0.5092	-0.3832	-0.6579	0.215	-0.636	-0.452
Overcrowding	-0.3367	-0.4374	-0.5242	0.2553	-0.493	-0.508

To measure these socioeconomic indicators we use information from the CASEN survey and the “Registro Social de Hogares” dataset.

Table 1.16 sheds lights on the relationship between inequality measures and the indices. The correlations with the Gini coefficient are strong with persistence measures but weak with upward mobility. However, an alternative measure of inequality like the share of the top 1 percent correlates positively with persistence and absolute mobility. This is an intriguing result because the correlation between the share of the top 1 percent and the upward mobility measures is positive. This is different to what Chetty et al.

(2014) found for the USA but it is line with evidence provided by Accarci et al. (2020) for Italy.

The proportion of immigrants is positively correlated with persistence at the top and upward mobility but is negatively related to persistence at the bottom. One of the strongest correlations is the proportion of people with more than 18 years of schooling. Municipalities with more educated people tend to see more mobility, and tend to be more persistent at the top and less so at the bottom. The proportion of monoparental households correlates positively with persistence at the bottom, but negatively with persistence at the top and absolute mobility.

Both municipal per capita expenditure and per capita square meters of green areas correlate positively with persistence at the top, upward mobility, and correlates negatively with persistence at the bottom. The proportion of indigenous populations correlates negatively with persistence at the top and upward mobility but correlates positively (albeit weakly) with persistence at the bottom. Finally, the socioeconomic characteristics correlate positively with upward mobility and persistence at the top but negatively with persistence at the bottom.

We also notice a weak correlation between absolute upward mobility and relative mobility. This is explained by the fact that the variance in relative mobility is higher compared to the variance in absolute upward mobility across municipalities. This means that there is more variance in persistence at the top than upward mobility. For the Metropolitan region, this finding supports the claim that, in terms of intergenerational upward mobility, where to live matters more for children from richer families than for children from middle- and lower-earnings families.

It is worth nothing that the correlation between the proportion of people with public health plan and intergenerational mobility indicators is very high. This means that the type of health that a child can benefit is a main variable that can explain intergenerational mobility in Chile. Thus, one way to improve the imputation-based procedures used to improve intergenerational mobility is by including this variable in the analysis.

1.7 Conclusion

This is the first paper that studies intergenerational mobility in Chile using administrative records. We build a data set that links parental and child earnings using information from the formal labour sector and the place of residence of children during their adolescence. Our analysis reveals that intergenerational mobility at the national level is significantly lower than what was estimated in previous research. However, intergenerational mobility is extremely non-linear. We found that mobility is very high for the bottom 80 percent of the earnings distribution but is very persistent at the upper tail of the parental and child distributions.

In addition, Chile is a highly heterogeneous country in its intergenerational mobility measures at the regional level. For instance, Antofagasta, which is a mining region, has a probability of rags to riches higher than 0.3. This result is in line with what Conolly et al. (2018) founds for the US and Canada. Meanwhile, regions like Araucanía or El Maule have a circle of poverty probability higher than 0.3. It is worth digging a little deeper in future research to understand why those regions are so persistent in poverty.

We also find heterogeneity within the Metropolitan region, with municipalities having a circle of privilege probability higher than 0.7, and other municipalities with a circle of poverty probability closer to 0.3. We also learn that the variance of persistence at the top is higher than the variance of upward mobility. This means that the place of residence affects children of upper-earnings parents more than middle- or poor-class parents. Future research should focus on understanding the causes behind these differences. Although our work is descriptive in nature, it sheds lights on intergenerational mobility in a highly unequal country that does not belong to the advanced economies.

Moreover, we make a some methodological contributions. We use RIF regressions and Kernel conditional densities to study intergenerational mobility at the top. Those tools help usus to show that intergenerational mobility is very persistent at the top in Chile. In addition, we differentiate the Gatsby curve for Chile and Santiago using two measures of intergenerational mobility: absolute intergenerational mobility and relative intergenerational mobility. We show that the Gatsby curve is valid for persistence and

upward mobility for Chile but only for persistence for la Región Metropolitana. This help us to differentiate different mechanisms that may affect intergenerational mobility for Chile.

This work builds on previous national literature and brings the state of research up to the robustness of analysis seen among works in developed economies. As such, not only does it provide more useful information for academics; it also provides an important counterpoint to similar works from developed economies by analyzing intergenerational earnings mobility in a non-developed [o developing] economy in a way that can be contrasted with the results of that literature. We believe that, by providing a clearer picture of how intergenerational earnings mobility occurs in Chile at a regional level, this work can both inspire further research on the matter both in Chile and other developing economies. These results can also help Chilean authorities better understand how and where to apply certain related social/economic programs in order to improve their impact, as well as provide input for drawing up and discussing proposed bills affected by this study's results.

Chapter 2

Income inequality, taxes, and undistributed corporate profits: evidence from Canada

2.1 Introduction

For the majority of individuals measuring income is straightforward this is because their income is equal to their earnings. However, as long as individuals start to obtain income though assets it is more difficult to measure it. In particular, it is known that the line between labour and capital is inherently imprecise for some business owners and some top earners (such as CEOs), and it is certainly possible that tax accounting differs from the common-language way of separating labour from capital (Kopczuk, 2016).

This imprecision is even greater in the case of corporate earnings and wages of small business owners. In some cases, part of the labour income is left inside the firm as retained earnings and accounted as capital income (Smith, Yagan, Zidar and Zwick, 2019). This point is important in inequality measurement; ignoring retained earnings may imply that a significant part of income is not correctly measured, and thus, the resulting inequality statistics do not reflect the true income (or wealth) distribution.[1]

[1]This is also shown by Flores (2018) who demonstrates that household surveys only account for the

In this context, the more recent literature on income (Piketty, Saez and Zucman, 2018; Smith, Yagan, Zidar and Zwick, 2019) inequality aims to measure all the labour and capital income from national accounts. However, can retained earnings be included as part of income?

To answer this question we use a comprehensive definition of income based on the Haig-Simons concept:[2] the flow of resources that can be consumed while leaving the stock of wealth unchanged; or in simple terms, income is the sum of consumption and the change in wealth. This chapter follows this definition for the case of business retained earnings. We should notice however that retained earnings is just a part of capital stock.[3]

This definition is important because there is a relationship between retained earnings and capital gains. More retained earnings mean higher future dividends, implying an increase in the value of the firm which turns into capital gains.[4] By definition, capital gains are changes in wealth. Thus, retained earnings are an indirect source of income.

Additionally, the use of retained earnings to measure capital gains is an attractive method to account for capital gains. In particular, there is an open debate about how one should measure capital gains. Indeed, in the income inequality literature there are two approaches to do this: one considers a realized approach while the other one uses an accrual approach. However, the use of realized capital gains as opposed to accrued capital gains has been criticized for several reasons: i) It often reflects capital income that occurred in different years in which the gains are being measured (Armour et al. 2013; Smeeding and Thompson, 2010). ii) Capital gains obtained by individual investors in a particular period may not be realized in the same year and therefore are omitted when the conventional approach of accounting for only realized capital gains is used.[5]

70% of labour income and 20% of capital income.

[2]For more details on this definition see Haig (1921) and Simons (1938).

[3]This chapter can only provide a partial glimpse of inequality in the Haig-Simons concept of income. In the national accounts, the capital stock includes the plant and equipment of firms, owner- occupied housing, the rental housing stock and publicly owned capital. For example, housing and speculative capital gains are not included in this chapter's inequality measurement.

[4]This argument was developed by Lopez et al. (2016) and Gutierrez et al. (2015). They refer to the capital gains generated in this manner as fundamental accrued capital gains.

[5]As Burkhauser et al. (2014) argue, taxable realized capital gains are not a good proxy for the accounting of yearly accrued capital gains because changes in the tax legislation within countries may affect the definition of taxable capital gains over time; thus, there exists considerable variation of income between years.

Thus, the use of retained earnings allows us to have an accrual measure of capital gains. However, the money inside the firm has a different value from that outside the firm. If an agent wants to get the money out of the company, he or she has to pay personal taxes (depending on the tax system he or she may receive a tax credit for the corporate taxes paid by the firm). The financial market may adjust for those future taxes decreasing the capital gains generated by this retained earning. That is, the tax system should be taking into account to measure the income from capital gains associated with retained earnings. In addition, if we consider that the marginal investor is a foreigner, as Boadway and Bruce (1992) states, the domestic tax system is irrelevant for the marginal investor. Thus, it is crucial to identify what is the right tax rate that should be used to value retained earnings as capital gains. The same is true if there are some transaction costs to get the money out of the firm. Thus, the first contribution of this work is to develop a conceptual framework that analyzes the effect that the ownership of the firm, the tax system and the equilibrium in capital markets have on the capital gain generated via retained earnings.

To compute income inequality and to measure the value of retained earnings as capital gains, one needs data about the ownership of retained earnings. Previous works (Alstæaeter et al. 2017 for Norway, Wolfson et al. 2016 for Canada, Fairfield and Jorrat 2016, for Chile and Austen and Splinter, 2016 for the USA) use administrative tax data, allowing them to attribute directly retained earnings to personal taxpayers. However, obtaining access to these data is quite complicated, and it is difficult to make a comparative study across countries. Moreover, even with the ideal dataset, some imputation (or assumption) might still be needed to have a defined distribution of retained earnings.

In this paper, we do not have access to administrative data. To overcome this limitation, we propose a parametric methodology to impute corporate retained earnings to families using household survey microdata and aggregate national account information. This procedure is based on the exponential-Pareto model established by Dragulescu and Yakovenko (2000) and Silva and Yakovenko (2004).[6] In addition, this method follows the

[6]Others studies that use an exponential distribution for the bottom part of the income distribution are Banerjee et al. (2006) and Jagielski and Kutner (2013).

spirit of Jenkins (2017) and Hundenborn et al. (2018). It uses survey data for a fraction p of the population and aggregate national account data as an additional source for the remaining $1-p$.[7] To evaluate the pertinence of this parametric imputation methodology, we compare it with a non-parametric imputation procedure a capitalization approach similar to that applied by Saez and Zucman (2016).[8] This method contributes to an extensive literature that estimate income and wealth inequality using parametric methods such as Kleiber and Kotz (2003), Chotikapanich, Griffiths and Rao (2007), Clementi and Gallegati (2016), among many others.

The valuation and imputation methodology is tested empirically for Canada, where we impute corporate retained earnings and then use the generated data to compute income inequality measures. To do so, we use the Survey of Consumer Finances (SCF) for 1984 and the Survey of Financial Security(SFS) for 1999, 2005, 2012 and 2016. The justification for using these surveys is that they can be harmonized, allowing us to make a comparison with the capitalization method used by Saez and Zucman (2016) because those surveys have rich information on assets in addition to income.[9] The inequality measures estimated here are not as precise as those estimated by Saez and Veall (2005), Veall (2012) and Wolfson et al. (2016). Despite this, it has the value of correcting a household survey for a form of under-reporting such as Burkhauser et al. (2011), Bourguignon (2018), Blanchet, Flores and Morgan (2018) among many others.

Empirically, we find that the inclusion of corporate retained earnings and its measure as accrued capital gains increases the estimated measure of income inequality and that this also affects the trend in income inequality. Indeed, for 2005, the share of the top 1% increases by 4.5 percentage points (from 7.8% to 12.3%), and the Gini coefficient

[7]The use of survey data is justified because in some countries it is not mandatory for individuals making certain income levels to file a tax declaration; thus, data that are just generated only from tax declarations may have a bias in the lower tail of the income distribution. Also, some developing countries do not have another reliable data source than a household survey, or it is politically difficult to get access to administrative data.

[8]One reason to establish another imputation procedure is that, as Kopczuk (2016) states, "In an environment with a low rate of return, a small bias in the estimated rate of return has large consequences on the estimations of wealth inequality."

[9]Dividends could also be used to impute corporate retained earnings. However, as Alstæaeter et al. (2017) shows, this is not a good procedure to impute retained earnings because (i) retained earnings and dividends move in different directions, (ii) mechanically, the imputations results are not adequate in periods in which the aggregated retained earnings are negative.

increases by 4.4 points (from 47.5 to 51.9) which implies higher income inequality than in 2012 and 2016; this was not the case before accrued capital gains were considered. Those results are robust to the method used to impute corporate retained earnings. In this context, this work contributes to a broader literature that study Canadian income inequality such as Saez and Veall (2005), Fortin et al. (2012), Lemieux and Riddell (2015), Milligan and Smart (2015), Wolfson et al (2016), Green, Riddell and St-Hilaire (2016), among many others.

The remainder of the paper proceeds as follows. In the next section, we develop a theoretical model for valuing retained earnings as income. We discuss how the ownership of the firm, the marginal investor and transaction costs affect the value of retained earnings as income. Section 3 describes the parametric imputation methodology. Section 4 presents the data and estimation of income inequality before and after the parametric imputation procedure. Section 5 evaluates the performance of the parametric imputation procedure by comparing it with a capitalization procedure and studies how reliable the stated assumptions are for building the parametric imputations. Section 6 concludes.

2.2 Measure of personal income derived from retained earnings

This section presents a conceptual and theoretical development of some issues regarding the measure of retained earnings as income. More specifically, if a firm belongs to some individuals, then the associated retained earnings belong indirectly to those agents. This money inside the company means that in the future, there will be dividends that must be paid to those agents, which implies an increase in the value of the firm. Thus, a capital gain is generated, even if not yet realized.

Following Burkhauser et al. (2015), Lopez et al. (2016), Gutierrez et al. (2015), Smeeding and Thompson (2010), Wolfson et al. (2016) and Piketty, Saez and Zuckman (2016), we use a comprehensive definition of income based on the Haig-Simons concept

of income.[10] A capital gain is income generated via a change in wealth. However, the transformation of retained earnings into capital gains depends primarily on the tax system. The intuition comes from the fact that the shareholders of the firm have to pay taxes to take the money out of the firm, thus a potential buyer of this firm is willing to pay the value of retained earnings discounting the taxes that she has to pay to get her income out of the firm. So, even if no tax is paid right away, the capital gains derived from retained earnings would be affected by taxes.

2.2.1 Retained earnings and ownership of the firm

In this section we analyze two scenarios. The first is the case of a publicly traded firm, which has previously been treated in the literature. The second is the case of a privately owned firm. The novelty here is to describe the case of a closely held firm where there exists the possibility of using part of the goods inside the firm for the owner's own consumption. In the case in which an agent is an atomistic owner, it is complicated for her to use retained earnings for her consumption. Moreover, given the law of one price, there should be just one value for a unit of retained earnings for each agent.

Retained earnings in publicly traded firms (corporate sector)

Starting from the Haig-Simons definition of income:

$$y_t \equiv c_t + \Delta V_t, \tag{2.1}$$

where y_t is the total income accrued by an agent in time t, c_t is the total consumption in time t, and $\Delta V_t \equiv V_{t+1} - V_t$ is the change in wealth in period t.

In the context of capital market equilibrium, a change in wealth (ΔV_t) is equivalent to a change in the value of the firm. In order to obtain this value, one can start from the following non-arbitrage condition in the capital markets:

[10] This concept refers to the flow of resources that can be consumed while leaving the stock of wealth unchanged with respect to the previous period. In simple terms, the Haig-Simon definition of income in period t is the consumption plus the change in wealth in period t.

$$rV_t = d_t^0 + (V_{t+1} - V_t)\,, \tag{2.2}$$

where r is the interest rate, V_t is the value of the firm in t and d_t^0 are the pre-tax dividends paid by the firm. This equation reflects the fact that the return from investing the value on the firm in another project (rV_t) should be equal to the return on the money inside the firm ($d_t^0 + (V_{t+1} - V_t)$). Now, following King (1974), one can write this equation after taxes:[11]

$$\frac{1 - \tau_s + \iota\tau_s}{1 - \tau_s}(1 - \tau_e)rV_t = d_t + (V_{t+1} - V_t)\,(1 - \tau_k), \tag{2.3}$$

where d_t is the after-tax dividends paid by the firm, τ_s is the entity tax rate (corporate tax), τ_e is the tax faced by securities holders, τ_k is the capital gains tax on an accrued basis, ι is the percentage of tax integration between the corporate tax and the income tax,[12] r is the interest rate, V_t is the value of the firm in t and d_t are the after-tax dividends paid by the firm. Now with this information, we can state the following proposition.

Proposition 1. *In a publicly traded firm, the capital gains generated by retained earnings (G_f) are given by*

$$G_f(\theta(\tau_e, \tau_s, \iota, \tau_k), \pi^r) = \theta(\tau_e, \tau_s, \iota, \tau_k)\pi^r, \tag{2.4}$$

where $\theta(\tau_e, \tau_s, \iota, \tau_k) = \frac{(1-\tau_s+\iota\tau_s)[1-\tau_e]}{(1-\tau_s)(1-\tau_k)}$ *and* π^r *are the retained profits indirectly owned by the individual.*

Proof. First, we follow Gutierrez et al. (2015) using the equilibrium in capital markets:

[11] This equilibrium condition assumes that the only value that generates retained earnings is the future dividend that could be given today and not the value of the expected future returns that a prospective investment could yield using retained earnings as a source of financing. We do not consider either the problem of the cost of capital or corporate financial policy because given the existence of tax credits and liquidity constraints, a unit of retained earnings could have a different opportunity cost. That is, the only relevant comparison is what an external agent that buys the firm could do today with the money inside the company by taking the money out of it.

[12] An integrated tax system is a system in which the personal tax takes account of corporate tax already paid. In Canada, the system is an imputation system; that is, the idea is to "impute" the gross profits that an individual receives in the form of a dividend, and she has to pay taxes using her personal tax rate; however, for this amount, a tax credit applies to the taxes that the firm previously paid.

$$\frac{1-\tau_s+\iota\tau_s}{1-\tau_s}(1-\tau_e)rV_t = d_t + (V_{t+1}-V_t)(1-\tau_k).$$

Now, using the definition of $d_t = \frac{(1-\tau_s+\iota\tau_s)[1-\tau_e]}{(1-\tau_s)} \cdot (\pi(1-\tau_s) - \pi^r)$ where π is the total profits, and noting that in equilibrium $rV = (1-\tau_s)\pi$, we have that

$$G_f \equiv V_{t+1} - V_t = \frac{(1-\tau_s+\iota\tau_s)[1-\tau_e]}{(1-\tau_s)(1-\tau_k)} \cdot \pi^r = \theta(\tau_e, \tau_s, \iota, \tau_k) \cdot \pi^r.$$

□

In this context, $\theta(\tau_e, \tau_s, \iota, \tau_k)$ is a ratio that represents the opportunity cost of the money left inside the firm in terms of the value outside the firm. This value is decreasing in τ_e, that is taxes that reduce the value of the money outside the firm may negatively affect the value of retained earnings as income. However, $\theta(\tau_e, \tau_s, \iota, \tau_k)$ increases with ι, τ_s or τ_k. This is due to the fact that ι and τ_s are related to a tax credit that reduces the value of the personal tax paid on dividends. The case of τ_k is more complex: having the money inside the firm implies more capital gains; thus, a higher τ_k also increases the tax that the agent pays due to capital gains. Therefore, the opportunity cost of not receiving dividends increases.[13]

$G_f(\theta(\tau_e, \tau_s, \iota, \tau_k), \pi^r)$ is the amount of capital gains derived from retained earnings. Moreover, because of the law of one price, τ_e, τ_s, ι and τ_k should be equal for each individual and equal to the highest values. Otherwise, there will be an arbitrage opportunity.

Retained earnings in closely held firms

Making the distinction between closely held firms and publicly traded companies is important to understand the implications of leaving the money in the company and the consumption possibilities derived from this money. Wolfson et al. (2016) estimate that

[13]Leaving money in the firm as retained earnings comes with some risk that the firm will have future losses and the anticipated future will not fully materialize. Thus if we include this type of risk, the valuation of retained earnings should change. However, θ reflects the opportunity cost of retained earnings related to dividends today, so there should be no risk included.

36% of all retained earnings were in Canadian-controlled private corporations (CCPC) in 2010.[14] The major distinction between closely held firms and publicly traded firms is that in the former the owner has some opportunity to buy goods using retained earnings for her consumption (for example, a chair, a computer, or a car for transportation). However, she can not do this for all the types of goods (for example, she cannot buy a trip to Hawaii for holiday or a luxury car). The distinction between which goods could be consumed inside the firm or outside the firm depends on the tightness of the tax administration. A tighter tax administration will require more documents to allow for some goods to be deductible. Thus, the amount of consumption made within the firm will be lower. For instance, in Canada, to deduct the car consumption it is necessary to justify that each mile is used for corporate purposes.

To illustrate these implications and to understand the equivalence between goods bought inside and outside the firm, consider a firm owner who has the problem of choosing how much to consume between goods inside and outside the firm using only retained earnings.[15] She faces the following maximization problem:

$$\max u(x_1, x_2)$$
$$s.t. \quad p_1\theta(\tau_e, \tau_s, \iota, \tau_k)x_1 + p_2x_2 \leq \theta(\tau_e, \tau_s, \iota, \tau_k)\pi^r,$$

where x_1 are the goods that she buys inside the firm, x_2 are goods that she buys outside the company and $\theta(\tau_e, \tau_s, s, z)\pi_r$ is the value of retained earnings (π^r) outside the firm in terms of forgone dividend. The amount of x_2 is restricted by the tax code. Assuming that a good has the same price inside the firm or outside the firm and that the goods bought inside the firm do not incur value added tax (VAT), we can assume that $p_1 = (1 - \tau_v)p_2$, where τ_v is the VAT tax. Without loss of generality, we can normalize p_2 to 1. Now, if $\theta = 1 - \tau_v$ there are no differences between buying goods inside or outside the firm.

[14] A Canadian-controlled private corporation is a firm whose shares are not publicly traded and that is not controlled by a public corporation or non-residents.

[15] This consumption constitutes a residual consumption that is not related to other income apart from retained earnings.

Define u^* as the optimal level of utility. The object of interest is $C(1,1,u^*)$ which is the cost of getting a level of utility u^* outside the firm. Using duality theory, it is possible to write $C((1-\tau_v)\,\theta(\tau_e,\tau_s,\iota,\tau_k),1,u^*) = \theta(\tau_e,\tau_s,\iota,\tau_k)\pi^r$.

With this in mind, define the Konüs (1924) true cost of living index $P_k(u^*,p_0,p_1)$ as the ratio of the minimum costs of achieving the same utility level u^* when an individual faces two sets of prices p_0 and p_1.

$$\frac{C(1,1,u^*)}{C((1-\tau_v)\,\theta(\tau_e,\tau_s,\iota,\tau_k),1,u^*)} = P_k. \tag{2.5}$$

Then, the value of retained earnings in term of the Haig-Simons definition of income is

$$G_{ch}(\theta(\tau_e,\tau_s,\iota,\tau_k),\tau_v,\pi^r) = P_k \cdot \theta(\tau_e,\tau_s,\iota,\tau_k)\pi^r. \tag{2.6}$$

Using equation (2.1), we define the value of retained earnings as consumption as $G_{ch} - G_f$ which is equal to:

$$(P_k - 1) \cdot \theta(\tau_e,\tau_s,\iota,\tau_k)\pi^r. \tag{2.7}$$

The next lemma summarizes the finding of the previous subsection:

Lemma 1. *For closely held firms, the value of a monetary unit of retained earnings as income is given by θ and for a closely held firm this value is given by $P_k \cdot \theta(\tau_e,\tau_s,\iota,\tau_k)$.*

2.2.2 Retained earnings and the marginal investor

As we demonstrated, the value of retained earnings as income depends on the value of taxes. Given this result, and assuming the law of one price, it is important to understand the value of the tax rate the marginal investor is paying. This issue is even more important for an integrated tax system. In particular, the Canadian tax system is an imputation tax system, with a tax credit used against the personal tax paid on dividend taxation.

There are important implications if the marginal investor is a foreigner or not

subject to domestic taxes. Following Boadway and Bruce (1992), Fuest and Huber (2000) and Edwards and Shevlin (2011), in a small open economy with a stock market that is fully integrated, the required rate of return on domestic shares may be determined by the behavior of foreign rather than domestic investors. In that case, the required return on shares issued by domestic companies may not be affected by changes in domestic personal tax rules. One of the reasons for this is that the dividend tax credit applies against the personal income tax which is typically levied on a residence basis, which implies that the tax integration benefit is irrelevant to determine the return of a particular stock from a corporate firm. However, this view is challenged by Sørensen (2014), who argues that, even in a small open economy, not all shares are traded internationally. Hence, one might expect that tax relief for domestic shareholders will at least reduce the cost of capital for small closely-held companies controlled by one or a few domestic residents. Given this debate in the literature, the question about who is the marginal investor is still open. Thus, we analyze both cases: (i) the marginal investor is a foreign agent and (ii) the marginal investor is a resident.

For the foreigner marginal investor, the arbitrage condition in a small open economy is given by

$$(1-\tau_e^f)rV_t = d_t + (V_{t+1} - V_t)(1-\tau_k), \tag{2.8}$$

where τ_e^f is the foreigner tax rate. This implies that the value of retained earnings will be:

$$\frac{\left[1-\tau_e^f\right]}{(1-\tau_k)}\pi^{re}. \tag{2.9}$$

In the case of a domestic marginal investor, for which the valuation will include the tax credit, the integration has to be included. Assuming an imputation tax system, the arbitrage condition will be

$$\frac{(1-\tau_e^r)}{1-\tau_s}rV_t = d_t + (V_{t+1} - V_t)(1-\tau_k),$$

where τ_e^r is the tax rate on dividends applied to the residents. Using this equation,

the value of retained earnings for a resident marginal investor is

$$\frac{[1-\tau_e^r]}{(1-\tau_s)(1-\tau_k)}\pi^{re}. \tag{2.10}$$

Quantitatively, this could be quite relevant. Following Edward and Shevlin (2011) prior to 2006, $\tau_e^f = 0.15$ but for domestic investors $\tau_e^r = 0.46$ and $\tau_s = 0.32$. Using those numbers, the value of retained earnings inside the firm will be 0.85 for the foreign investor and 0.79 for the national investor.

2.2.3 Retained earnings and transaction costs

Transaction costs are a relevant component of the financial system. The genesis of those costs is that investors require extra compensation to cover the costs of buying and selling a security. These transactions costs tend to be lower for more frequently traded and more liquid stocks. These costs are both theoretically and empirically relevant. Indeed, as Fisher (1994) showed, transaction costs explain part of the equity premium puzzle reducing it from 6.2%(Mehra and Prescott, 1985) to 0.4%. Spreads are not the only cost associated with trading stocks. Equity investors must also pay brokerage commissions. For instance, Jones and Lipson (2001) find that one-way institutional fees on NYSE-listed stocks during 1997 are about 0.12% of the amount transacted. However, as Jones (2002) documented, there are two types of commissions, a proportional commission and a fixed fee.[16]

The existence of a transaction cost implies that getting the money out of the firm is costly, that is, the opportunity cost of having the money in the firm decreases. For instance, if we assume a proportional transaction cost of χ_p, the no-arbitrage condition in the capital markets becomes

[16]Jones (2002) documented that for the USA between 1925 and 2002 proportional commissions ranged from 3% of the total transaction to 0.1% depending on the amount of the transaction, and the fixed commission was from 3 USD to 148 USD. When summed together, transaction costs and commissions represent a substantial and variable friction in trading US equities during the 20th century. The total costs averaged 0.84% over the 1925-2000 period.

$$\frac{1-\tau_s+\iota\tau_s}{1-\tau_s}(1-\tau_e)rV_t = d_t + (V_{t+1}-V_t)\,(1-\tau_k)[1+\chi_p]. \tag{2.11}$$

This implies that the value of a unit of retained earnings inside the firm is given by:

$$\frac{(1-\tau_s+\iota\tau_s)\,[1-\tau_e]}{(1-\tau_s)\,(1-\tau_k)\,(1+\chi_p)}\cdot\pi^r. \tag{2.12}$$

In the case of a fixed transaction cost of χ_f per monetary unit, the non-arbitrage condition is given by:

$$\frac{1-\tau_s+\iota\tau_s}{1-\tau_s}(1-\tau_e)rV_t = d_t + (V_{t+1}-V_t)\,(1-\tau_k) + \chi_f. \tag{2.13}$$

Then, the value of a unit of retained earnings inside the firm is given by

$$\left(\frac{(1-\tau_s+\iota\tau_s)\,[1-\tau_e]}{(1-\tau_s)\,(1-\tau_k)} - \chi_f\right)\cdot\pi^r. \tag{2.14}$$

Combining the two cases, we arrive at

$$\left(\frac{(1-\tau_s+\iota\tau_s)\,[1-\tau_e]}{(1-\tau_s)\,(1-\tau_k)\,(1+\chi_p)} - \chi_f\right)\cdot\pi^r. \tag{2.15}$$

As we can see, an increase of the transaction costs, decreases the value of retained earnings as income.

2.2.4 Summary of the contexts used in the valuation and imputation of retained earnings and drawbacks of the methodology

This section extends the results of Gutierrez et al. 2015 and Lopez et al. 2016. We give an extended theoretical framework to value retained earnings in the following contexts: (i) ownership (corporate vs non-corporate), (ii) marginal investor (domestic vs foreigner) and

(iii) transaction costs (transaction costs vs no transaction costs). For clarity, we present the following table as a summary of this section:

Table 2.1: Value of retained earnings (θ) given different contexts

Type of assumption	Yes	No
Corporate Firm	$\frac{(1-\tau_s+\iota\tau_s)[1-\tau_e]}{(1-\tau_s)(1-\tau_k)}$	$P_k\frac{(1-\tau_s+\iota\tau_s)[1-\tau_e]}{(1-\tau_s)(1-\tau_k)}$
Domestic marginal investor	$\frac{[1-\tau_e^r]}{(1-\tau_s)(1-\tau_k)}$	$\frac{\left[1-\tau_e^f\right]}{(1-\tau_k)}$
Transaction Costs	$\frac{(1-\tau_s+\iota\tau_s)[1-\tau_e]}{(1-\tau_s)(1-\tau_k)(1+\chi_p)} - \chi_f$	$\frac{(1-\tau_s+\iota\tau_s)[1-\tau_e]}{(1-\tau_s)(1-\tau_k)}$

The valuation method described so far has two important drawbacks. First, it does not include speculative capital gains, which are part of income. Ignoring them may generate a bias in the computed inequality measures. Second, this methodology relies on the non-arbitrage condition; if the market allows for arbitrage, then this method gives an imprecise value of the capital gains generated through retained earnings. However, showing that the value of retained earnings as income is different from the value of the retained earnings is a relevant contribution to this literature. In particular, previous works that try to measure income inequality (Wolfson et al. 2016; Fairfield and Jorrat, 2016; and Alstæaeter et al. 2017) ignore completely the effects of the tax system in the value of retained earnings as income. They simply add retained earnings directly to income. This implies that there could be a bias in the estimation of income inequality measures.

2.3 Imputation procedure

2.3.1 Overview of the imputation procedure

This section presents a parametric procedure to impute retained earnings from an aggregate source (national accounts totals) into a microdata source (household survey). This procedure contributes to a broader literature that uses parametric methods to estimate income and wealth inequality such as Kleiber and Kotz (2003), Chotikapanich, Griffiths and Rao (2007), Clementi and Gallegati (2016) among many others.

We start by describing the general procedure and the data sources; then we proceed to describe each stage in more detail. As a starting point, suppose that income including capital gains (h_i) of a family i is given by:

$$h_i \equiv x_i + y_i^{cg}, \tag{2.16}$$

where x_i are income observed in a household survey (without including any capital gains) and y_i^{cg} are income derived from accrued capital gains; this income is generally not observed in microdata. Additionally, from section 2 we know that:

$$y_i^{cg} = \theta \cdot \pi_i^{re}, \tag{2.17}$$

where π_i^{re} are retained earnings accrued to the family i. In addition, we assume that θ is common over the whole population, thus it is irrelevant if we use individual income data or family income data. With this assumption and some other distributional assumptions we can develop a parametric methodology that allows us to have an estimate of π_i^{re} only using x_i and $\sum_i^n \pi_i^{re}$ as inputs.

Assumption 1: Minimum threshold. Corporate undistributed profits π^{re} are a function $\psi : \mathcal{R}^+ \to \mathcal{R}^+$ that takes as input x_i, with

$$\psi(x_i) = \begin{cases} 0 & x_i \leq \bar{w} \\ \psi^*(x_i) & x_i > \bar{w} \end{cases}, \tag{2.18}$$

where $\bar{w}$ is the minimum income required to own corporate retained earnings.[17]

Assumption 2: Ranking preservation. $\psi^*(x_i)$ is a strictly increasing and continuous function of x_i.

Assumption 3: Parametric form. Conditional on $x_i > \bar{w}$, x_i and $h_i = x_i + \psi(x_i)$ are Pareto distributed, with exponents η_x and η_h, respectively. Additionally, conditional on $x_i \leq \bar{w}$, x_i is exponentially distributed.

[17]This is a simplification. In real life there are some people, such as retirees, with low income and some assets.

With this set of assumptions, it is possible to find a closed form for ψ as a function of x_i, $\bar{w}$, η_x and η_h.

Proposition 2. *If Assumptions 1-3 hold, then $\psi^*(x)$ has a close form and is given by:*

$$\psi^*(x) = x^{\frac{\eta_x}{\eta_h}} \frac{\bar{w}}{\bar{w}^{\frac{\eta_x}{\eta_h}}} - x. \tag{2.19}$$

Proof. See appendix C. □

Proposition 2 means that the only plausible way that, conditional on $x > \bar{w}$, the distribution of x and h are Pareto, and, if assumptions 1 and 2 holds (minimum threshold and ranking preservation) the amount of retained earnings owned by each family must be a deterministic function of income without retained earnings (the ψ function or equation (2.19)).

Before describing the estimation procedure, it is essential to discuss the pertinence and intuition of each of the assumptions stated above. Assumption 1 means that it is mandatory to earn a minimum amount of income to be an owner of retained earnings. There are some strong empirical arguments to support the idea that there is a non-trivial proportion of the population that do not hold any capital income.[18]

Assumption 2 means that an agent (or family) that earns more income owns more retained earnings. That is, the richer the household, the higher the amount of retained earnings owns. Empirically, there is a positive relationship between capital and labour income; this correlation is higher for top incomes.

Assumption 3 is a parametric one; this seems slightly arbitrary. However, Pareto type 1 distribution is commonly used in the inequality literature because of its simplicity and precision. Also, by construction, retained earnings increase inequality, this is something that can be justified empirically.[19] Those three assumptions are necessary for getting an appealing closed form for π^{re}. In particular, Proposition 2 states that it is possible to have an approximate measure of retained earnings simply by knowing the income

[18] Also, it makes sense that there could be two thresholds, one for small business and another for large corporations ownerships. In principle, our methodology can be adjusted to this case.

[19] See for instance Wolfson et al. (2016), Fairfield and Jorrat (2016), Lopez et al. (2016) and Alstæaeter et al. (2017). Those works show that including retained earnings increase inequality.

without retained earnings.

Now, we describe the procedure to estimate the parameters of the ψ function, η_h, η_x and $\bar{w}$ by using household survey data and national account data.

i Using household survey microdata, we fit a parametric model to estimate $\bar{w}$ and η_x.

ii Using the estimators generated in the previous step it is possible to compute η_h combining household survey and the national accounts aggregates.

iii With estimates of η_x, $\bar{w}$ and η_h it is possible to estimate $\psi(x_i)$ for each x_i as $\hat{\psi}\left(x_i|\hat{\eta}_x, \hat{\eta}_h, \hat{\bar{w}}\right)$.

iv Now, with iii) it is possible to estimate the total income including retained earnings as $\hat{h}_i = x_i + \theta\hat{\psi}\left(x_i|\hat{\eta}_h, \hat{\eta}_x, \hat{\bar{w}}\right)$

2.3.2 Estimation of η_x and $\bar{w}$

We assume that the true income distribution is a combination of two parametric distributions. When $x_i \leq \bar{w}$, x_i is drawn from an exponential distribution and if $x_i \geq \bar{w}$, then x_i is drawn from a Pareto distribution. From an empirical perspective, it is not hard to justify a Pareto distribution for the upper part of the income distribution, but fitting an exponential distribution for the bottom part is not a perfect distribution for this. However, this distribution is frequently used in inequality research, for instance, Dragulescu and Yakovenko (2000), Silva and Yakovenko (2004), Banerjee et al. (2006), Jagielski and Kutner (2013) among many others. Also, in Appendix A, we develop an economic model to justify those parametric functions. With this in mind, the cumulative distribution function for x is:

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \leq \bar{w} \\ 1 - e^{-\lambda\bar{w}} + e^{-\lambda\bar{w}}\left(1 - \left(\frac{\bar{w}}{x}\right)^{\eta_x}\right) & x > \bar{w} \end{cases}. \tag{2.20}$$

In addition, we define $p = F(\bar{w})$ as the proportion of people (or families) receiving an observed income lower than $\bar{w}$. We assume that those people do not own any retained

earnings (assumption 1). Thus, we impute retained earnings using national account data for those that have a market income higher than $\bar{w}$ (the richest $1-p$ proportion of the population).

Also, notice that the CDF in (2.20) is not differentiable in $\bar{w}$. Thus, the classical theory of extremum estimators fails. To overcome this issue, we use a threshold model to estimate $\Theta = (\bar{w}, \lambda, \eta_x)$. This estimation procedure is similar to those of Coles (2001), Clauset et al. (2009) and Jenkins (2017).[20] To estimate the parameters of the model, we minimize the mean square error (MSE) between the empirical cumulative distribution and the theoretical distribution derived from the model presented in the previous subsection.

$$MSE = \sum_{i=1}^{n} \left(F^{emp}(x_i) - \hat{F}(x_i, \Theta) \right)^2, \tag{2.21}$$

where $F^{emp}(x_i)$ is the empirical CDF defined as:

$$F^{emp}(x_i) = \frac{\sum_{j=1}^{n} 1(x_j \leq x_i)}{n}, \tag{2.22}$$

and $\hat{F}(x_i, \Theta)$ is the estimate of the theoretical distribution from equation (2.20).

Now, to estimate Θ, a finite grid (Gr) is built for $\bar{w}$. For each $w_k \in Gr$, we set $w_k = \bar{w}$. We assume that each $x_i \leq w_k$ are exponentially distributed and each $x_i > w_k$ are Pareto distributed. We use each sub-sample to estimate maximum likelihood estimators for λ and η_x, call them λ^k and η_x^k. Then, we compute the MSE for each $(w_k, \lambda^k, \eta_x^k)$. We choose $(\hat{\bar{w}}, \hat{\lambda}, \hat{\eta}_x)$ as $(w_k, \lambda^k, \eta_x^k)$ that minimizes the MSE.

Knowing $\hat{\bar{w}}$, we define the estimated proportion of families who receive observed income lower than $\bar{w}$ as $\hat{p} = F^{emp}\left(\hat{\bar{w}}\right)$. This is the estimated proportion of people that do not own any retained earnings.

[20] Cowell and Van Kerm (2015) argue that "In practice, however, this threshold is generally determined heuristically, selecting by eye the amount of the upper tail that needs to be replaced by inspecting a Pareto diagram showing the linear relationship between the log of wealth and the log of the inverse cumulative distribution function". In this context, Clauset et al. (2009) argues in favour of a more objective and principled approach based on minimizing the distance between a power-law model and the empirical data.

2.3.3 Estimation of η_h

To estimate η_h it is necessary to combine the total amount of retained earnings Π^{re} (from national accounts) with $\hat{\bar{w}}$ and apply assumption 3. Proposition 3 shows how.

Proposition 3. *Assuming that assumption 3 holds, we know $\bar{w}$ and $E(h|h>\bar{w})$. Then, η_h is uniquely identified.*

Proof. Using assumption 3 we know that if $h_i > \bar{w}$ then h_i distributes Pareto with exponential parameter η_h. Thus, we can use the following relationship derived by Atkinson et al. (2011).

$$\beta_h = \frac{E(h|h>\bar{w})}{\bar{w}} = \frac{\eta_h}{\eta_h - 1}. \tag{2.23}$$

Then, $\eta_h = \frac{\beta_h}{\beta_h - 1}$ □

Now, to estimate η_h, it is sufficient to use the sample analogue estimator for $E(h|h>\bar{w})$ defined by:

$$\hat{E}\left(h|h>\hat{\bar{w}}\right) = \frac{\sum_{x_i>\hat{\bar{w}}} h_i + \Pi^{re}}{\sum_{i=1}^{n} 1(h_i \geq \bar{w})}. \tag{2.24}$$

Then, $\hat{\beta}_h = \frac{\hat{E}\left(h|h>\hat{\bar{w}}\right)}{\hat{\bar{w}}}$ and $\hat{\eta}_h = \frac{\hat{\beta}_h}{\hat{\beta}_h - 1}$.

2.3.4 Estimation of ψ

From i) and ii), we have estimated values for η_x, $\bar{w}$, and η_h. We just need to replace those values in the ψ function. However, the ψ function is derived in a context of continuous random variables and microdata in household surveys is presented as a discrete random variable. To account for this, we need to add a normalization constant c, such that the following restriction holds.

$$\Pi^{re} = \sum_{i}^{n} \hat{\psi}(x_i) \tag{2.25}$$

Then, the estimator for ψ is given by:

$$\hat{\psi}\left(x_i|\hat{\eta}_h,\hat{\eta}_x,\hat{\bar{w}}\right)=\begin{cases}0 & x_i\leq\hat{\bar{w}}\\ c\cdot\left[x_i^{\frac{\hat{\eta}_x}{\hat{\eta}_h}}\frac{\hat{\bar{w}}}{\hat{\bar{w}}^{\frac{\hat{\eta}_x}{\hat{\eta}_h}}}-x_i\right] & x_i>\hat{\bar{w}}\end{cases}. \tag{2.26}$$

with $c=\frac{\Pi^{re}}{\sum_{x_i\geq\hat{\bar{w}}}\left[x_i^{\frac{\hat{\eta}_x}{\hat{\eta}_h}}\frac{\hat{\bar{w}}}{\hat{\bar{w}}^{\frac{\hat{\eta}_x}{\hat{\eta}_h}}}-x_i\right]}$.

With this, we estimate a parametric imputation function for each x_i (income known in a household survey).

2.3.5 Estimation of h_i

The final step is to estimate the total income for a family. We know x_i from the survey data and $\hat{\psi}$ from iii). Also, we need to transform retained earnings into accrued capital gains, we can do this just by multiplying $\hat{\psi}$ by θ (the transformation rate discussed in section 3). With this, we can build an estimation for h_i as:

$$\hat{h}_i=x_i+\theta\cdot\hat{\psi}\left(x_i|\hat{\eta}_h,\hat{\eta}_x,\hat{\bar{w}}\right). \tag{2.27}$$

Now, we can compute inequality measures using parametric imputed income ($\hat{h}_i$) as input. The following section describes the data used in this application and the inequality measures estimated for $\hat{h}_i$.

2.4 Estimation of inequality measures with imputed corporate undistributed profits for Canada

In this section, we show the estimated measures of inequality for Canada using household survey data and aggregate data from national account. The two objectives of this practical application are i) to show the effect of include accrued capital gains (derived from retained earnings) on the measurement of economic inequality and ii) to apply the parametric

imputation method developed in the previous section. The results exposed here should be studied with caution because they are not methodologically comparable with studies that use administrative data to estimate income inequality for Canada.[21] For instance, Saez and Veall (2005), Veall (2012) and Wolfson et al. (2016) do a much more rigorous measure of inequality trends and top incomes measures. Despite this, this empirical estimation shed lights on the importance of correcting a household survey for under-reported capital income and the consequences on the measures of income inequality.[22]

2.4.1 Data and definition used

Data used

We use three data sources. First, we use the Survey of Consumer Finances (SCF) for 1984. In that year, this survey had a supplementary questionnaire that asks wealth related questions at the family level. Second, we use the Survey of Financial Security (SFS) for 1999, 2005, 2012 and 2016. Those surveys measure income and wealth at the family level. Finally, we use annual data of the change in corporate retained earnings from CANSIM table 36-10-0117-01.[23].

[21] Also, there are two other sources of error, the sampling error (from the household survey) and the non-sampling-error (that coming from the imputation process via data combination). This total error should be taken into account when we analyze the results stated in this section.

[22] Burkhauser et al. (2011), Bourguignon (2018) and Blanchet, Flores and Morgan (2018) study the effect of correcting household survey for under-reported income and non-response rates.

[23] Following CANSIM description of corporate savings: "Retained earnings of a corporation or quasi-corporation are equal to the distributable income less the dividends payable or withdrawal of income from the corporation or quasi-corporation respectively. If the foreign direct investment enterprise is wholly owned by a single foreign direct investor (for example, a branch of a foreign enterprise), the whole of the retained earnings is deemed to be remitted to that investor and then reinvested, in which case the saving of the enterprise must be zero. When a foreign direct investor owns only part of the equity of the direct investment enterprise, the amount that is deemed to be remitted to, and reinvested by, the foreign investor is proportional to the share of the equity owned. Retained earnings are equal to the net operating surplus of the enterprise plus all property income earned less all property income payable (before calculating reinvested earnings) plus current transfers receivable less current transfers payable and less the item for the adjustment for the change in pension entitlements. Reinvested earnings accrued from any immediate subsidiaries are included in the property income receivable by the direct investment enterprise." That is, this work has flow of retained earnings for Canadian residents

θ and accrued capital gains before taxes

In section 2, we showed that retained earnings could be transformed into accrued capital gains by multiplying them by θ. This factor reflects the valuation that the financial market makes for a unit of retained earning inside the firm rather than the taxes that the shareholder pays. Those taxes will be paid if and only if the shareholder decides to sell her stocks. For this reason, the accrued capital gains estimated in this work are measured pre-personal taxes. Also, the tax parameters used to obtain θ are from Milligan (2016). We use the highest personal tax rate in Canada for each year. For the sake of simplicity, we only present the result for one case, a domestic marginal investor without transaction costs. However, the results are quite similar using any combination from Table 2.1.

Income Definition.

Most of our interest in inequality from a 'social welfare' point of view is in some definition of post-fiscal income. However, to be consistent among the different income sources, the income used here is pre-tax and transfers. Indeed, accrued capital gains imputed using retained earnings are pre-tax. Thus, the income definition used here is similar to the pre-tax and transfers income used by Piketty and Saez (2003), Saez and Veall (2005) and Burkahuser et al. (2012).

Also, during this application we use the Haig-Simons comprehensive definition of income; that is, we treat income as the sum of consumption and the change in wealth during a defined period of analysis. We can write this as the sum of two components: family market income (x) and accrued capital gains (y^{cg}). Family market income x is defined as the sum of employment income (wages and salaries, net farm income and net income from a non-farm unincorporated business and/or professional practice), investment income, retirement pensions, superannuation and annuities (including those from Registered Retirement Savings Plans [RRSPs] and Registered Retirement Income Funds [RRIFs]) and other money income. It is equivalent to total income minus all government transfer payments. It is also referred to as income before transfers and taxes. Also, market income does not include net capital gains or losses. Thus, the use of retained earnings

as a source of accrued capital gains does not generate double accounting issues.

In addition we use accrued capital gains derived from retained earnings y^{cg} as a measure of family capital gains. That is:

$$y_i^{cg} = \theta \cdot \pi_i^{re},$$

where θ is the transformation rate of retained earnings to accrued capital gains and π_i^{re} is the flow of corporate retained earnings on a given year accrued to family i. However, we do not know the distribution of retained earnings across families. For this reason, we use the estimated retained earnings using the parametric estimator $\hat{\psi}$ defined in section 3. Then, the parametric imputed income ($\hat{h}_i$) is defined as:

$$\hat{h}_i = x_i + \theta \cdot \hat{\psi}(x_i).$$

This income concept along with market income x_i are the objects of study during this section.

Units of Analysis

The units of analysis are the economic families of two or more individuals and unattached individuals. That is, the imputation is made for the family group than the individual. Because we assume that θ is common across individuals, there are no taxation discrepancies of using economic family instead individuals to value retained earnings as income. This assumption is consistent with the fact that we do not know the distribution of family adjusted retained earnings. Thus, we are not able to correct neither y_i^{cg} nor $\hat{h}_i$ for family composition.[24]

[24]See Chanfreau and Burchard (2008) for more details about equivalence scales and family size.

Control for Total Income

The control for total income (H) is the total family market income (X) plus the total accrued capital gains imputed to families Y^{cg}. Moreover, we do not impute all corporate retained earnings to families because, following Bédard-Pagé et al. (2016), approximately 15 percent of the Canadian stock market is controlled by pension plans funds. Thus, we impute 85 percent of total corporate retained earnings (Π^{re}) to families. Table 2.2 is a summary of the information used to construct the control for total income.

Table 2.2: Totals used in the estimation of income inequality

Years	X (2)	Π^{re} (3)	Π^{re} imp. (4)	θ (5)	Y^{cg} (6)	H (7)	(6)/(7)
SCF 1984	248,100	8,061	6,852	0.84	5,756	253,856	2.3 %
SFS 1999	525,300	15,272	12,981	1.14	14,779	540,099	2.7 %
SFS 2005	705,600	94,809	80,588	0.89	71,723	777,323	9.2 %
SFS 2012	959,800	57,892	49,208	0.94	46,256	1,006,056	4.6 %
SFS 2016	1,132,400	10,720	9,112	0.92	8,383	1,140,783	0.7%

Total market income X from SFS and SCF and the annual flow of corporate retained earnings Π^{re} from CANSIM table 36-10-0117-01. Values for X, Π^{re}, Y^{cg} and H are in millions of nominal CAD. Values for θ from Milligan (2016).

We can observe that the share of total corporate retained earnings as the control for total income changes significantly over time. For instance, in 2005 accrued capital gains are 9.2 percent of the control for total income but in 2016 are just 0.7 percent. .

Accounting period for income

The accounting period for market income is income generated from January 1 to December 31 of the corresponding year. This is consistent with the work done by Brzozowski et al. (2010) and Davies, Fortin and Lemieux (2017) which used the SFS for inequality analysis. The same accounting period is used for retained earnings; we use the annual flow of corporate retained earnings starting from January 1 to December 31. The data

used for this purpose is from Statistics Canada table 36-10-0117-01 (formerly CANSIM 380-0078).[25]

2.4.2 Inequality measures including accrued capital gains

Estimates of the parametric model

Table 2.3 shows the estimate of the parameters of the model presented in section 3.

Table 2.3: Parameter estimatets

Year	$\hat{p^*}$	$\hat{\bar{w}}$	$\hat{\eta_x}$	$\hat{\eta_h}$
1984	0.69	32,993	2.7	2.5
	(0.009)	(473.3)	(0.047)	(0.001)
1999	0.78	67,448	2.78	2.54
	(0.009)	(453.5)	(0.027)	(0.001)
2005	0.82	87,518	2.59	1.93
	(0.006)	(641.3)	(0.037)	(0.001)
2012	0.76	92,507	2.29	2.04
	(0.008)	(1,213.8)	(0.046)	(0.001)
2016	0.76	102,542	2.23	2.22
	(0.007)	(1,642.4)	(0.036)	(0.001)

Standard errors in parenthesis. Note: Standard errors were generated using a semi-parametric bootstrap following Cowell and Van Kerm (2015). All the estimation procedures use sample weights.

From Table 2.3 we observe that roughly between 25 percent and 20 percent of the population had some type of corporate savings ($1-\hat{p^*}$). This proportion changes over time; in particular, this ratio decreases between 1984 and 2005, increases by 6 percent

[25]However, Statistics Canada measures market income as the amount earned the year before the survey is asked. For example, for 1984, the income reflected are incomes earned between January 1 and December 31 of 1983. This is a common issue in household surveys. Also, the wealth variables are measured in the middle of the reference year (the year of the survey). Despite this issue, we use values for the actual year for retained earnings. The rationale for this approach is that retained earnings are a consequence of assets bought before the reference year.

between 2005 and 2012 and stays constant between 2012 and 2016. We can also observe that $\hat{\eta_h}$ (the Pareto parameter including accrued capital gains), is lower for years in which retained earnings are more important related to market income.

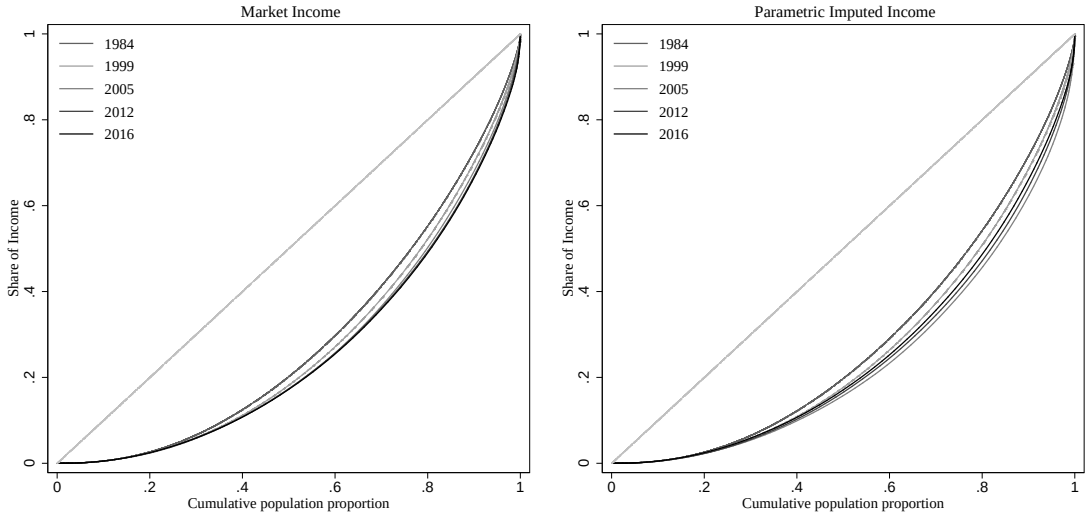
Estimate of the Lorenz curve, Gini coefficient and p-shares with and without parametric imputation

We can estimate retained earnings by replacing the estimated parameters presented in Table 2.3 along with x_i into equation (2.28).[26]

$$\hat{\psi}\left(x_i|\hat{\eta}_h,\hat{\eta}_x,\hat{\tilde{w}}\right)=\begin{cases}0 & x_i\leq\hat{\tilde{w}}\\ c\cdot\left[x_i^{\frac{\hat{\eta}_x}{\hat{\eta}_h}}\frac{\hat{\tilde{w}}}{\hat{\tilde{w}}^{\frac{\hat{\eta}_x}{\hat{\eta}_h}}}-x_i\right] & x_i>\hat{\tilde{w}}\end{cases}. \tag{2.28}$$

Next, we can compare the Lorenz curve using market income with the Lorenz curve estimated using the parametric imputed income $\hat{h}_i$.

Figure 2.1: Lorenz curves for market income and parametric imputed income



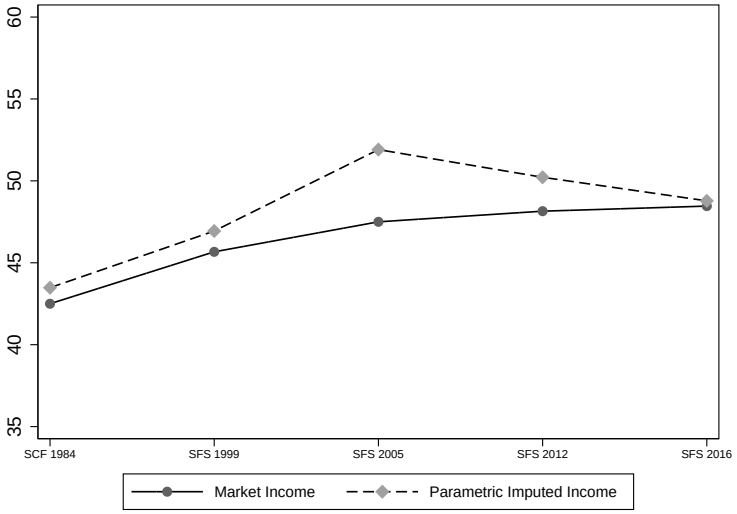
Based on the Survey of Consumer Finances, Survey of Financial Security and CANSIM table 36-10-0117-01.

[26] Recall that $c=\frac{\Pi^{re}}{\sum_{x_i\geq\hat{\tilde{w}}}\left[x_i^{\frac{\hat{\eta}_x}{\hat{\eta}_h}}\frac{\hat{\tilde{w}}}{\hat{\tilde{w}}^{\frac{\hat{\eta}_x}{\hat{\eta}_h}}}-x_i\right]}$.

From Figure 2.1, we observe that after the imputation of accrued capital gains it is easier to differentiate between years; in particular, after the imputation of retained earnings as accrued capital gains, the year that is clearly unequal is 2005 followed by 2012. The latter fact is not valid for market income; it is not easy to observe using Figure 2.1 that the unequal year is 2016. This result is consistent using other inequality measures. Figure 2.2 shows the Gini coefficient using those two income definitions. For 1984 the Gini coefficient after imputing accrued capital gains increases by 2.3 percent relative to the Gini coefficient using market income, which was similar to the increase in 1999 (2.8 percent). In 2005, the Gini coefficient increases by 9.3 percent (from 47.5 to 51.9). However, for the following years, the increase in the Gini coefficient is lower, 4.3 percent in 2012 and 0.7 percent in 2016.[27] Moreover, in Figure 2.2 we observe a change in the inequality trend. There is an increase in inequality between 2005 and 2016 using market income; that is, inequality increases after the Great Recession. However, using the parametric imputed income, inequality decreases between 2005 and 2016. This result is consistent with the wealth inequality trend documented by Davies, Fortin and Lemieux (2017) using the same databases.

[27]The change in the Gini coefficient can be explained using the following formula $G = (1-p) \cdot G_{1-p} \cdot S_{1-p} + p \cdot G_p \cdot S_p + S_p - p$ where p is the proportion of people that is in the top p percent S_x is the share of the x proportion of the population and G_x is the Gini coefficient of the x percentage of the population. Thus, retained earnings affect the Gini coefficient because it changes S_p without changing G_{1-p} thus, despite retained earnings belongs to the top part of the distribution, the Gini coefficient is affected by retained earnings.

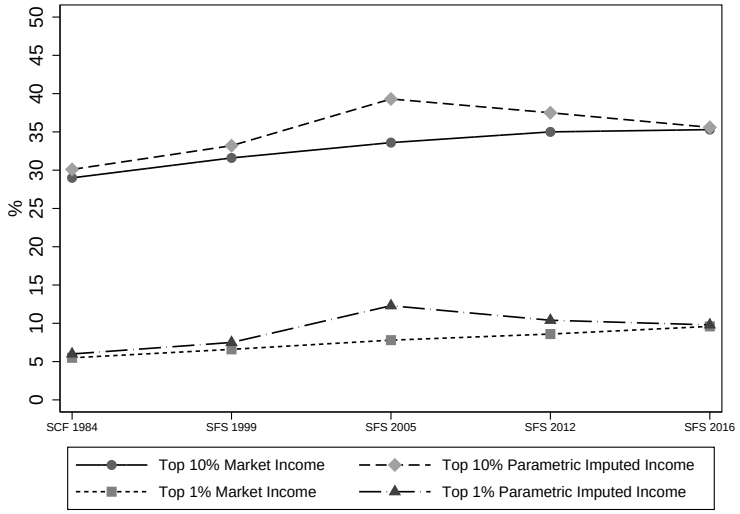
Figure 2.2: Gini coefficient with and without parametric imputed income



Based on the Survey of Consumer Finances, Survey of Financial Security and CANSIM table 36-10-0117-01.

In addition, Figure 2.3 shows the share of the top 10 percent and top 1 percent. Qualitatively, this figure shows the same as that Figure 2.2: we observe an increase of 57.7 percent in 2005 in the share of the top 1 percent after imputing accrued capital gains (from 7.8 percent to 12.3 percent) but only of 2.1 percent for 2016 (from 9.6 percent to 9.8 percent).

Figure 2.3: Top 10% and top 1% with and without parametric income.



Author calculations based on the Survey of Consumer Finances, Survey of Financial Security and CANSIM table 36-10-0117-01.

2.5 Discussion of the methodology and assumptions

The imputation procedure presented so far gives us an appealing parametric procedure to estimate retained earnings. However, there are non-trivial limitations that lead to potential weaknesses of this methodology. First, how well does this parametric approach adjusts to the real income distribution? Second, How valid is the ranking preservation assumption? Third, how precise are the estimations of p presented in Table 2.3? To answer those questions, we contrast our estimated measures of inequality with those generated by a non-parametric procedure: the capitalization imputation method.

2.5.1 Contrasting the parametric estimation using a capitalization approach

One of the limitations of the methodology proposed is to assume a parametric exponential-Pareto model for income distribution. Indeed, the estimates for $\hat{p}$ (incomes where a

Pareto distribution starts) are far away from those used by the literature (Atkinson, 2017; Bourguignon, 2018; Jenkins, 2017). This cast doubts that the distribution of the top incomes is well-described by a Pareto distribution. One way to address this limitation is to use a non-parametric procedure to impute retained earnings. In particular, we use the capitalization imputation approach used by Atkinson and Harrison (1974) and, more recently, by Saez and Zucman (2016).

To use the capitalization method we must know a wealth source to impute an income source. In particular, we know ownership of corporate stocks as a wealth source to impute the flows of retained earnings. Assume that a monetary unit of corporate stocks generates a stream of α_{cap} retained earnings, this value is constant for the whole population and the different types of corporate stocks. Then, the amount of retained earnings imputed to a family i is equal to:

$$\pi_i^{re} = \alpha_{cap} \cdot v_i, \tag{2.29}$$

where π_i^{re} are the retained earnings accrued to the family i, α_{cap} is the capitalization factor that is assumed constant for each i and v_i is the corporate stock owned by the family i.

Summing over i both sides of equation (2.29), we have:

$$\hat{\alpha}_{cap} = \frac{\Pi^{re}}{V}, \tag{2.30}$$

where $\hat{\alpha}_{cap}$ is the estimated capitalization factor, $\Pi^{re} = \sum_{i=1}^{n} \pi_i^r$ and $V = \sum_{i=1}^{n} v_i$. Then, we have,

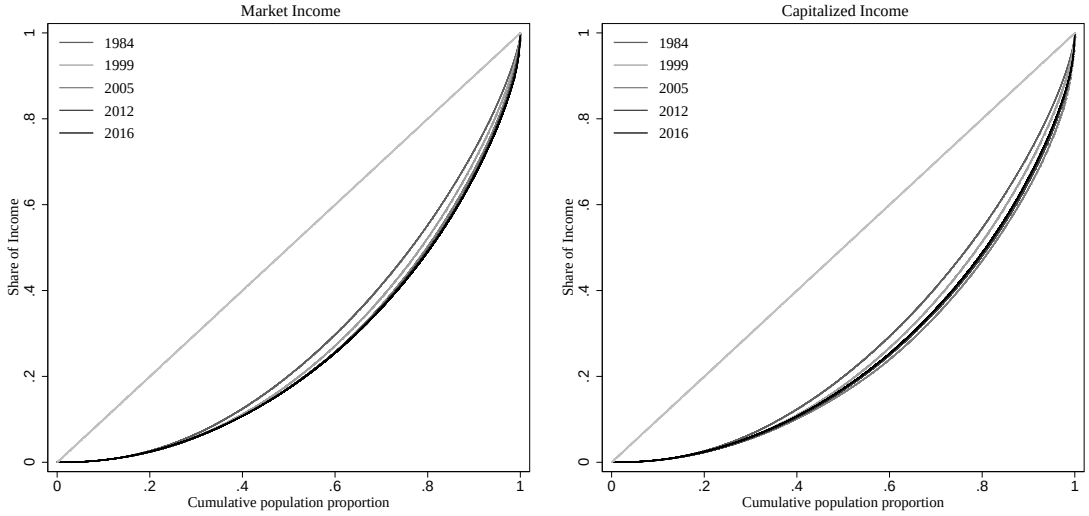
$$\hat{\pi}_i^{cap} = \hat{\alpha}_{cap} \cdot v_i, \tag{2.31}$$

where $\hat{\pi}_i^{cap}$ are the imputed retained earnings using the capitalization method.

For our example, we use national accounts data for total retained earnings and data from the SCF-SFS for corporate stock totals held by families. With this information, we define the capitalized income $\hat{h}_i^{cap}$ as the sum of market income x_i and capitalized accrued capital gains $(\theta \cdot \hat{\pi}_i^{cap})$.

Figure 2.4 shows the Lorenz curves of market income and capitalized income. We observe that the unequal year is 2005 following by 2012. Also, comparing with Figure 2.1, the effect of including capitalized capital gains is not as evident as using the parametric imputation procedure to impute capital gains.

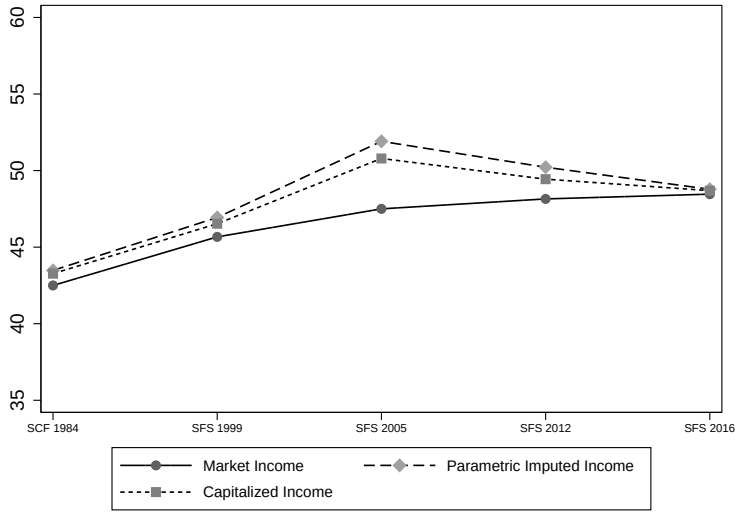
Figure 2.4: Lorenz curves of capitalized income vs market income



Based on the Survey of Consumer Finances, Survey of Financial Security and CANSIM table 36-10-0117-01.

Figure 2.5 shows the Gini coefficient of the capitalized income, parametric imputed income, and market income. We observe that the Gini coefficient computed using the capitalized income confirms that the income inequality trend changes after we include a measure of accrued capital gains. On the other hand, computing inequality using capitalized income shows lower levels of inequality than using parametric imputed income. Despite this, the differences are very small (the Gini coefficient computed using capitalized income is on average 1.1 percent lower than using parametric imputed income).

Figure 2.5: Gini coefficients of capitalized income, parametric imputed income and market income

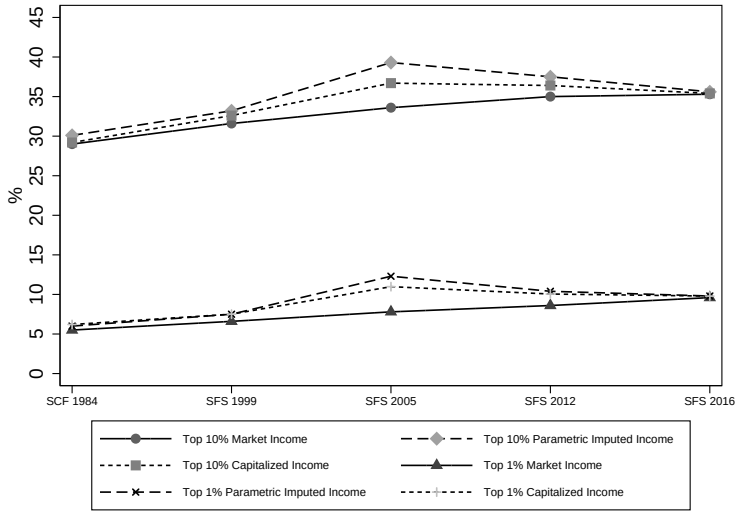


Author calculations based on the Survey of Consumer Finances, Survey of Financial Security and CANSIM table 36-10-0117-01

Figure 2.6 shows the share of the top 10 percent and the top 1 percent using capitalized income, parametric income, and market income. Again, using capitalized income shows the same trend than using parametric imputed income, which means that income inequality is decreasing after the Great Recession.

Moreover, the top income shares estimated using the parametric imputed income are higher than the capitalized income shares (3.3 percent higher for the top 10 percent and 2.2 percent higher for the top 1 percent). This difference is considerable for 2005 where the top 1 percent using the parametric imputed method is 1.3 percentage points higher than the same measure using capitalized income (12.3 vs. 11 percent, i.e., 10.7 percent higher) and 2.9 percentage points higher for the top 10 percent (39.3 vs. 36.7, i.e., 8.4 percent higher). For 1984 and 1999, the share of the top 1 percent using the capitalization method is slightly higher than the same share estimated using parametric imputed income. Thus, estimate inequality measures after impute accrued capital gains using the capitalization method gives similar results than estimating inequality measures using the parametric procedure developed in section 4.

Figure 2.6: Top 10% and top 1% of capitalized income, parametric imputed income and market income.



Based on the Survey of Consumer Finances, Survey of Financial Security and CANSIM table 36-10-0117-01.

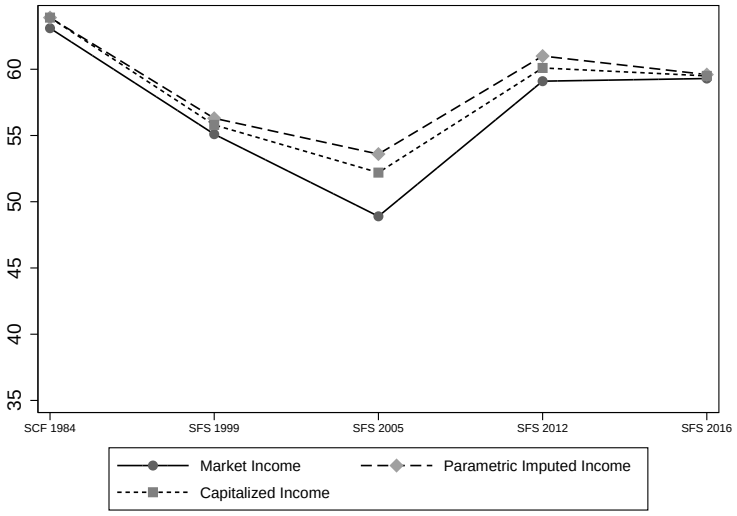
2.5.2 Evaluation of the ranking preservation assumption

Another strong assumption used to build the parametric imputation methodology is that the ranking on the income distribution is the same before and after including accrued capital gains or the "ranking preservation assumption". This assumption is in practice not true; some people are owners of corporate stock shares, but they do not receive any income, and other people receive a very high labour income, but they do not own any assets. Thus, using the ranking preservation assumption could imply a bias in the estimation of inequality measures. Indeed, we can see from Figures 2.5 and 2.6 that the inequality measures computed using the parametric imputed income are slightly higher than using capitalized income. One way to evaluate the pertinence of the ranking preservation assumption is by computing the share on the total income of the $1 - \hat{p}$ percent (those who in theory are the owners of corporate retained earnings). We take $\hat{p}$ from Table 2.3. If the ranking preservation assumption is true, then there should be no differences in the share of the top $1 - \hat{p}$ percent using parametric imputed income or the capitalized

income.

Figure 2.7 shows that the difference between the share of the top $1 - \hat{p}$ percent between the parametric imputed approach and the capitalization approach is small, on average is less than 1 percentage point.

Figure 2.7: Share of the income before $\hat{p}$ and after $\hat{p}$



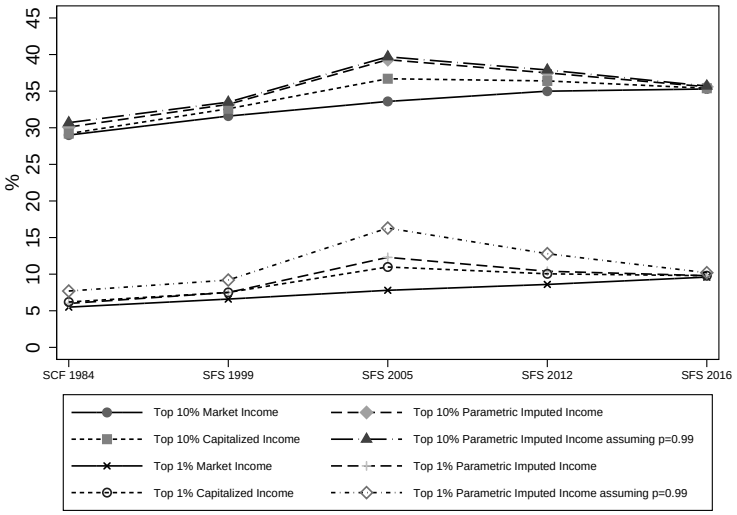
Based on the Survey of Consumer Finances, Survey of Financial Security and CANSIM table 36-10-0117-01. $\hat{p}$ estimates are based on Table 2.3

2.5.3 The effect of changing $\hat{p}$

One of the lessons of Jenkins (2017) is that the threshold above which incomes are Pareto is higher than often assumed. Following Jenkins, we estimate the parametric imputation procedure assuming $(1 - p)$ is 1 percent of the population. To do so, we can fix $\bar{w}$ such a p proportion of the population is driven by an exponential distribution. Figure 2.8 shows the share of the top 10 percent and the share of the top 1 percent after assuming that $p = 0.99$. We observe that the estimates for the top 10 percent are very similar between the baseline parametric imputed income (with $\hat{p}$ from Table 2.3) and the new imputation (with $p = 0.99$). However, the share of the top 1 percent for this new imputation gives higher estimates than the same inequality measure computed using the baseline

parametric imputation. Thus, although our baseline estimates for p are very far from what the literature had found, our inequality measures estimated using the parametric imputed methodology developed here gives results that are closer to what we estimate using capitalized income.[28]

Figure 2.8: Top 10% and top 1% of capitalized income, parametric imputed income and market income (with $\hat{p}$ and $p = 0.99$)



Author calculations based on the Survey of Consumer Finances, Survey of Financial Security and CANSIM table 36-10-0117-01.

2.6 Conclusion

This study argues in favour of using retained earnings as a measure of accrued capital gains, clarifies the difference between left money inside the firm for closely held corporations and for publicly traded firms. In particular, we make a conceptual contribution by showing that for closely held corporations retained earnings could have a consumption value in addition to the change of wealth value that the literature acknowledged. The intuition is that goods can be bought inside the firm to satisfy personal consumption.

In addition, we propose a methodology to impute corporate retained earnings

[28]If we assume $p = 0.9$ or $p = 0.95$ the results are qualitatively the same.

to households. The convenience of this methodology is that it only needs market income information from a household survey and aggregated retained earnings information from a household survey. We apply this methodology to Canada. We show that including accrued capital gains increases measured income inequality and, more importantly, changes the observed trend of inequality which is similar to what Davies, Fortin and Lemieux (2017) found for Canadian wealth.

We compare our imputation method with the broadly used capitalization imputation method. We obtain similar results by using the capitalization method to impute retained earnings. Even though the parametric imputation procedure developed here is not perfect, it gives results that are close enough to the capitalization method. This fact suggests that in the context of scarcity of data, for example, developing countries where we cannot access easily to administrative data, and this procedure could be useful for imputing capital income in such contexts.

Another application of the imputation methodology developed here can be used in the context of undercoverage at the upper tail. Recent literature, such as Flores (2018), show that because of nonresponse and underreporting, household surveys only account for 20 percent of the capital income that appears in national account data. Thus, because the parametric imputation procedure developed here only need a household survey and national account data, could be used to correct for underreporting of capital income in household surveys. Future research is needed to study the imputation procedure in such contexts.

Chapter 3

Gini and undercoverage at the upper tail: a simple approximation

3.1 Introduction

The Gini coefficient is an indispensable index to measure income inequality.[1] Institutions like the World Bank, the Economic Commission for Latin America and the Caribbean, and other important policy institutions routinely use the Gini coefficient to analyze income-distribution changes.[2] Given its popularity, this index should be measured properly; however, data used to compute this coefficient are mainly derived from household surveys, which typically undercover the upper part (top) of the income distribution (Atkinson et al., 2011; Jenkins, 2017; Flores, 2018; Hlasny and Verme, 2018)[3] and, leading to biases in the calculation of the Gini coefficient. As Burkhauser et al. (2017) show, estimating inequality without correcting for such bias may result in an inaccurate analysis of historical

[1]The Gin coefficient is not the only inequality index used. For a discussion about the comparison of two income distributions and inequality indexes see Atkinson (1970).

[2]For a broader discussion about the limitations of the Gini index, see Cowell and Flachaire (2018), Alvaredo et al. (2017), and Osberg (2017). For elegant usages of this index, see Corvalán (2014) and Modalsli (2017), and for an introduction to the Gini coefficient, see Ceriani and Verme (2012).

[3]Undercoverage not only exists at the top of the distribution: Higgins, Lustig, and Vigorito (2018) and Bollinger et al. (2018) show that the entire distribution faces undercoverage, and Ceriani and Verme (2019) show "inequality almost invariably increases by adding observations on the tails of an income distribution but that missing a few observations at the top is much more relevant than missing many observations at the bottom".

inequality changes.

We study effects of two types of undercoverages at the top of the income distribution: underreporting (i.e., missing income) and nonresponse (i.e., missing people). Underreporting occurs when individuals in a population report less income or wealth than they earn (e.g., tax evasion, top coding, information omissions in household surveys).[4] On the other hand, nonresponse occurs when individuals in a population are unrecorded in the data source (i.e., truncated data; e.g., people not submitting their household surveys or not declaring taxes). Bourguignon (2018), Lustig (2018), Blanchet, Flores and Morgan (2018) recently studied these two missing-information types. Their works discuss how adjustments for these biases affect income-inequality measures. In particular, Lustig (2018) develops a taxonomy to differentiate the different types of undercoverage at the upper tail. Bourguignon (2018) shows, in a didactic manner, how different adjustments in the upper tail affect the income distribution. In particular, he argue that the adjustments of the original data relies on three key parameters: i) How much is to be allocated to the top of the distribution; ii) how broad should the top b; iii) what share of the population should be added to the top. Finally, Blanchet, Flores and Morgan develops a novel methodology to find the point where tax data describes better the income distribution than survey data. Their method can be used to correct for underreporting and nonresponse at the top.

In this paper, we depart from Bourginon (2018) and Blanchet, Flores and Morgan, instead of studing the whole income distribution, we only study the effects underreporting and nonresponse in the Gini coefficient. Our first contribution is that we demonstrate that not correcting for underreporting and nonresponse at the top does not necessarily result in an underestimated Gini coefficient.[5]

To correct the Gini coefficient for undercoverage at the top, Atkinson (2007) proposes a simple and pragmatic approximation. He uses household-survey information and tax data, and he approximates the Gini index as $G = G_{1-p}(1 - S_p) + S_p$, where G_{1-p}

[4]Flores (2018) also shows that household surveys do not adequately account for capital income. This problem is also considered underreporting.

[5]Regarding the effect of under coverage on the Gini coeffcient, Ceriani and Verme (2019) shows that adding unit at the bottom or the top not necessary increase the Gini coefficient. Thus, our contribution extends and complements their results.

is the Gini coefficient computed from a household survey representative of a population's poorest $1-p$ percent, and where S_p is the income share owned by the population's top p percent (e.g., the share of the top 1%) and computed from income-tax data (the size of the top). Alvaredo (2011) further develops this procedure and analytically derives and extends his formula, proposing an exact Gini decomposition to be used when a p proportion of the population is not well measured in a data source but is better measured in another source.[6] However, Alvaredo's exact decomposition requires to know: (i) how broad the size of p is and (ii) the income distribution within the top p population. As was discussed by Cowell and Flachaire (2015) and Higgins, Lustig and Vigorito (2018) estimating this p is a major challenge.

Some scenarios lack information on either of these elements (e.g., measuring either income inequality adding undistributed profits or tax-haven wealth[7]). that is, we only know i) how much should be allocated to the top of the distribution. In this context, and without knowing p. A modified version of the Atkinson approximation can be used under such information scarcity and thereby can correct the Gini coefficient.

Thus, this paper's second contribution is that it proposes a simple approximation of the Gini coefficient in the case of underreporting at the top which is a slightly modified version of the traditional Atkinson approximation without the necessity of knowing the size of the top p. In addition, the approximation's analytical bias is computed. In addition, we show that the bias is higher when the traditional Atkinson approximation is used for solving nonresponse and underreporting instead of the new adjusted formula to correct underreporting. It shows, numerically, that the proposed approximation is near exact when used to correct the Gini coefficient for underreporting but may be heavily upward biased for correcting the Gini coefficient for nonresponse. That is, in order to use the underreporting methodology we need to first correct for missing people at the top.

Thus, this paper's third contribution is to propose and apply a methodology for estimating the missing proportion at the upper tail. Thus, we can estimate an un-

[6]Diaz-Bazan (2015) present another exact interpolation which is more exact than the Atkinson approximation, but requires additional data.

[7]Issues that are tremendously relevant for inequality measurement, see for instance, Alstadsæter et al. (2017, 2018).

derreporting and nonresponse corrected Gini coefficient by estimating the proportion of nonrespondants and then apply the underreporting correction. It applies this methodology to two countries: Chile and Canada, and corrects the income Gini coefficient by adding undistributed business profits, a source of capital income underreported in household surveys, and administrative-tax-declaration data. Indeed, as Smith, Yagan, Zidar and Zwick (2019) argue "a primary source of top income is private "pass-through" business profit, which can include entrepreneurial labour income for tax reasons", thus some part of labour income is transformed into capital income and left inside the firm. Indeed, some tax reforms induces to keep business income inside the firm whereas others generate incentives to take out profits as dividends (see for instance the 2003 US dividend tax reform). Thus, not accounting for undistributed business income when we measure income inequality could lead to artificial changes that bias levels and trends of income inequality estimates. Thus, the methodology developed here could be used to estimate level and trends of income inequality that are robust to tax changes and tax avoidance behaviour.

The next section discusses the Gini coefficient in an undercoverage context. Section 3 proposes a Gini approximation when used for solving undercoverage at the top. Section 4 proposes a methodology to correct the Atkinson non nonresponse approximation. Section 5 presents empirical aplications to test the methodology developed here, and Section 6 concludes the paper.

3.2 Gini coefficient and undercoverage at the top

This chapter studies the Gini coefficient in the case of underreporting and non-response at the top.[8] The underreporting and non-response problems studied are discrete and only affect the upper tail of the income distribution. The intuition to do such analysis is that, in some cases, there are some tax avoidance technologies that are costly and only available for individuals at the top of the income distribution. Thus, only those

[8]There are other biases that may affect measurement of top income inequality. For instance small sample bias, that is, incomes at the very top span very large dollar ranges - the difference between the top and bottom of the top 0.01% is measured in multiple millions. But only a few hundred dollars separate people whose incomes are near the median - e.g. within the 51st percentile. When a random sample is drawn, sampling variability can matter at the top end even if it is inconsequential for middle incomes.

who can afford this technology can underreport their income or, in the most extreme cases, disappear completely from the income distribution. Another reason to assume a discrete undercoverage is that in general income taxes are progressives with an important proportion of the population that are exempt from income taxes, and those that are exempt from paying taxes do not have any incentives to under declare their income to the tax authority. Finally, and this is relevant for capital income, in order that people can start saving money (and receiving benefits from capital income), one needs first to get enough to consume the basics or to get certain income such that liquidity constraints are not binding. After that people can satisfy their basics needs, people can start buying assets to receive capital income. Thus, only after certain income level, people have some capital income to underreport.

3.2.1 Underreporting at the top

Suppose the top p proportion of a population underreports their income, due to survey-measurement error, tax evasion, or top coding. Let y_i be the true income of each individual or household unit i, and y_i^* the observed income for this unit i. Thus,

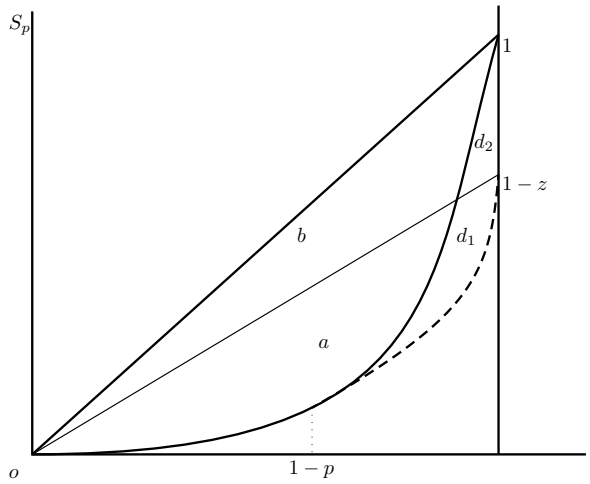
$$y_i = y_i^* + z_i, \tag{3.1}$$

with z_i being the amount underreported for each unit i and $z_i \geq 0 \ \forall i$. In addition, $z \equiv \frac{\sum_{i=1}^{n} z_i}{\sum_{i=1}^{n} y_i}$ is the proportion of the total underreported income over the total true income. Figure 3.1 represents this type of under-reporting graphically. The dashed Lorenz curve describes the observed information y_i^*, and the bold Lorenz curve represents the distribution of true income, y_i. Also, $\phi \equiv \frac{\sum_{i=n\cdot p+1}^{n} y_i}{\sum_{i=1}^{n} y_i}$ is the share of the total income held by the top p proportion of the population.

Note that the distribution for people whose income is lower than the $1-p$ percentile is the same with and without underreporting. In this context, the true income distribution is not necessarily more unequal than the observed one. Proposition 1 established the circumstances under which this type of outcome occurs.

Proposition 4. *Given nonnegative underreporting at the top, the Gini coefficient (mea-*

Figure 3.1: Relative Lorenz curves with underreported income at the top



The dashed relative Lorenz curve describes the observed income distribution and the bold Lorenz curve describes the true income distribution, a, b, d_1 and d_2 are areas; and $1-z$ is the proportion of the total observed income divided by the total true income

sured using the observed income distribution) is not necessarily lower than the Gini coefficient measured using the true income distribution. The true Gini coefficient is lower than the observed Gini coefficient when underreporting sufficiently reduces inequality inside the underreported group.

Proof. Suppose that a fraction z of income is underreported, that S_{1-p} is the income share of the group reporting their income accurately, and, S_p is the income share of the group underreporting their income. Total reported income is the denominator for both shares. $\bar{S}_{1-p}$ and $\bar{S}_p$ are the income shares of the same groups when underreported income is included. Then, the following relationships hold:

$$\bar{S}_{1-p} = S_{1-p}\,(1-z)$$

$$\bar{S}_p = S_p\,(1-z) + z$$

From Alvaredo (2011), the Gini coefficient G^* (uncorrected for underreporting) is

$$G^* = (1-p)\,S_{1-p}G_{1-p} + pS_pG_p + S_p - p \tag{3.2}$$

where G_{1-p} is the Gini coefficient of those who report accurately; G_p is the Gini coefficient of reported income for underreporters (i.e., this income is lower than their true income); and p is the fraction of underreporters. The Gini coefficient for true income G is

$$G = (1-p)\, S_{1-p}\, (1-z)\, G_{1-p} + p\, (S_p\, (1-z) + z)\, \bar{G}_p + S_p\, (1-z) + z - p \tag{3.3}$$

with $\bar{G}_p$ as the income underreporters' Gini coefficient, including their underreported income. Thus,

$$G^* - G = zS_{1-p}\, (1-p)\, G_{1-p} - p\left[S_p\left(\bar{G}_p - G_p\right) + \bar{G}_p S_{1-p} z\right] - zS_{1-p}. \tag{3.4}$$

And, $G^* - G > 0$ if:

$$G_{1-p} > \frac{p\left[S_p\left(\bar{G}_p - G_p\right) + \bar{G}_p S_{1-p} z\right] + zS_{1-p}}{zS_{1-p}\,(1-p)}. \tag{3.5}$$

Thus, G_{1-p} can only be less than 1 only if $\left(\bar{G}_p - G_p\right) < 0$. □

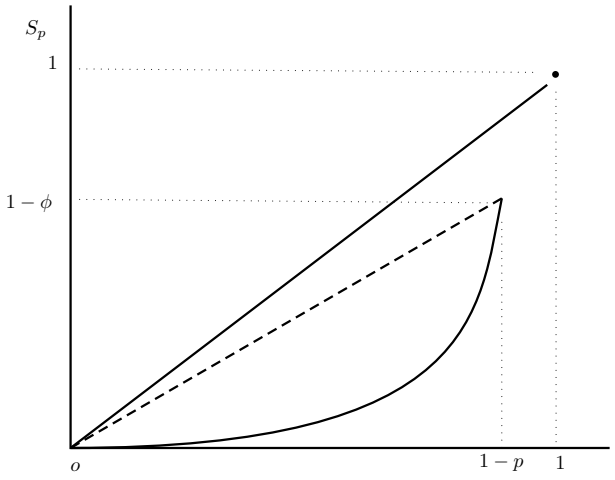
As an example, assume the observed income distribution is $(1, 20, 50)$ and the true income distribution is $(1, 120, 150)$. In the first distribution, the Gini coefficient is 0.460, and in the second distribution, it is 0.367; however, if underreporting increases inequality between the underreporters, then the Gini coefficient of the true income distribution will be higher than the observed distribution Gini coefficient.

3.2.2 Nonresponse at the top

Suppose that the richest p proportion of the population does not report any information (e.g., if p equals 0.01, the income information for the top 1% is truncated), that this nonreported population holds a proportion ϕ of the total income, and that the true income y_i for the other $1-p$ fraction of the population is observed. Figure 2 characterizes

this situation using Lorenz curves.

Figure 3.2: Lorenz curve in a context of nonresponse at the top



.

p is the proportion missing, ϕ is the proportion of the total income held by missing population, the bold black line is the equality line for the entire population, and the dashed black line is the equality line for the population observed

Since the shape of the missing Lorenz curve segment from $1-p$ to 1 is unknown, we cannot say anything a priori about the Gini of the reporters versus the true distribution, as described in proposition 2, not necessarily the true income distribution is more unequal than than the observed income distribution.

Proposition 5. *Given nonresponse at the top, the Gini coefficient (measured using observed information) can be greater than the Gini coefficient of the true-income distribution (including nonrespondents' information).*

Proof. The true Gini coefficient can be written as

$$G = (1-p)\, S_{1-p} G_{1-p} + p S_p G_p + S_p - p, \tag{3.6}$$

where G_{1-p} is the observed Gini coefficient, $1-p$ is the proportion of people observed, S_{1-p} is the share of the total income (observed and unobserved) received by the observed people, and G_p is the Gini coefficient of the unobserved people. Thus,

$$G_{1-p} - G = pG_{1-p}S_{1-p} - pS_pG_p - S_p + p. \tag{3.7}$$

And, $G_{1-p} - G > 0$ if:

$$G_{1-p} > \frac{S_p\left(G_p p + 1\right) - p}{pS_{1-p}} \tag{3.8}$$

□

For example, if the truncated distribution is $(0, 1)$ and the true distribution is $(0, 1, 1)$, the Gini coefficient for the truncated is 0.5, while for the observed is 0.33. However, if the observed Gini coefficient is low enough, the Gini coefficient, computed after including nonrespondents, will be higher than the observed Gini coefficient.[9] Ceriani and Verme (2019) prove a similar version of Proposition 2: they show that including additional people in the upper part of the income distribution, each of them earning the maximum observed income does not necessarily increase the Gini coefficient.

3.2.3 Nonnegative underreporting and nonresponse: the joint case

In real-world applications underreporting and nonresponse at the upper tail are often both presented. Figure 3.3 uses Lorenz curves to describe this situation.

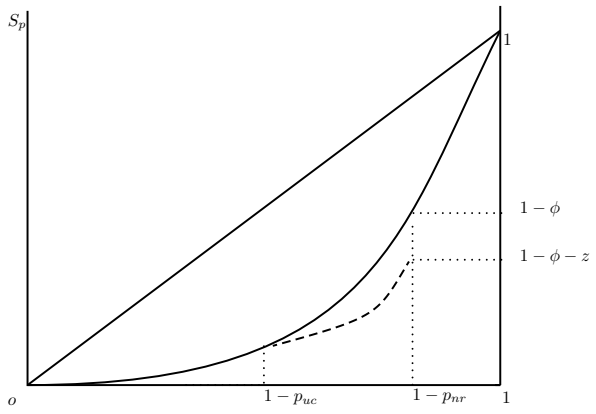
Proposition 6. *Given underreporting and nonresponse at the top, the Gini coefficient (measured using the observed-income distribution) can be greater than the Gini coefficient of true-income distribution.*

Proof. Follows directly from propositions 1 and 2. □

From Propositions 1, 2, and 3, it can be concluded that the existence of undercoverage at the top does not imply an over or underestimated Gini coefficient if this

[9]Bollinger et al. (2018) present evidence that the income distribution, including nonrespondents, is more unequal than the observed distribution.

Figure 3.3: Relative Lorenz curves in a context of underreported income and nonresponse at the top



The dashed relative Lorenz curve describes observed-income distribution; the bold Lorenz curve describes true income; $1-\phi-z$ is the proportion of total income observed in a dataset; z is the proportion of total income underreported in that data set; ϕ is the propotion of income belonging to nonrespondants; $1-p_{uc}$ is the proportion of people reporting their total income in the dataset; $p_{uc}-p_{nr}$ is the proportion of people who appear in the dataset but underreport their income; and p_{nr} is the proportion of people not appearing in the dataset.

undercoverage is uncorrected.

3.3 A Gini approximation for undercoverage at the top

3.3.1 An approximation as a solution for underreporting at the top

One way to correct for underreporting is to use the Atkinson approximation (Atkinson, 2007; Alvaredo, 2011); G^{atk}, defined as:

$$G^{atk} = G_{1-p} \cdot (1 - S_p) + S_p, \tag{3.9}$$

where, G_{1-p} is the Gini coefficient for the bottom $1-p$ proportion of the population, and S_p is the share of the total income held by the top p proportion. Alvaredo (2011) shows that if p is infinitesimal, then the true Gini coefficient G is identical to the Atkinson

approximation.

Equation (3.9) can be modified slightly to estimate the Gini coefficient when underreporting is presented using the observed (and underreported) income, y_i^*, and the proportion of total income that is underreported, z. We simply replace G_{1-p} with G^*, the Gini coefficient of y_i^*, and S_p with z. Denote this modified Atkinson approximation by G_{ur} :

$$G_{ur} = G^* (1 - z) + z. \tag{3.10}$$

Equation (3.10) can be shown to be an upper bound of the true Gini coefficient G.

Proposition 7. *G_{ur} is an upper bound of the true Gini coefficient G.*

Proof. Using Figure 3.1,

$$\frac{a + b + d_1 + d_2}{\frac{1}{2}} = G + 2 \cdot (d_1 + d_2), \tag{3.11}$$

where G is the true Gini coefficient. Notice that

$$\frac{(a + d_1)}{\frac{1}{2}} = \frac{(a + d_1)}{\frac{(1-z)}{2}} \cdot (1 - z) = G^* \cdot (1 - z), \tag{3.12}$$

where G^* is the Gini coefficient of the observed data. In addition,

$$\frac{b + d_2}{\frac{1}{2}} = z \tag{3.13}$$

Thus,

$$G_{ur} - G = 2 \cdot (d_1 + d_2).$$

□

In Figure 3.1, the bias generated by (3.10) is equal to $2 (d_1 + d_2)$. This bias stems from the Atkinson approximation's assumption that one individual owns all the underreported income. Proposition 5 computes this bias analytically.

Proposition 8. *$Bias^{ur} = G_{ur} - G$ is given by:*

$$p\left[z\cdot(\,1-G_p\,)-\theta(z)\left(S_p\left(1-z\right)+z\right)\right], \tag{3.14}$$

where p is the proportion of people underreporting their income; z is the proportion of total income that is underreported; G_p is the Gini coefficient of the underreporters' observed income; $\theta(z)$ is the difference between the true Gini coefficient of underreporters $\bar{G}_p$, and the observed Gini coefficient of the underreporters G_p; and S_p is the observed-income share of the p proportion that underreports their income.

Proof. From Proposition 1, the true Gini coefficient can be written as

$$G=\left(1-p\right)S_{1-p}\left(1-z\right)G_{1-p}+p\left(S_p\left(1-z\right)+z\right)\bar{G}_p+S_p\left(1-z\right)+z-p, \tag{3.15}$$

where S_{1-p} and S_p are the observed income shares of the people who, respectively, do and do not underreport their income. Those shares use total reported income as the denominator. Also, $\bar{G}_p=G_p+\theta(z)$; that is, the true Gini coefficient of the underreporters $\bar{G}_p$ is the Gini coefficient computed using those underreporters' reported income G_p plus the change in the underreporters' Gini coefficient after including their underreported income $\theta(z)$. Thus,

$$G=\left(1-z\right)\left[\left(1-p\right)S_{1-p}G_{1-p}+pS_pG_p+S_p-p\right]+p\theta(z)\left(S_{1-p}\left(1-z\right)+z\right)+pG_{1-p}z+z-zp. \tag{3.16}$$

Given that the observed Gini coefficient is equal to

$$G^*=\left[\left(1-p\right)S_{1-p}G_{1-p}+pS_pG_p+S_p-p\right]$$

and that $G_{ur}=G^*\left(1-z\right)+z$,

$$G=G_{ur}+p\left[z\left(G_p-1\right)+\theta(z)\left(S_p\left(1-z\right)+z\right)\right], \tag{3.17}$$

and,

$$G_{ur}-G=p\left[z(\,1-G_p\,)-\theta(z)\left(S_p\left(1-z\right)+z\right)\right]. \tag{3.18}$$

□

In Proposition 5, the underreporting approximation bias increases with p but decreases with $\theta(z)$ and S_p; however, an increase in z does not necessarily increase the bias.

3.3.2 A Gini approximation as a solution for nonresponse at the top

The Atkinson approximation can be directly used to correct the Gini coefficient for nonresponse. Assuming the nonrespondents' income share is ϕ, the Atkinson approximation for the full population's Gini coefficient in the case of nonresponse at the top is:

$$G_{nr} = G_{1-p}\,(1-\phi) + \phi, \tag{3.19}$$

where, G_{1-p} is the observed population's Gini coefficient. This is an upper bound of the Gini coefficient.

Proposition 9. *Assuming nonrespose at the top, G_{nr} is greater than the true Gini coefficient, and the bias of this approximation is given by:*

$$Bias^{nr} \equiv G_{nr} - G = p\left[G_{1-p}\,(1-\phi) + 1 - \phi G_p\right], \tag{3.20}$$

where p is the proportion of people underreporting their income; ϕ is the proportion of total income that is nonreported; G_{1-p} is the Gini coefficient of those with known income; and G_p is the Gini coefficient of the nonrespondents.

Proof. Using the previously applied Gini decomposition, the true Gini coefficient is

$$G = (1-p)\,S_{1-p}G_{1-p} + pS_pG_p + S_p - p, \tag{3.21}$$

and the Atkinson approximation for solving nonresponse is

$$G_{nr} = G_{1-p}\,(1 - S_p) + S_p. \tag{3.22}$$

Thus,

$$G_{nr} - G = p\,[G_{1-p}\,(1 - S_p) + 1 - G_p S_p]\,.$$

Recall $S_p = \phi$. Also, because $\phi < 1$ this is always greater than 0. Thus, $G_{nr} > G$. □

An increase in p increases the nonresponse bias, but an increase in the share of nonrespondents reduces this bias. Greater inequality among nonrespondents also reduces the bias.

3.3.3 A Gini approximation for underreporting and nonresponse at the top

In Figure 3.3, underreporting and nonresponse are both presented. Using the Gini approximation proposed to correct underreporting without correcting for nonresponse, or correcting only for nonresponse without correcting for underreporting, could lead to Gini coefficients lower or higher than the true one.

Proposition 10. *When underreporting and nonresponse both exist in the top of the income distribution, estimating G_{ur} to correct only for underreporting without correcting nonresponse does not necessarily establish an upper bound of the true Gini coefficient.*

Proof. Follows directly from proposition 2 and proposition 4. □

Proposition 11. *When underreporting and nonresponse jointly exist in the top of the income distribution, use of the Atkinson approximation to correct only for nonresponse without correcting for underreporting does not necessarily establish an upper bound of the true Gini coefficient.*

Proof. Follows directly from proposition 1 and proposition 6. □

Another way to correct for both underreporting and nonresponse is to use the Atkinson approximation to consider underreporting and nonresponse both as a form of nonresponse. This procedure generates a Gini coefficient that is an upper bound of the true Gini coefficient.

Proposition 12. *When underreporting and nonresponse jointly exist in the top of the income distribution, use of the nonresponse Atkinson approximation to correct for non-response and underreporting generates an upper bound of the Gini coeffient. The bias is given by*

$$Bias^{nr} + (1-\phi)Bias^{ur}_{1-p_{nr}},$$

where $Bias^{nr}$ *is the bias for non-response Atkinson approximation and* $Bias^{ur}_{1-p_{nr}}$ *is the Gini approximation for correcting underreporting in the proportion of the population that underreports,* p_u*.*

Proof.

$$(1-z-\phi)G_{p_u} + \phi + z$$

$$(1-\phi)\left(\frac{(1-\phi-z)}{1-\phi}G^*_{p_u} + \frac{z}{1-\phi}\right) + \phi$$

$$(1-\phi)\left(G_{p_u} + Bias^{ur}_{1-p_{nr}}\right) + \phi$$

$$G + Bias^{nr} + (1-\phi)\,Bias^{ur}_{1-p_{nr}}$$

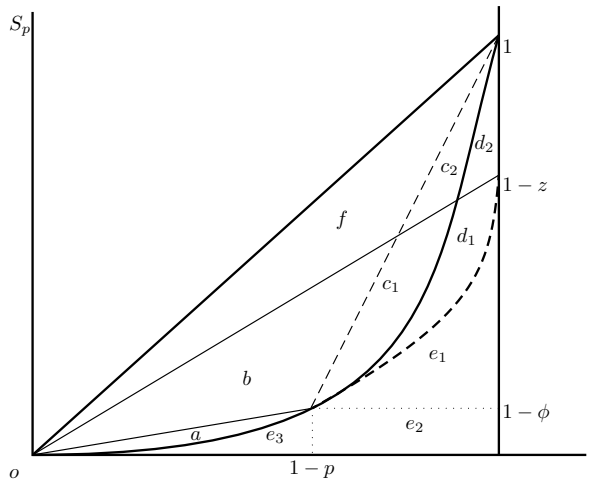
□

3.3.4 The underreporting vs the nonresponse approximation

As established in propositions 4, 6 and 9, the approximations for correcting underreporting G_{ur} and/or nonresponse G_{nr} are upper bounds of the Gini coefficient. There are two ways to correct for underreporting, one is to correct using the whole population even if that population is not well measured and the other is to eliminate the proportion that is not well measured and correct the gini coefficient using the straight Atkinson approximation. Assuming 1) ϕ and z are known, 2) the true income distribution until the

$1-p$ percentile is known, and 3) an income level y_i^* lower than the true income is known for the top p proportion of the population, then either G_{nr} or G_{ur} can be used to correct the Gini coefficient for undercoverage. G_{nr} can be estimated using the Gini coefficient of those accurately reporting their income and ϕ. G_{ur} is estimated by computing the Gini coefficient of the observed distribution and z.[10] Proposition 10 establishes that the nonresponse approach G_{nr} (i.e., dropping underreporters' information) generates a bigger bias than using the underreporting approach G_{ur}. Figure 3.4 illustrates these results.

Figure 3.4: Relative Lorenz curves of both underreporting and nonresponse problems



a, b, c_1, c_2, d_1, d_2, e_1, e_2, e_3 and f are areas; $1-p$ is the proportion of the population that is well measured, $1-\phi$ is the proportion of the total income held by the first $1-p$ percentiles; $1-z$ is the proportion of the total income observed; the bold line is the equality line for the whole population; the black line is the equality line for the reported population; and the dashed line is the equality line for the p richest percentiles

Proposition 13. *G_{nr} is higher than G_{ur}.*

Proof. From Figure 3.4, given $G_p \equiv \frac{a}{\frac{(1-\phi)\cdot p}{2}}$. Thus,

$$\frac{a+b+c_1+c_2+d_1+d_2+e_1+f}{\frac{1}{2}} = G_p\left(p\cdot(1-\phi)\right)+\phi-(1-p)+\phi\left(1-p\right) \leq G_p\cdot(1-\phi)+\phi. \tag{3.23}$$

[10] ϕ is greater than z because ϕ also includes the proportion of underreporters captured by the household survey. One way to know ϕ and z is by having two harmonized income-distribution information sources: a household survey and income-tax data. Harmonized sources are essential, but in practice, such harmonized sources may not exist —e.g., Atkinson, Piketty, and Saez (2011) used nonharmonized data, but Jenkins (2017), Burkhauser et al. (2018), and Piketty, Saez, and Zucman (2018) did.

And,

$$G_{ur} + 2 \cdot e_1 = G + 2 \cdot (d_1 + d_2 + e_1) \leq G_{nr}. \tag{3.24}$$

□

For nonresponse, the bias is greater than $2\,(d_1 + d_2 + e_1)$. This bias is higher than the underreporting case $2\,(d_1 + d_2)$ because the proportion of income assumed to be held by one individual is larger when using the Atkinson approximation to correct for nonresponse instead of the approximation proposed for underreporting (ϕ is greater than z). This shows that an imperfect measure of the income held by the top part of the distribution is preferable to not having any measure of that same income. For instance, it is better to have the upper part with top coding, even if the top coded value is far from the real income. Thus, before applying the corrected underreporting approximation, we need to correct for nonresponse first.

3.3.5 Montecarlo simulation

To evaluate the magnitude of the benefits from using the underreporting approximation over the nonresponse approximation we performed Montecarlo simulations. In these simulations, we assumed that (i) the bottom $1 - p$ of the population is drawn from an exponential distribution;[11] that (ii) the true income for the richest p of the population is drawn from a Pareto type I distribution (with α_r as the Pareto parameter and $\bar{w}$ as the threshold income); that (iii)the observed income for the top p proportion is drawn from a Pareto type I distribution with α_s parameter and $\bar{w}$ income threshold with $\alpha_s > \alpha_r$ (i.e., the income distribution for the observed top tail is less unequal than the real income distribution).[12] In addition, this modified Atkinson approximation can be implemented in a top-coded environment (assuming top coding at $\bar{w}$). We denote to this estimated

[11]Results are presented via a single parametric distribution; however, the bottom part of the distribution could also be approximated with Singh-Maddala, Dagum, or GB2 distributions, which produce similar results.

[12]For simplicity, results are presented using a Type I Pareto distribution; however, as Atkinson (2017) shows, Pareto type I is not a perfect tool for studying top income shares and is rather "at best a convenient first summary of the extent of the income concentration." In addition, Jenkins (2017) showed (for the United Kingdom) that a Pareto type II is preferable to the Pareto type I typically used at the thresholds. Blanchet et al. (2018) also used a Pareto type II to fit income inequality at the top.

Gini coefficient assuming top coding by G_{top}.

Table 3.1: Montecarlo simulations

$1-p$	G_{ur}	G_{top}	G_{nr}
	$\alpha_r = 1.1$		
0.9	0.0057	0.0082	0.0752
0.95	0.0023	0.0033	0.0501
0.99	0.0002	0.0003	0.0156
	$\alpha_r = 1.5$		
0.9	0.0042	0.0085	0.1244
0.95	0.0015	0.0030	0.0733
0.99	0.0001	0.0002	0.018
	$\alpha_r = 2$		
0.9	0.0017	0.0068	0.1457
0.95	0.0006	0.0023	0.0818
0.99	0.0001	0.0002	0.0188
	$\alpha_r = 2.4$		
0.9	0.0003	0.0057	0.15
0.95	0.0001	0.0019	0.085
0.99	0.0001	0.0001	0.016

Mean Square Error of the difference between the real Gini coefficient (G), G_{ur}, G_{top} and G_{nr}. Average over 1000 simulations with $n = 100,000$, $\lambda = \frac{1}{40,000}$ (exponentinal parameter), α_r (real Pareto parameter) and $\alpha_s = 2.5$ (survey Pareto parameter).

Table 3.1 shows that a higher p generates a larger bias; however, the bias caused by underreporting or top coding is quite small, which implies that knowing an income distribution for underreporters is unnecessary and only the proportion of total underreported income is needed to correct the Gini coefficient.[13]

In contrast, the bias in a nonresponse context can approach 15 percentage points of the Gini coefficient. This bias depends on the proportion of nonrespondents, p, which, in practice, is difficult to determine. The next section proposes a methodology to estimate the missing population at the top p when only ϕ, μ_{1-p} (the average income of truthfull reporters), a top-coded value χ, and an estimate for the Gini coefficient of nonrespondent G_p are known.

[13]In this theoretical exercise we assume that we know $1-p$ but in real world applications p is very hard to estimate. The recent literature (Atkinson, 2017; Jenkins, 2017; Bourguignon, 2018) suggest to try with p=0.01, p=0.05 and p=0.1.

3.4 An extension of the Atkinson approximation in the case of nonresponse

This section proposes a methodology to estimate the proportion of the missing population p.[14] To understand the estimator's construction, assume that p (the proportion of nonrespondents), ϕ (the proportion of income belonging to nonrespondents), and G_p (the Gini coefficient of nonrespondents) are all known. Alvaredo (2011) shows the true Gini coefficient G is equal to

$$G = (1-p)(1-\phi)G_{1-p} + p\phi G_p + \phi - p, \tag{3.25}$$

where G_{1-p} is the observed population's Gini coefficient. Also, the observations of respondents are given by $(x_1, ..., x_n)$, and those of nonrespondents are $(x_{n+1}, ..., x_N)$, with $x_i \leq x_{i+1}\ \forall i$. We begin by constructing a synthetic distribution Ω for the income distribution such that the new distribution is composed by

$$\Omega \equiv ((x_1, ..., x_n), (x^*, ..., x^*)),$$

where x^* is a nonnegative value such that $x_n \leq x^* \leq \mu_p$ and μ_p is the average nonrespondent income. Notice that ϕ is divisible into two proportions, δ and λ, and δ can be defined as

$$\delta \equiv \frac{px^*}{(1-p)\mu_{1-p} + p\mu_p},$$

with μ_{1-p} being the average income of the respondents or average observed income. Also, define $\chi \equiv x_n$, the maximum value from the observed $1-p$ proportion of the population. Then, for instance, if $\delta = \phi$, then δ is formed by attributing the average nonrespondent income to each nonrespondent, or δ could be formed by assuming each nonrespondent earns the highest reported value χ. Consequently, λ is defined as $\lambda \equiv \phi - \delta$. Thus, (3.25)

[14] Blanchet, Flores, and Morgan (2018) propose a new methodology to find this proportion, what they call a merging point—the point at which tax data becomes more representative than household data.

can be written as

$$G = (1-\lambda)\left[(1-p)S^*_{1-p}G_{1-p} + S^*_p - p\right] + pS_pG_p + \lambda - \lambda p, \tag{3.26}$$

where $S^*_{1-p}(1-\lambda) = 1-\phi$, $S^*_p(1-\lambda) = \delta$. The Gini coefficient of the Ω synthetic income distribution is $\left[(1-p)S^*_{1-p}G_{1-p} + S^*_p - p\right]$.[15] Then, (3.26) can be written as:

$$G = (1-\lambda)\left[G^{**}\right] + \lambda + p\left(\phi G_p - \lambda\right), \tag{3.27}$$

where G^{**} is the Gini coefficient of the income distribution that includes the observed people and the synthetic nonrespondents. Also, $G^{**}_{ur} \equiv (1-\lambda)\left[G^{**}\right] + \lambda$ is the Gini approximation for correcting underreporting in this new synthetic income distribution. Thus,

$$G = G^{**}_{ur} + p\left[\phi G_p - \lambda\right]. \tag{3.28}$$

This means the Gini coefficient can be written as the sum of two terms: an underreporting Gini approximation and $p\left[\phi G_p - \lambda\right]$. If $\lambda = \phi G_p$, then $G = G^{**}_{nr}$. Choosing λ (i.e., the missing-income proportion left as underreporting in the synthetic income distribution Ω) transforms the true Gini coefficient into an underreporting Gini approximation.

Though, in real world applications, neither p nor G_p are known, p can be estimated via an estimated value for G_p (this estimator is $\hat{G}_p$). To estimate p assume that δ is composed of nonrespondents, each with income equal to χ, chose $\hat{\lambda} = \phi\hat{G}_p$, and use the fact that $S^*_p\left(1-\hat{\lambda}\right) = \hat{\delta}$. Moreover, $S^*_p = \frac{p\chi}{(1-p)\mu_{1-p}+p\chi}$ and $\hat{\delta} = \phi - \hat{\lambda}$. Thus,

$$\hat{p} = \frac{\phi\left(1-\hat{G}_p\right)\frac{\mu_{1-p}}{\chi}}{1-\phi+\phi\frac{\mu_{1-p}}{\chi}\left(1-\hat{G}_p\right)}. \tag{3.29}$$

Notice that if $\hat{G}_p = 1$, then $\hat{p} = 0$, the empirical counterpart to (3.27) returns to the Atkinson approximation for nonresponse. In addition, the bias of this approximation can

[15]Note that the Gini coefficient of $(x^*, ..., x^*)$ is 0.

be written as

$$\hat{G}^{**}_{nr} - G = -p\phi\left(G_p - \hat{G}_p\right) - (1-\lambda)(\hat{p}-p)(1+(1-\phi)G_p). \tag{3.30}$$

Let assume that adding nonrespodants did not change the Gini coefficient for the top of the distribution. In particular, let $\hat{G}^{*}_{1\%}$ the estimated Gini coefficient of the top 1% using observed data. Assume that the Gini coefficient of the true upper tail $\hat{G}_p$ (the upper tail that includes nonrespondants) is the same that $\hat{G}^{*}_{1}$.[16] An algorithm can thus be written to generate upper and lower bounds for the Gini coefficient under nonresponse.

Algorithm 1: Correcting for nonresponse *If the distribution for the bottom p fraction is known, then the following are also known: the average of the known distribution μ_{1-p}, the Gini coefficient G_{1-p}, and the maximum observed value in the observed income distribution, χ.*

i Compute $\hat{G}_{1-p}$ using observed data

ii Knowing ϕ, χ and assuming $G_p = \hat{G}^{*}_{1\%}$, $\hat{\lambda}$ and $\hat{p}$ can be obtained from (3.28) and (3.29) respectively.

iii Having $\hat{\lambda}$, μ and $\hat{p}$ means $\hat{\delta}$ can be computed from $\hat{\delta} = \phi - \hat{\lambda}$.

iv Having $\hat{\delta}$ means G^{**} can be computed as $G^{**} = (1-\hat{p})\left(1-\hat{\delta}\right)G_{1-p} + \hat{\delta} - \hat{p}$. The approximation for the Gini coefficient G can be estimated as $G^{**}_{nr} = \left(1-\hat{\lambda}\right)G^{**} + \lambda$.

Notice that surveys with high maximum values (χ) related to the average (μ) will estimate very low missing proportions, this is because an additional synthetic individual will cover a higher proportion of the missing income. The following section tests this methodology on household surveys.

[16] The Gini coefficient of the observed top 1% and the nonrespondants is $G_{1\%+p} = \frac{S_{1\%}}{S_{1\%}+S_p}\frac{0.01}{0.01+p}G^{*}_{1} + \frac{S_p}{S_{1\%}+S_p}\frac{p}{0.01+p}G_p + \frac{S_p}{S_{1\%}+S_p} - \frac{p}{0.01+p}$.

3.5 Empirical Applications

This section presents an empirical application, using data from two countries, Canada and Chile. The applications use the Atkinson approximation to solve underreporting and nonresponse at the top. The application is meant to test the methodology developed in the previous section.

3.5.1 Application. Income inequality and undistributed business profits.

It is known that the line between labour and capital is inherently imprecise for the small business sector, and it is certainly possible that tax accounting differs from the common-language way of separating labour from capital (Kopczuk, 2016). In some cases, part of the labour income is left inside the firm as undistributed business profits and not accounted in the year that income was generated. Indeed, undistributed business profits are typically an underreported component of income. It is not declared in personal tax records nor requested in surveys. Recent works (Adalæster et al., 2017; Flores et al., 2019; Fairfield and Jorrat, 2016; Gutierrez et al., 2015; Smith, Yagan Zidar and Zwick, 2019; Wolfson et al., 2016) show that including undistributed business profits in inequality measurements increases income inequality measures.

Moreover, the level of undistributed business profits is closely related to tax changes, as Smith, Yagan, Zidar and Zwick (2019) argue, an important proportion of top income is private "pass-through" business profit for tax proposes. Indeed, some tax reforms induces to keep business income inside the firm whereas others generate incentives to take out profits as dividends (see for instance the 2003 US dividend tax reform). Thus, not accounting for undistributed business income could lead to artificial changes that bias levels and trends of measured income inequality. Thus, the methodology developed here could be applied to estimate level trends of income inequality that are robust to tax changes and tax avoidance behaviour.

The Canadian data we used in this application comes from the Survey of Finan-

cial Security, while Chilean data comes from Encuesta de Caracterización Socioeconómica Nacional. The Survey of Financial Security, which has four available waves (1999, 2005, 2012, and 2016), provides a record of Canadian residents' market incomes, assets, and debts. The Encuesta de Caracterización Socioeconómica Nacional survey is a household-level survey that measures different types of income in Chile. We used household market income to compute each country's Gini coefficient. Canadian undistributed business profits data comes from CANSIM Table 36-10-0117-01 and Chilean application data comes from Flores et al. (2019).[17] We use Chile and Canada because both countries have integrated tax systems where the corporate tax paid by firms could be partially used as a credit for the dividends received by individuals. This creates incentives for tax planning.[18]

With the above data G_{ur} can be computed by estimating the Gini coefficient of the household survey G^*, and the proportion of total income corresponding to undistributed business profits z can be used to adjust the Gini coefficient. Here we are not correcting for nonresponse by assuming that undistributed business profits are pure underreporting G_{ur}. Otherwise, we can correct for nonresponse at the top and then correct for underreporting G^{**}_{nr}. For the latter estimation we need to estimate the proportion of nonrespondant, to do that, we will use the methodology developed in the previous section. Then, we compare G_{ur}, G^{**}_{nr} with the uncorrected Gini coefficient G^*.

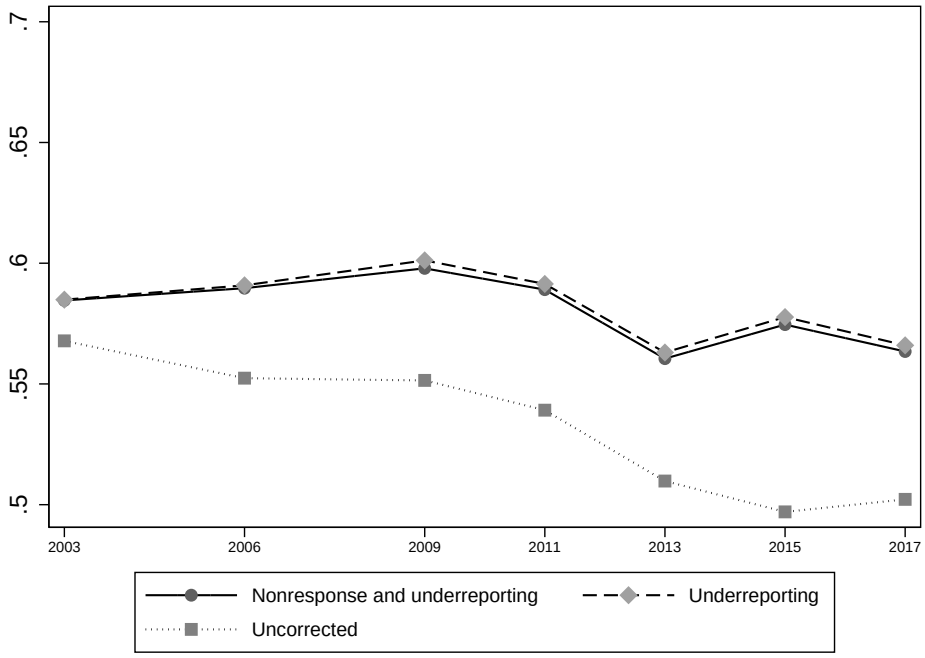
Results

Figure 3.5 shows the empirical application's Chilean results. The Gini coefficient decreased throughout 2003 to 2013, then increases from 2013 to 2015 to return to 2013 levels in 2017. This can be caused by retained earnings. In particular, retained earnings (as a proportion of total income) increased between 2003 and 2015. In this last year, retained earnings were 16 percent of the total income. However,retained earnings fell to 12 percent in 2017. This can be explained by the 2014 tax reform that reduced the integration rate between corporate and personal taxes along with others tax changes that reduces the incentives to retain earnings inside the firms.

[17]Canadian retained earnings data comes mostly from financial corporations.

[18]Tax planning using small businesses in Canada and Chile was broadly studied in the literature. See for instance Fairfield and Jorrat (2016); Lopez et al. (2016); Wolfson et al. (2016).

Figure 3.5: Corrected vs uncorrected Gini coefficient for Chile

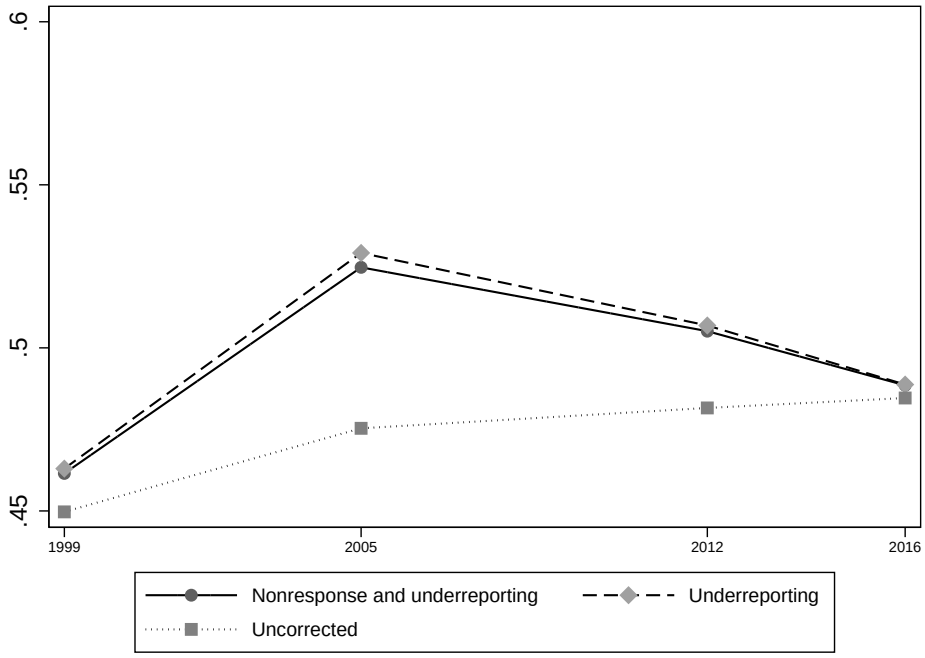


The Chilean Gini coefficient for income using Encuesta de Caracterización Socioeconómica Nacional information and Flores et al. (2016)

This result is robust for G_{ur} and G_{nr}^{**}. The difference between the adjusted Gini coefficient and the observed Gini coefficient increases during the analyzed period. In addition, the difference between G_{ur} and G_{nr}^{**} are tiny, this is because the maximum income reported in CASEN survey is very high compared to the average income.

Figure 3.6 shows Canadian results. The adjusted Gini coefficient increases between 1999 and 2005 and decreases between 2005 and 2016. However, the unadjusted Gini coefficient always increases between 1999 and 2016. One possible explanation for the huge jump in the Gini coefficient in 2005 is because retained earnings in 2005 where much more bigger than they were in other years. In particular, 2005 retained earnings were more than 10 percent of total income, whereas for other years, retained earnings were at most 5 percent. Retained earnings could be bigger in 2005 because it was that it was a year before the 2006 dividend tax reform that increases the dividend tax credits received from publicly traded firms and large CCPC. This reform creates incentives to pay more dividends instead of retain earnings inside firms.

Figure 3.6: Corrected vs uncorrected Gini coefficient for Canada



The Canadian Gini coefficient for income using SFS information and CANSIM Table 36-10-0117-01

Both examples shows that trends of the nonresponse and underreporting Gini coefficients for both countries are different than the observed Gini coefficient. In addition, given the survey structure and maximum values, we can conclude that retained earnings are more an underreporting issue than a nonresponse one.

3.6 Conclusion

This work studied the Gini coefficient for underreporting and nonresponse at the top two issues that recently attract the attention of several scholars that estimate incore inequality. The first contribution is that this work proves that underreporting or nonresponse does not necessarily result in a true Gini coefficient that is higher than the estimated Gini. That is, we are not necessarily estimating underestimated Gini coefficients when we use a household survey.

In addition, a correction of the Atkinson approximation approximates the Gini

coefficient for correcting underreporting or top coding well —this approximation can be used to correct inequality measurements of income sources (e.g., undistributed business profits) concentrated at the top but unreported in household surveys. A key feature of this underreporting approximation is that it is not necessary to know the proportion of underreporters and sometimes it is also not necessarily to know the missing population either. Thus, despite not being estimating the whole income distribution, contribute to Bourginon (2018) and Blanchet, Flores and Morgan (2018) by developing an adjustment where it is not necesseraly to find the size of the top neither the size of the missing people to obtain results close to the true ones.

Moreover, we developed a simple adjustment that combines top coding and the underreporting approximation to construct an estimation of the Gini coefficient in the presence of nonresponse and underreporting at the top. We estimate the missing population and then correcting for underreporting income. We applied this methodology to two countries, Chile and Canada where we show that retained earnings is more an underreporting issue than a nonresponse one. Our methodology can be easily replicated for other countries and additional undercoverage examples.

Conclusion

In this thesis, I present new evidence on economic inequality and intergenerational mobility in the Canadian and Chilean context. Chapter 1 first study that estimate intergenerational mobility in Chile using administrative records. We build a data set that links parental and child earnings using information from the formal labour sector and the place of residence of children during their adolescence. Our analysis reveals that intergenerational mobility at the national level is significantly lower than what was estimated in previous research. However, intergenerational mobility is extremely non-linear. We found that mobility is very high for the bottom 80 percent of the earnings distribution but is very persistent at the upper tail of the parental and child distributions.

In addition, Chile is a highly heterogeneous country in its intergenerational mobility measures at the regional level. For instance, Antofagasta, which is a mining region, has a probability of rags to riches higher than 0.3. This result is in line with what Conolly et al. (2018) founds for the US and Canada. Meanwhile, regions like Araucanía or El Maule have a circle of poverty probability higher than 0.3. It is worth digging a little deeper in future research to understand why those regions are so persistent in poverty.

We also find heterogeneity within the Metropolitan region, with municipalities having a circle of privilege probability higher than 0.7, and other municipalities with a circle of poverty probability closer to 0.3. We also learn that the variance of persistence at the top is higher than the variance of upward mobility. This means that the place of residence affects children of upper-earnings parents more than middle- or poor-class parents. Future research should focus on understanding the causes behind these differences. Although our work is descriptive in nature, it sheds lights on intergenerational mobility

in a highly unequal country that does not belong to the advanced economies.

Moreover, we make a some methodological contributions. We use RIF regressions and Kernel conditional densities to study intergenerational mobility at the top. Those tools help usus to show that intergenerational mobility is very persistent at the top in Chile. In addition, we differentiate the Gatsby curve for Chile and Santiago using two measures of intergenerational mobility: absolute intergenerational mobility and relative intergenerational mobility. We show that the Gatsby curve is valid for persistence and upward mobility for Chile but only for persistence for la Región Metropolitana. This help us to differentiate different mechanisms that may affect intergenerational mobility for Chile.

This work builds on previous national literature and brings the state of research up to the robustness of analysis seen among works in developed economies. As such, not only does it provide more useful information for academics; it also provides an important counterpoint to similar works from developed economies by analyzing intergenerational earnings mobility in a non-developed [o developing] economy in a way that can be contrasted with the results of that literature. We believe that, by providing a clearer picture of how intergenerational earnings mobility occurs in Chile at a regional level, this work can both inspire further research on the matter both in Chile and other developing economies. These results can also help Chilean authorities better understand how and where to apply certain related social/economic programs in order to improve their impact, as well as provide input for drawing up and discussing proposed bills affected by this study's results.

Chapter 2 argues in favour of using retained earnings as a measure of accrued capital gains, clarifies the difference between left money inside the firm for closely held corporations and for publicly traded firms. In particular, we make a conceptual contribution by showing that for closely held corporations retained earnings could have a consumption value in addition to the change of wealth value that the literature acknowledged. The intuition is that goods can be bought inside the firm to satisfy personal consumption.

In addition, we propose a methodology to impute corporate retained earnings to households. The convenience of this methodology is that it only needs market income

information from a household survey and aggregated retained earnings information from a household survey. We apply this methodology to Canada. We show that including accrued capital gains increases measured income inequality and, more importantly, changes the observed trend of inequality which is similar to what Davies, Fortin and Lemieux (2017) found for Canadian wealth.

We compare our imputation method with the broadly used capitalization imputation method. We obtain similar results by using the capitalization method to impute retained earnings. Even though the parametric imputation procedure developed here is not perfect, it gives results that are close enough to the capitalization method. This fact suggests that in the context of scarcity of data, for example, developing countries where we cannot access easily to administrative data, and this procedure could be useful for imputing capital income in such contexts.

Another application of the imputation methodology developed here can be used in the context of undercoverage at the upper tail. Recent literature, such as Flores (2018), show that because of nonresponse and underreporting, household surveys only account for 20 percent of the capital income that appears in national account data. Thus, because the parametric imputation procedure developed here only need a household survey and national account data, could be used to correct for underreporting of capital income in household surveys. Future research is needed to study the imputation procedure in such contexts.

Chapter 3 studied the Gini coefficient for underreporting and nonresponse at the top two issues that recently attract the attention of several scholars that estimate incore inequality. The first contribution is that this work proves that underreporting or nonresponse does not necessarily result in a true Gini coefficient that is higher than the estimated Gini. That is, we are not necessarily estimating underestimated Gini coefficients when we use a household survey.

In addition, a correction of the Atkinson approximation approximates the Gini coefficient for correcting underreporting or top coding well —this approximation can be used to correct inequality measurements of income sources (e.g., undistributed business

profits) concentrated at the top but unreported in household surveys. A key feature of this underreporting approximation is that it is not necessary to know the proportion of underreporters and sometimes it is also not necessarily to know the missing population either. Thus, despite not being estimating the whole income distribution, contribute to Bourginon (2018) and Blanchet, Flores and Morgan (2018) by developing an adjustment where it is not necesseraly to find the size of the top neither the size of the missing people to obtain results close to the true ones.

Moreover, we developed a simple adjustment that combines top coding and the underreporting approximation to construct an estimation of the Gini coefficient in the presence of nonresponse and underreporting at the top. We estimate the missing population and then correcting for underreporting income. We applied this methodology to two countries, Chile and Canada where we show that retained earnings is more an underreporting issue than a nonresponse one. Our methodology can be easily replicated for other countries and additional undercoverage examples.

Bibliography

[1] Acciari, Paolo and Polo, Alberto and Violante, Gianluca, 'And Yet, it Moves': Intergenerational Mobility in Italy (April 2019). CEPR Discussion Paper No. DP13646. Available at SSRN: https://ssrn.com/abstract=3368143.

[2] Alesina, A., Hohmann, S., Michalopoulos, S., and Papaioannou, E. (2019). Intergenerational Mobility in Africa (No. w25534). National Bureau of Economic Research.

[3] Alstadsæter, A., Jacob, M., Kopczuk, W., and Telle, K. (2017). Accounting for Business Income in Measuring Top Income Shares: Integrated Accrual Approach Using Individual and Firm Data from Norway (No. w22888). National Bureau of Economic Research.

[4] Alstadsæter, A., Johannesen N. and Zucman G. (2018). Who owns the wealth in tax havens? Macro evidence and implications for global inequality. Journal of Public Economics.

[5] Alvaredo, F. (2011). A note on the relationship between top income shares and the Gini coefficient. Economics Letters, 110(3), 274-277.

[6] Alvaredo, F., Chancel, L., Piketty, T., Saez, E., and Zucman, G. (2017). World inequality report 2018. The World Inequality Lab, http://wir2018. wid. world

[7] Aoki, S., and Nirei, M. (2016). Zipf's Law, Pareto's Law, and the Evolution of Top Incomes in the US.

[8] Armour, P., Burkhauser, R. V. and Larrimore, J.: Deconstructing income and income inequality measures: A crosswalk from market income to comprehensive income. Am. Econ. Rev. 103(3), 173-177 (2013).

[9] Asher, S., Novosad, P., and Rafkin, C. (2018). Intergenerational Mobility in India: Estimates from New Methods and Administrative Data. World Bank Working Paper. Available at: http://www. dartmouth. edu/ novosad/anr-india-mobility. pdf (accessed December 2018).

[10] Atkinson, A.B., (1970). On the measurement of inequality. Journal of economic theory, 2(3), 244-263.

[11] Atkinson, A. B. (2007). Measuring Top Incomes: Methodological Issues, in A. Atkinson and T. Piketty (editors) Top Incomes over the Twentieth Century: A Contrast Between Continental European and English-Speaking Countries, Oxford: Oxford University Press.

[12] Atkinson, A. B. (2017). Pareto and the upper tail of the income distribution in the UK: 1799 to the present. Economica, 84(334), 129-156.

[13] Atkinson, A. B., Piketty, T., and Saez, E. (2011). Top incomes in the long run of history. Journal of economic literature, 49(1), 3-71.

[14] Atria, J., Flores I., Sanhueza C., and Mayer R. (2018). Top Income in Chile: A Historical Perspective of Income Inequality (1964-2015), WID.world Working Paper 2018/11.

[15] Auten, G., and Splinter D. (2016). Using Tax Data to Measure Long-Term Trends in U.S. Income Inequality. mimeo.

[16] Banerjee, A., Yakovenko, V. M., and Di Matteo, T. (2006). A study of the personal income distribution in Australia. Physica A: statistical mechanics and its applications, 370(1), 54-59.

[17] Becker, Gary S., and Nigel Tomes. (1979). An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility. Journal of Political Economy 87 (6): 1153-89.

[18] Becker, Gary S., Scott Duke Kominers, Kevin M. Murphy, and Jörg L. Spenkuch. (2018). A Theory of Intergenerational Mobility. Journal of Political Economy 126 (S1): S7-S25.

[19] Bédard-Pagé, G., Demers, A., Tuer, E., and Tremblay, M. (2016). Large Canadian public pension funds: A financial system perspective. Bank of Canada Financial System Review, 33-38.

[20] Bernheim, B. D., Ray, D., and Yeltekin, (2015). Poverty and Self-Control. Econometrica, 83(5), 1877-1911.

[21] Björklund, A., and Jäntti, M. (1997). Intergenerational income mobility in Sweden compared to the United States. The American Economic Review, 87(5), 1009-1018.

[22] Black, S.., and Devereux, J. (2011). Recent Developments in Intergenerational Mobility. In Handbook of Labor Economics, edited by David Card and Orley Ashenfelter, 4, Part B:1487-1541. Elsevier.

[23] Blanchet, T., Garbinti, B., Goupille-Lebret, J. and Martinez-Toledano, C. (2018). Applying Generalized Pareto Curves to Inequality Analysis, WID.world Working Paper 2018/3.

[24] Blanchet, T., Flores, I. and Morgan, M. (2018). The Weight of the Rich: Improving Surveys Using Tax Data World Inequality Lab. WID Working Paper Series No. 2018/12.

[25] Boadway, R., and Bruce, N. (1992). Problems with integrating corporate and personal income taxes in an open economy. Journal of Public Economics, 48(1), 39-66.

[26] Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., and Ziliak, J. P. (2018). Trouble in the Tails? What We Know about Earnings Nonresponse Thirty Years after Lillard, Smith, and Welch. The Journal of Political Economy. Forthcoming.

[27] Bourguignon, F. (2018). Simple adjustments of observed distributions for missing income and missing people. The Journal of Econonomic Inequalily.

[28] Bratsberg, B., Røed, K., Raaum, O., Naylor, R., Jäntti, M., Eriksson, T., and Österbacka, E. (2007). Nonlinearities in intergenerational earnings mobility: consequences for cross-country comparisons. The Economic Journal, 117(519), C72-C92.

[29] Burkhauser, R. V., Feng, S., Jenkins, S. P., and Larrimore, J. (2011). Estimating trends in US income inequality using the Current Population Survey: the importance of controlling for censoring. The Journal of Economic Inequality, 9(3), 393-415

[30] Burkhauser, R. V., Feng, S., Jenkins, S. P. and Larrimore, J. (2012). Recent trends in top income shares in the United States: reconciling estimates from March CPS and IRS tax return data. Review of Economics and Statistics 94(2), 371-388.

[31] Burkhauser, R. V., Hahn, M. H., and Wilkins, R. (2015). Measuring top incomes using tax record data: A cautionary tale from Australia. The Journal of Economic Inequality, 13(2), 181-205.

[32] Burkhauser, R. V., Hérault, N., Jenkins, S. P., and Wilkins, R. (2017). Survey under-coverage of top incomes and estimation of inequality: what is the role of the UK's SPI adjustment?. Fiscal Studies.

[33] Burkhauser, R. V., Hérault, N., Jenkins, S. P., and Wilkins, R. (2018). Top incomes and inequality in the UK: reconciling estimates from household survey and tax return data. Oxford Economic Papers.

[34] Cahill, S. A. (2007). Corporate Income Tax Rate Database-Canada and the Provinces, 1960-2005 (No. 52726). Agriculture and Agri-Food Canada.

[35] Caner, M., and Hansen, B. E. (2004). Instrumental variable estimation of a threshold model. Econometric Theory, 20(5), 813-843.

[36] Celhay, P., Sanhueza, C., and Zubizarreta, J. (2010). Intergenerational Mobility of Income and Schooling: Chile 1996-2006.

[37] Ceriani, L. and P. Verme (2012). The origins of the gini index: Extracts from variabilita' e mutabilita' (1912) by corrado gini. Journal of Economic Inequality 10 (3), 421-443.

[38] Ceriani, L. and Verme, P., (2019). The inequality of extreme incomes. Ecineq WP 490.

[39] Chadwick, L., and Solon, G. (2002). Intergenerational income mobility among daughters. American Economic Review, 92(1), 335-344.

[40] Chanfreau, J., and Burchardt, T. (2008). Equivalence scales: rationales, uses and assumptions. Scottish Government. Available at www.scotland.gov.uk/Resource/Doc/933/0079961.pdf.

[41] Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R. and Narang, J., (2017). The fading American dream: Trends in absolute income mobility since 1940. Science, 356(6336), pp.398-406.

[42] Chetty, R., Hendren, N., Kline, P. and Saez, E., (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. The Quarterly Journal of Economics, 129(4), pp.1553-1623.

[43] Chetty, R. and Hendren, N., (2018a). The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects. The Quarterly Journal of Economics, 133(3), pp.1107-1162.

[44] Chetty, R., and Hendren N. (2018b). The Impacts of Neighborhoods on Intergenerational Mobility II: County Level Estimates. Quarterly Journal of Economics 133 (3): 1163-1228.

[45] Chotikapanich, D., Griffiths, W. E., and Rao, D. P. (2007). Estimating and combining national income distributions using limited data. Journal of Business and Economic Statistics, 25(1), 97-109.

[46] Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. SIAM review, 51(4), 661-703.

[47] Clementi, F., and Gallegati, M. (2016). Distribution of Income and Wealth. Springer International Publishing.

[48] Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. London: Springer.

[49] Corak, M. (2013). Income inequality, equality of opportunity, and intergenerational mobility. Journal of Economic Perspectives, 27(3), 79-102.

[50] Corak M., (2019) The Canadian Geography of Intergenerational Income Mobility, The Economic Journal, , uez019, https://doi.org/10.1093/ej/uez019

[51] Corak, M., and Heisz, A. (1999). The intergenerational earnings and income mobility of Canadian men: Evidence from longitudinal income tax data. Journal of Human Resources, 504-533.

[52] Corvalan, A. (2014). The impact of a marginal subsidy on Gini Indices. Review of Income and Wealth, 60(3), 596-603.

[53] Cowell, F. A. and E. Flachaire (2015). Statistical methods for distributional analysis. In Handbook of income distribution, Volume 2. Elsevier, pp. 359-465.

[54] Cowell,F.A. and Flachaire, E. (2018) Inequality Measurement and the Rich: Why inequality increased more than we thought. Public Economics Programme Discussion Paper, 36, STICERD, LSE.

[55] Cowell, F. A., and Van Kerm, P. (2015). Wealth inequality: A survey. Journal of Economic Surveys, 29(4), 671-710.

[56] Daude, C., and Robano, V. (2015). On intergenerational (im) mobility in Latin America. Latin American Economic Review, 24(1), 9.

[57] Davies, J. B., and Di Matteo, L. (2018). Filling the gap: Long run Canadian wealth inequality in international context (No. 2018-1). Research Report.

[58] Davies, J. B., Fortin, N. M., and Lemieux, T. (2017). Wealth inequality: Theory, measurement and decomposition. Canadian Journal of Economics/Revue canadienne d'économique, 50(5), 1224-1261.

[59] Davidson, R., and Flachaire, E. (2007). Asymptotic and bootstrap inference for inequality and poverty measures. Journal of Econometrics, 141(1), 141-166.

[60] Deutscher, N., and Mazumder, B. (2020). Intergenerational mobility across Australia and the stability of regional estimates. Labour Economics, 66, 101861.

[61] Diaz-Bazan, T. (2015). Measuring inequality from top to bottom. World Bank Policy Research Paper no. 7237. Washington, DC: World Bank.

[62] Díaz, J. D., Gutierrez, P., and, Tapia, P. (2020). The exponential Pareto model with hidden income processes: Evidence from Chile. Physica A: Statistical Mechanics and its Applications, 125196.

[63] Diebold, F. X., and Chen, C. (1996). Testing structural stability with endogenous breakpoint a size comparison of analytic and bootstrap procedures. Journal of Econometrics, 70(1), 221-241. Chicago.

[64] Dragulescu, A., and Yakovenko, V. M. (2000). Statistical mechanics of money. The European Physical Journal B-Condensed Matter and Complex Systems, 17(4), 723-729.

[65] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics, 7(1), 1-26.

[66] Emran, M. S., Ferreira, F. H., Jiang, Y., and Sun, Y. (2019). Intergenerational Educational Mobility in Rural Economy: Evidence from China and India. Available at SSRN 3393904.

[67] Fairfield, T., and Jorratt De Luis, M. (2016). Top Income Shares, Business Profits, and Effective Tax Rates in Contemporary Chile. The Review of Income and Wealth, 62(1), S120-S144.

[68] Fisher, S. J. (1994). Asset trading, transaction costs and the equity premium. Journal of Applied Econometrics, 9(S1).

[69] Flores, I. (2018). Income Under the Carpet: What Gets Lost Between the Measure of Capital Shares and Inequality, Working Paper, Dec. 2018

[70] Flores, I., Atria, J., Sanhueza, C. and Mayer, R., (2019). Top Incomes in Chile: A Historical Perspective of Income Inequality (1964-2017). Review of Income and Wealth (forthcoming).

[71] Fortin, N., Green, D. A., Lemieux, T., Milligan, K., and Riddell, W. C. (2012). Canadian inequality: Recent developments and policy options. Canadian Public Policy, 38(2), 121-145.

[72] Fortin, N., and Lefebvre, S. (1998). Intergenerational income mobility in Canada. Labour markets, social institutions, and the future of Canada's children, (89-553).

[73] Fuest, C., and Huber, B. (2000). Can corporate-personal tax integration survive in open economies? Lessons from the German tax reform. FinanzArchiv/Public Finance Analysis, 57(4), 514-524.

[74] Gabaix, X. (2009). Power laws in economics and finance. Annu. Rev. Econ., 1(1), 255-294.

[75] Gabaix, X., Lasry, J. M., Lions, P. L., and Moll, B. (2016). The dynamics of inequality. Econometrica, 84(6), 2071-2111.

[76] Genicot, G., and Ray, D. (2017). Aspirations and inequality. Econometrica, 85(2), 489-519.

[77] Green, D. A. Riddell, W.C. and S-Hilaire F. (2016).Income Inequality: The Canadian Story. Montreal: Institute for Research on Public Policy.

[78] Güell, M., Rodríguez Mora, J.V. and Telmer, C.I., (2015). The informational content of surnames, the evolution of intergenerational mobility, and assortative mating. The Review of Economic Studies, 82(2), pp.693-735.

[79] Güell, M., Pellizzari, M., Pica, G. and Rodríguez Mora, J.V., (2018). Correlating social mobility and economic outcomes. The Economic Journal, 128(612), pp.F353-F403. (2018): Correlating Social Mobility and Economic Outcomes, The Economic Journal, 128(July), F353-F403.

[80] Gutierrez, P., Lopez, R. E., and Figueroa, E. (2015). Top income measurement and undistributed profits. Economics Letters, 134, 138-140.

[81] Haile, G. A. (2018). Intergenerational Mobility in Socio?economic Status in Ethiopia. Journal of International Development, 30(8), 1392-1413.

[82] Haig, R. M. (1921). The concept of income-economic and legal aspects. The Federal Income Tax, 1(7).

[83] Hansen, B. E. (2000). Sample splitting and threshold estimation. Econometrica, 68(3), 575-603.

[84] Heidrich, S. (2017). Intergenerational mobility in Sweden: a regional perspective. Journal of Population Economics, 30(4), 1241-1280.

[85] Higgins, S., N. Lustig, and A. Vigorito (2018). The rich underreport their income: Assessing bias in inequality estimates and correction methods using linked survey and tax data. Ecineq WP 475.

[86] Hlasny, V. and P. Verme (2018). Top incomes and inequality measurement: A comparative analysis of correction methods using the eu silc data. Econometrics 6 (2), 1-21.

[87] Hundenborn, J., Woolard, I., and Jellema, J. (2018). The effect of top incomes on inequality in South Africa. International Tax and Public Finance, 1-30.

[88] Jagielski, M., and Kutner, R. (2013). Modelling of income distribution in the European Union with the Fokker-Planck equation. Physica A: Statistical Mechanics and its Applications, 392(9), 2130-2138.

[89] Jäntti, M., Bratsberg, B., Roed, K., Raaum, O., Naylor, R., Osterbacka, E., Bjorklund, A. and Eriksson, T., (2006). American exceptionalism in a new light: a comparison of intergenerational earnings mobility in the Nordic countries, the United Kingdom and the United States.

[90] Jäntti, M., and Jenkins, S. P. (2015). Income mobility. In Handbook of income distribution (Vol. 2, pp. 807-935). Elsevier.

[91] Jenkins, S. P. (2017). Pareto Models, Top Incomes and Recent Trends in UK Income Inequality. Economica, 84(334), 261-289.

[92] Jones, C. M. (2002). A century of stock market liquidity and trading costs.

[93] Kapetanios, G. (2010). Testing for exogeneity in threshold models. Econometric Theory, 26(1), 231-259.

[94] Kim, J. (2015). The Effect of the Top Marginal Tax Rate on Top Income Inequality. mimeo.

[95] King, M. A. (1974). Taxation and the cost of capital. The Review of Economic Studies, 41(1), 21-35.

[96] Kleiber, C., and Kotz, S. (2003). Statistical size distributions in economics and actuarial sciences (Vol. 470). John Wiley and Sons.

[97] Konüs, A.A. (1924), The Problem of the True Index of the Cost of Living, translated in Econometrica 7, (1939), 10-29.

[98] Kopczuk, W. (2016). Measuring Income and Wealth at the Top Using Administrative and Survey Data: Comment, Brookings Papers on Economic Activity, Spring 2016, 321-27.

[99] Lambert, S., Ravallion, M., and Van de Walle, D. (2014). Intergenerational mobility and interpersonal inequality in an African economy. Journal of Development Economics, 110, 327-344.

[100] Lee, C. I., and Solon, G. (2009). Trends in intergenerational income mobility. The Review of Economics and Statistics, 91(4), 766-772.

[101] Lemieux, T., and Riddell C. (2015). Who are Canada‘s Top 1 percent?. Income Inequality: The Canadian Story, edited by David A. Green, W. Craig Riddell and France St-Hilaire. The Institute for Research on Public Policy, Montreal.

[102] Lopez, R., E. Figueroa and P. Gutierrez (2016). Fundamental accrued capital gains and the measurement of top incomes: An application to Chile. The Journal of Economic Inequality 14.4 (2016): 379-394.

[103] Lustig, N. (2018), The Missing Rich in Household Surveys: Causes and Correction Methods. CEQ Working Paper 75

[104] Luttmer, E. G. (2015). An Assignment Model of Knowledge Diffusion and Income Inequality. Federal Reserve Bank of Minneapolis Research Department Staff Report, 715.

[105] Meghir, C., and Pistaferri, L. (2011). Earnings, consumption and life cycle choices. Handbook of labor economics, 4, 773-854.

[106] Mehra, R., and Prescott, E. C. (1985). The equity premium: A puzzle. Journal of monetary Economics, 15(2), 145-161.

[107] Milligan, K. (2016), Canadian Tax and Credit Simulator. Database, software and documentation, Version 2016-2.

[108] Milligan, K., and Smart, M. (2015). Taxation and top incomes in Canada. Canadian Journal of Economics/Revue canadienne d’économique, 48(2), 655-681.

[109] Modalsli, J. (2017). Decomposing Global Inequality. Review of Income and Wealth 63.3 : 445-463.

[110] Neidhöfer, G. (2019) Intergenerational mobility and the rise and fall of inequality: Lessons from Latin America. Available at SSRN 2740395 .

[111] Neidhöfer, G., Serrano, J., and Gasparini, L. (2018). Educational inequality and intergenerational mobility in Latin America: A new database. Journal of Development Economics, 134, 329-349.

[112] Nirei, M. (2009). Pareto Distributions in Economic Growth Models (No. 09-05). Institute of Innovation Research, Hitotsubashi University.

[113] Núñez, J. I., and Miranda, L. (2010). Intergenerational income mobility in a less-developed, high-inequality context: The case of Chile. The BE Journal of Economic Analysis and Policy, 10(1).

[114] Ñúñez, J., and Miranda, L. (2011). Movilidad intergeneracional del ingreso y la educación en zonas urbanas de Chile. Estudios de economía, 38(1), 195-221.

[115] Osberg, L. (2017). On the limitations of some current usages of the Gini Index. Review of Income and Wealth, 63(3), 574-584.

[116] Paredes D., Iturra V. and Lufin M. (2016) A Spatial Decomposition of Income Inequality in Chile, Regional Studies, 50:5, 771-789, DOI: 10.1080/00343404.2014.933798

[117] Piketty, T., Saez, E., and Zucman, G. (2018). Distributional national accounts: methods and estimates for the United States. The Quarterly Journal of Economics, 133(2), 553-609.

[118] Saez, E., and Veall, M. R. (2005). The evolution of high incomes in Northern America: lessons from Canadian evidence. The American Economic Review, 95(3), 831-849.

[119] Saez, E., and Zucman, G. (2016). Wealth inequality in the United States since 1913: Evidence from capitalized income tax data. The Quarterly Journal of Economics, 131(2), 519-578.

[120] Schnelle, K. (2015): Intergenerational Mobility in Norway: Transition Probabilities and Directional Rank Mobility, Mimeo, University of Bergen.

[121] Schneebaum, A., Rumplmaier, B., and Altzinger, W. (2015). Gender in intergenerational educational persistence across time and place. Empirica, 42(2), 413-445.

[122] Sehnbruch, K., and Carranza, R. (2015). The Chilean system of unemployment insurance savings accounts. Univ. de Chile, Department de Economía.

[123] Silva, A. C., and Yakovenko, V. M. (2004). Temporal evolution of the "thermal" and "superthermal" income classes in the USA during 1983-2001. EPL (Europhysics Letters), 69(2), 304.

[124] Simons, H. C. (1938). Personal income taxation: The definition of income as a problem of fiscal policy (pp. 28-30). Chicago: University of Chicago Press.

[125] Smeeding, T. M. and Thompson, J. P. (2011). Recent trends in income inequality: labor, wealth and more complete measures of income. Res. Labor Econ. 32, 1-50.

[126] Smith, M., Yagan, D., Zidar, O. and Zwick, E., (2019). Capitalists in the Twenty-first Century. The Quarterly Journal of Economics, 134(4), pp.1675-1745.

[127] Simard-Duplain, G., and St-Denis, X. (2020). Exploration of the Role of Education in Intergenerational Income Mobility in Canada: Evidence from the Longitudinal and International Study of Adults. Canadian Public Policy, (aop), e2019072.

[128] Solon, G. Cross-country differences in intergenerational earnings mobility. Journal of Economic Perspectives 16.3 (2002): 59-66.

[129] Sørensen, P. B. (2014). Taxation of shareholder income and the cost of capital in a small open economy. No. 5091. CESifo Working Paper.

[130] Toda, A. A., and Walsh, K. (2015). The double power law in consumption and implications for testing Euler equations. Journal of Political Economy, 123(5), 1177-1200.

[131] Torche, F. (2014). Intergenerational mobility and inequality: The Latin American case. Annual Review of Sociology, 40, 619-642.

[132] Torche, F. (2005). Unequal but fluid: social mobility in Chile in comparative perspective. American Sociological Review, 70(3), 422-450.

[133] Ugarte, S. M., Grimshaw, D., and Rubery, J. (2015). Gender wage inequality in inclusive and exclusive industrial relations systems: a comparison of Argentina and Chile. Cambridge Journal of Economics, 39(2), 497-535.

[134] Veall, M. R. (2012). Top income shares in Canada: recent trends and policy implications. Canadian Journal of Economics/Revue canadienne d'économique, 45(4), 1247-1272.

[135] Wolfson, M.,Veall M., Brooks N., and Murphy B. (2016). Piercing the Veil-Private Corporations and the Income of the Affluent. Canadian Tax Journal 64, no. 1 1-30.

[136] Yu, P. (2014). The bootstrap in threshold regression. Econometric Theory, 30(3), 676-714.

[137] Zimmerman, S. D. (2019). Elite colleges and upward mobility to top jobs and top incomes. American Economic Review, 109(1), 1-47.

Appendix A

Appendix for Chapter 1

A.1 Data appendix

This appendix discussed the linkage and cleaning process of the data used in this dissertation. Fist we deliver to the Social Registry Services with 10,000,000 of Tax Numbers (RUT) so that they can create the matching between parents and Children. Then, we work with the unemployment insurance program dataset and we use the RUT as linkage key. Finally, we use information from the education ministry to obtain children address, where we use RUT as linkage key.

Table A.1: Linkage units

Subset	N
Total children cohort	2,817,428
Children that worked more than 6 months	1,889,107
Total parental cohort	3,021,785
Parents that works more than 6 months	1,275,664
Total number of linkages (excluding duplicates)	505,524

A.1 shows the total number of children that are in the cohort studied. Then we show the total amount of children that works more than 5 months and who earned more than half the minimum wage in average. In addition, we included the total number of

potential parents and the number of parents that works more than 5 months receiving more than half the minimum wage. Finally, this table shows the number of total linkages used for the estimates. Comparing with Corak and Heisz (1999) we have more than 100,000 linkage for a country with half the size of the Canadian population.

Table A.2: Educational linkages

Grade included	Linkage rate
12th grade	76.17%
Including also 11th grade	84.41%
Including also 10th grade	87.77%
Including also 9th grade	91.96%
Including also 8th grade	92.87%
Including also 7th grade	93.37%

A.2 shows the proportion of the children that worked more than 5 months with wages in average greater than half the minimum wage. When we include only 12 grade, the linkage rate is only 76.17%. When we include 11th grade information, the linkage rate is 84.41% and when we include 12th to 7th grade, the linkage rate is 93.37%. This number is close to Chetty et al. (2014) and Acarci et al. (2020) linkage rates.

Table A.3: Age distribution

Age	CASEN	UID
28	297,739	313,696
29	257,271	301,839
30	292,403	282,430
31	205,791	254,602
32	233,878	234,762
33	246,575	211,548
Total	1,533,657	1,598,877

In addition, we can see from Table A.3 that the numbers per age are somewhat

similar between CASEN survey, which is a national representative survey and UID information.

Table A.4: Sex distribution

Source	Female percentage
UID	47.32%
CASEN 2017	51.41%

Finally, from Table A.4 we can see that the female proportion is lower in the UID database than the national proportion given that the female labour participation in the formal private sector is lower than the female proportion.

A.2 Additional regressions

Table A.5: Period robustness checks for IGE

	(1)	(2)	(3)	(4)	(5)	(6)
y_p	0.288***	0.297***	0.311***	0.323***	0.333***	0.354***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.005)
Constant	9.506***	9.426***	9.298***	9.193***	9.135***	8.942***
	(0.016)	(0.018)	(0.021)	(0.027)	(0.039)	(0.058)
Observations	505,524	416,818	282,979	173,683	83,668	39,160
R-squared	0.091	0.098	0.108	0.117	0.124	0.134

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Earnings are measured as average earnings over the months where a children-parents pair report positive earnings over the studied 5-year period. We keep individuals that appear at least 6 times with positive earnings in the dataset with average earnings greater than half of the corresponding minimum wage. Columns (1) to (4) report results for male and female children. (1) considers individuals with at least 6 months of positive earnings, (2) considers individuals with at least 12 months of positive earnings, (3) considers individuals with at least 24 months of positive earnings, (4)considers individuals with at least 36 months of positive earnings, (5) considers individuals with at least 48 months of earnings and (6) considers individuals with at least 54 months of earnings.

Table A.6: Age robustness checks for IGE

	(1)	(2)	(3)	(4)	(5)	(6)
y_p	0.289***	0.299***	0.313***	0.326***	0.335***	0.353***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)	(0.005)
Constant	9.496***	9.411***	9.278***	9.167***	9.116***	8.962***
	(0.017)	(0.019)	(0.023)	(0.029)	(0.042)	(0.062)
Observations	440,222	363,667	247,419	152,032	73,177	34,214
R-squared	0.093	0.101	0.112	0.121	0.128	0.137

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Earnings are measured as average earnings over the months where a children-parents pair report positive earnings over the studied 5-year period. We keep individuals that appear at least 6 times with positive earnings in the dataset with average earnings greater than half of the corresponding minimum wage. Columns (1) to (4) report results for male and female children. (1) considers individuals with at least 6 months of positive earnings, (2) considers individuals with at least 12 months of positive earnings, (3) considers individuals with at least 24 months of positive earnings, (4)considers individuals with at least 36 months of positive earnings, (5) considers individuals with at least 48 months of earnings and (6) considers individuals with at least 54 months of earnings.

A.3 Penn parade

To evaluate the validity of the earnings distribution, we will plot the Pen parade -the relationship between earnings percentiles and the percentiles, for the UID dataset and the CASEN dataset from 2003 (after correcting for inflation).

Table A.7: Period robustness checks for rank-rank correlation

	(1)	(2)	(3)	(4)	(5)	(6)
r^p	0.257***	0.264***	0.273***	0.279***	0.281***	0.293***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.005)
Constant	37.833***	39.075***	41.208***	43.635***	46.905***	49.420***
	(0.087)	(0.096)	(0.119)	(0.152)	(0.215)	(0.312)
Observations	440,222	363,667	247,419	152,032	73,177	34,214
R-squared	0.065	0.069	0.075	0.080	0.086	0.096

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Earnings are measured as average earnings over the months where a children-parents pair report positive earnings over the studied 5-year period. We keep individuals that appear at least 6 times with positive earnings in the dataset with average earnings greater than half of the corresponding minimum wage. Columns (1) to (4) report results for male and female children. (1) considers individuals with at least 6 months of positive earnings, (2) considers individuals with at least 12 months of positive earnings, (3) considers individuals with at least 24 months of positive earnings, (4)considers individuals with at least 36 months of positive earnings, (5) considers individuals with at least 48 months of earnings and (6) considers individuals with at least 54 months of earnings.

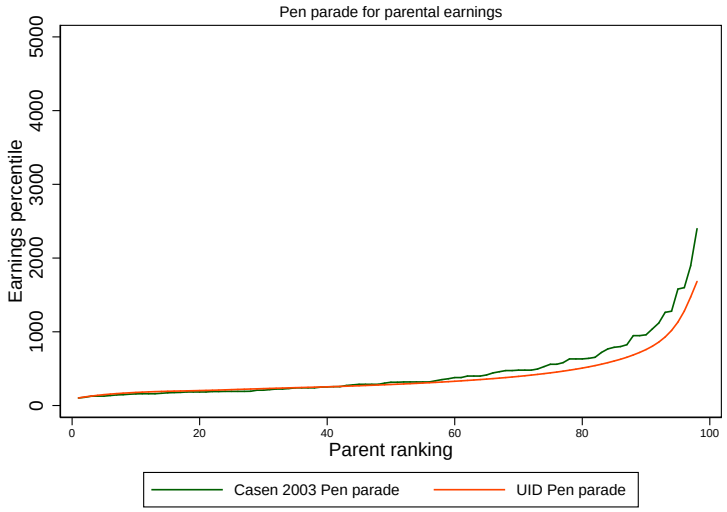
Table A.8: Age robustness checks for rank-rank correlation

	(1)	(2)	(3)	(4)	(5)	(6)
r^p	0.257***	0.264***	0.273***	0.279***	0.281***	0.293***
	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.005)
Constant	37.833***	39.075***	41.208***	43.635***	46.905***	49.420***
	(0.087)	(0.096)	(0.119)	(0.152)	(0.215)	(0.312)
Observations	440,222	363,667	247,419	152,032	73,177	34,214
R-squared	0.065	0.069	0.075	0.080	0.086	0.096

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

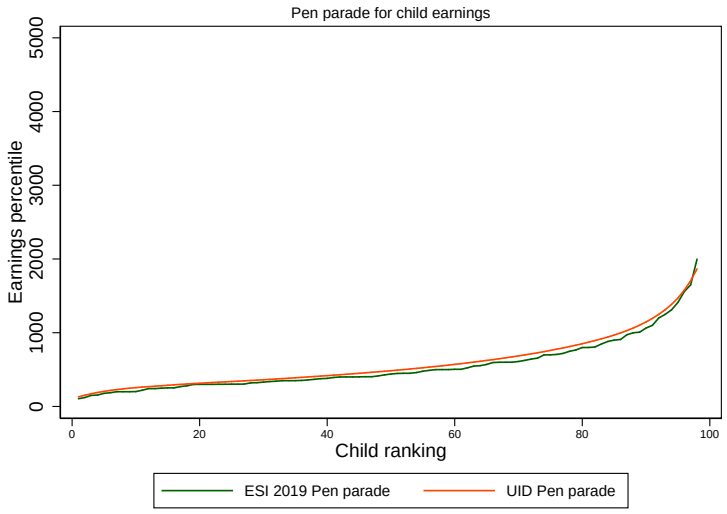
Earnings are measured as average earnings over the months where a children-parents pair report positive earnings over the studied 5-year period. We keep individuals that appear at least 6 times with positive earnings in the dataset with average earnings greater than half of the corresponding minimum wage. Columns (1) to (4) report results for male and female children. (1) considers individuals with at least 6 months of positive earnings, (2) considers individuals with at least 12 months of positive earnings, (3) considers individuals with at least 24 months of positive earnings, (4)considers individuals with at least 36 months of positive earnings, (5) considers individuals with at least 48 months of earnings and (6) considers individuals with at least 54 months of earnings.

Figure A.1: Pen parade for parental earnings



This figure shows the parental Pen parade using the Casen 2003 survey and the UID information.

Figure A.2: Pen parade for child earnings



This figure shows the children Pen parade using the ESI 2019 survey and the UID information.

As we can see, both parental and children earnings are close to other information sources that includes all workers.

A.4 Additional Tables

Table A.9: Descriptive statistics of the UIP database

Year	N	Min	Mean	Max	P1	P5	Median	P95	P99
2003	1,349,584	2	384,908	12,379,695	18,804	67,062	267,458	1,105,923	2,475,924
2004	1,849,529	2	428,928	12,432,321	23,508	76,579	289,927	1,316,578	2,486,467
2005	2,337,830	2	456,767	11,231,467	24,410	86,533	306,313	1,429,162	2,477,661
2006	2,701,345	2	491,342	7,432,997	27,325	99,207	327,247	1,576,341	2,477,666
2007	3,103,157	1	526,906	12,848,755	30,519	101,730	350,557	1,663,517	2,495,274
2008	3,309,297	1	538,045	34,705,504	26,991	97,940	353,730	1,720,412	2,479,460
2009	3,419,851	1	581,561	21,906,338	28,923	108,215	382,093	1,890,408	2,472,360
2010	3,742,474	1	622,926	8,101,156	30,870	110,580	407,878	2,038,509	2,678,778
2011	4,052,453	1	655,923	7,992,452	33,467	112,763	434,478	2,158,657	2,734,946
2012	4,286,460	1	698,879	10,768,408	38,684	118,068	468,135	2,296,852	2,777,116
2013	4,404,045	1	745,498	14,414,099	40,927	124,889	507,729	2,426,510	2,913,642
2014	4,502,329	1	758,235	16,735,737	38,477	125,925	518,157	2,476,340	2,990,906
2015	4,611,049	1	772,466	16,963,680	34,989	129,608	532,841	2,497,008	3,026,780
2016	4,695,182	1	790,532	19,047,738	35,946	134,800	546,676	2,528,353	3,073,080
2017	4,861,557	1	813,370	44,150,312	40,455	138,969	565,972	2,574,900	3,131,013
2018	5,010,358	1	835,629	38,659,216	40,000	144,000	580,000	2,623,760	3,238,980

Figure A.3: Regions of Chile

Regions Chile

Table A.10: Regional information

Region	Size KM^2	Population	Region Number	Poverty rate	% GDP Chile
Región de Arica y Parinacota.	16,873	226,068	15	8.4%	0.8%
Región de Tarapacá.	42,226	330,558	1	6.4%	2.5%
Región de Antofagasta.	126,049	607,534	2	5.1%	10.1%
Región de Atacama.	75,176	286,168	3	7.9%	2.6%
Región de Coquimbo.	40,580	757,586	4	11.9%	3.1%
Región de Valparaíso.	16,396	1,815,902	5	7.1%	9.1%
Región Metropolitana de Santiago.	15,403	7,112,808	13	5.4%	46.4%
Región del Libertador General Bernardo O'Higgins.	16,387	914,555	6	10.1%	4.8%
Región del Maule.	30,296	1,044,950	7	12.7%	3.4%
Región de Ñuble	13,179	480,609	16	16.1%	–
Región del Biobío.	23,890	1,556,805	8	12.3%	7.9%
Región de La Araucanía.	31,842	957,224	9	17.2%	2.8%
Región de Los Ríos.	18,430	384,837	14	12.1%	1.4%
Región de Los Lagos.	48,584	828,708	10	11.7%	3.3%
Región de Aysén del General Carlos Ibáñez del Campo.	108,494	103,158	11	4.6%	0.6%
Región de Magallanes y la Antártica Chilena	132,291	166,533	12	2.1%	1.2%

Table A.11: Intergenerational mobility indicators by municipality in the Metropolitan region. "Santiago" refers to the municipality, not the city.

Region	N	β_r	α_r	p_{15}	p_{11}	p_{55}	r_{abs}	r_{per}
Santiago	4711	0.212	43.360	0.183	0.207	0.388	48.654	63.474
Cerrillos	2207	0.175	43.128	0.165	0.184	0.291	47.514	59.792
Cerro Navia	4743	0.140	42.220	0.096	0.216	0.240	45.720	55.520
Conchalí	3876	0.131	46.015	0.159	0.163	0.314	49.286	58.444
El Bosque	5855	0.171	40.689	0.132	0.248	0.258	44.954	56.893
Estación Central	3277	0.207	42.257	0.154	0.203	0.359	47.434	61.931
Huechuraba	2523	0.269	36.537	0.123	0.226	0.366	43.263	62.098
Independencia	1664	0.197	43.278	0.170	0.175	0.368	48.207	62.009
La Cisterna	2259	0.197	44.585	0.236	0.189	0.379	49.506	63.283
La Florida	11614	0.226	42.181	0.154	0.192	0.388	47.819	63.607
La Granja	4257	0.162	43.585	0.158	0.208	0.300	47.646	59.015
La Pintana	6475	0.141	38.496	0.090	0.264	0.173	42.021	51.889
La Reina	2460	0.313	43.944	0.268	0.158	0.582	51.778	73.713
Las Condes	5619	0.409	41.259	0.235	0.166	0.676	51.489	80.132
Lo Barnechea	2353	0.551	29.397	0.149	0.222	0.771	43.172	81.741
Lo Espejo	3789	0.144	41.046	0.091	0.237	0.211	44.641	54.709
Lo Prado	3061	0.180	41.495	0.135	0.188	0.277	45.991	58.578
Macul	3074	0.247	39.830	0.145	0.215	0.372	46.008	63.307
Maipú	17481	0.186	45.625	0.197	0.190	0.351	50.286	63.337
Ñuñoa	3655	0.264	44.424	0.234	0.227	0.467	51.024	69.502
Pedro Aguirre Cerda	3350	0.174	41.523	0.130	0.235	0.283	45.879	58.075
Peñalolén	7261	0.288	34.718	0.109	0.263	0.400	41.917	62.075
Providencia	1627	0.248	50.776	0.284	0.205	0.589	56.983	74.363
Pudahuel	7465	0.151	43.272	0.137	0.221	0.265	47.055	57.646
Quilicura	5469	0.184	43.138	0.114	0.170	0.293	47.728	60.580
Quinta Normal	3050	0.169	44.525	0.153	0.195	0.302	48.760	60.618
Recoleta	4729	0.162	43.364	0.126	0.214	0.267	47.416	58.763
Renca	4892	0.142	44.065	0.135	0.186	0.266	47.620	57.573
San Joaquín	2650	0.160	43.306	0.102	0.208	0.273	47.304	58.498
San Miguel	2111	0.247	42.157	0.184	0.197	0.420	48.328	65.607
San Ramón	2890	0.153	38.985	0.098	0.250	0.179	42.799	53.478
Vitacura	1570	0.271	56.960	NA	NA	0.721	63.740	82.724
Puente Alto	19612	0.189	41.116	0.145	0.220	0.292	45.852	59.113
Pirque	510	0.262	38.296	0.103	0.191	0.509	44.857	63.227
San José de Maipo	385	0.332	36.005	NA	NA	0.381	44.297	67.516
Colina	3551	0.223	38.426	0.098	0.227	0.350	43.990	59.571
Lampa	1593	0.186	39.806	0.081	0.244	0.343	44.466	57.513
Tiltil	630	0.138	46.137	0.195	0.195	0.326	49.599	59.292
San Bernardo	9491	0.189	39.324	0.102	0.250	0.286	44.049	57.277
Buin	2918	0.235	36.048	0.086	0.299	0.313	41.932	58.408
Calera de Tango	650	0.360	32.700	0.144	0.288	0.568	41.697	66.887
Paine	1970	0.184	38.631	0.092	0.235	0.311	43.235	56.125
Melipilla	3228	0.182	39.959	0.110	0.266	0.286	44.498	57.207
Alhué	172	0.244	46.174	NA	NA	NA	52.280	69.380
Curacaví	783	0.134	43.136	0.133	0.235	0.216	46.489	55.876
María Pinto	327	0.136	41.334	0.058	0.256	NA	44.740	54.277
San Pedro	196	0.150	43.680	NA	NA	NA	47.426	57.914
Talagante	2241	0.234	38.799	0.113	0.242	0.363	44.641	60.999
El Monte	827	0.213	38.600	0.072	0.258	0.268	43.930	58.854
Isla de Maipo	1028	0.154	39.982	0.125	0.253	0.330	43.832	54.612
Padre Hurtado	1494	0.126	43.783	0.105	0.145	0.270	46.939	55.775
Peñaflor	2381	0.178	41.758	0.144	0.187	0.319	46.214	58.689

A.5 More on non-linearities

One way to study non-linearities is to study the effect that a change on parental earnings has on child earnings for different quantiles of the child earnings distribution. One way to do it is using the Unconditional quantile regressions developed by Firpo, Fortin and Lemieux (2009). First, let's assume that:

$$y_i^c = h(y^p, \epsilon) \tag{A.1}$$

where ϵ is the unobservable component and $h(\cdot)$ is stricly monotonic in ϵ. We can define the unconditional partial effect as the small location shift in the distribution of y^p on a distributional statistic $v(F_y)$. We can write this as:

$$\alpha(v) = \int \frac{dE[RIF(Y^c, v)|Y^p = y^p]}{dy^p} dF(y^p), \tag{A.2}$$

where $RIF(Y^c, v)$ is the recentered influence function. When the distributional statistic v is the rth quantile function, $q_\tau = inf\{q : F_{y^c}(q) \geq \tau\}$ and we can write the $RIF(Y^c, q_\tau)$ as:

$$RIF(Y^c, q_\tau) = q_\tau + \frac{\tau - 1\{y^c \leq q_t\}}{f_{Y^c}(q_\tau)}, \tag{A.3}$$

where $f_{Y^c}(q_\tau)$ is the density function of Y^c evaluated as q_τ. From Firpo, Fortin and Lemieux (2009) we know that the unconditional quantile partial effect (UQPE) can be expressed as the weighted average of the conditional quantile partial effects (CQPE):

$$UQPE(\tau) = E\left[\omega_\tau(Y^p) \cdot CQPE\left(\xi_\tau(y^p), y^p\right)\right] \tag{A.4}$$

where $CQPE(\tau, y^p) = \frac{\partial Q_\tau[h(y^p,\epsilon)|Y^p=y^p]}{\partial y^p}$, and define $CQPE(\tau) \equiv E\left[CQPE(\tau, y^p)\right]$, $\omega_\tau(Y^p) \equiv \frac{f_{Y^c|Y^p}(q_\tau|y^p)}{f_{Y^c}(q_\tau)}$ and $\xi_\tau : Y^p \to (0, 1)$ is given by:

$$\xi_\tau(y^p) = \{s : Q_s[Y^c|Y^p = y^p] = q_\tau\} = F_{Y^c|Y^p}(q_\tau|Y^p = y^p) \tag{A.5}$$

ξ_τ is a "matching function" that indicates when the unconditional quantile q_τ falls in the conditional distribution of Y^c given Y^p.

A.6 Why imputation-based IGE estimates may fail: the importance of administrative data use

The main challenge in studying intergenerational mobility is to link child and parental permanent income. In the absence of administrative records, studies on intergenerational mobility often rely on low-quality data, which do not allow to establish a parent-child link with adequate income information. This limitation is detrimental for the study of intergenerational mobility because it affects the credibility and precision of IGE estimates, which may cause to arrive at a misleading conclusion (Emran and Shilpi, 2019).

Two-Sample Two-Stage Least Squares (TSTSLS) estimator

With no access to administrative data, the most used methodology for IGE estimation is the Two-Sample Two-Stage Least Squares (TSTSLS). This estimator was originally introduced by Björklund and Jäntti (1997) for IGE estimation in a setting with missing parental income, and it has been used in several empirical studies (e.g., Aaronson and Mazumder, 2008; Gong et al., 2012; Olivetti and Paserman, 2015; Piraino, 2015). The TSTSLS estimator uses retrospective information on parents' socioeconomic background along with a sample of "pseudo"-parents to impute parental income through a Mincer's equation. Since background information of this type is more likely to be available in survey datasets (or historical censuses), the TSTSLS methodology has allowed the IGE estimation for a significantly larger number of countries and historical periods, especially in developing nations (Narayan et al., 2018; Brunori et al., 2020). The standard empirical specification for estimating intergenerational income mobility is given by the following equation:

$$y_i^c = \alpha + \beta y_i^p + \epsilon_i, \tag{A.6}$$

where y_i^c is the logarithm of a child's permanent individual income and y_i^p is the logarithm of her parent's permanent individual income.[1] The coefficient β is generally called "intergenerational elasticity" (IGE) and forms the basis for comparisons of intergenerational income mobility across countries. Among the existing IGE estimates in the literature, virtually all of those for developing countries are obtained through the TSTSLS methodology proposed by Björklund and Jäntti (1997). This estimation procedure is based on two samples. The main sample contains information on individual incomes and recall socioeconomic information about their parents. The auxiliary sample is typically derived from an earlier survey of the same population where individuals (pseudo-parents) report their income as well as socioeconomic information such as that recalled by respondents in the main sample. The estimation then proceeds in two steps. First, the auxiliary sample is used to estimate a Mincer's equation by OLS:

$$y_{it}^{sp} = \omega z_i + v_{it}, \tag{A.7}$$

where y_{it}^{sp} is the income of pseudo-parents for child i at time t, z_i is a vector of time invariant characteristics, and v_{it} is the residual component of (A.7)). Then, we estimate y_{it}^p as $\hat{y}_{it}^p = \omega \cdot z_i t^p$, that is, we impute the income of unseen parents. With this, we adjust the following relationship between y^c and $\hat{y}_i^p$

$$y_{it}^c = \alpha + \beta^{TSTSLS} \hat{y}_i^p + \psi_i \tag{A.8}$$

where β^{TSTSLS} is the elasticity between the imputed parental income and child income, while ψ_i is an error term. We estimate β^{TSTSLS} as:

$$\hat{\beta}^{TSTSLS} = \frac{\hat{cov}\left(y^c, \hat{y}^p\right)}{\hat{var}\left(\hat{y}^p\right)} \tag{A.9}$$

There are many ways to create $\hat{y}^p$. Let us call $\Omega = \{\hat{y}^p(j)\}_{j=1}^{\infty}$ the set of predictions

[1]Solon (1992) discuss that (1) could not be the true income model. In particular, there could be omitted variables that affect child earnings that are correlated with parental income. For instance, parents' education. Following Chetty et al. (2014), Corak (2018) and Accarci et al. (2020), we assume that (1) is the correct model. That is, there are not omitted variables problems. In addition, we also ignore any measurement error of the children's and parents' income.

for y^p. Each element $\hat{y}^p(j)$ of Ω defines a different parameter, call it $\beta(j)^{TSTSLS}$ with its respective TSTSLS estimator as $\hat{\beta}(j)^{TSTSLS}$. Significantly, we can stablish a relationship between $plim\hat{\beta}$ and $plim\hat{\beta}^{TSTSLS}(j)$, with

$$\hat{\beta} = \frac{\hat{cov}(y^c, y^p}{\hat{var}\,(y^p)} \tag{A.10}$$

Proposition 14. *$plim\hat{\beta}^{TSTSLS}(j)$ could be theoretically be higher or lower than $plim\hat{\beta}$.*

Proof. We know that

$$\begin{aligned} plim\hat{\beta} &= \tfrac{cov(y^c,y^p}{var(y^p)} = \tfrac{cov(y^c,y^p+\hat{y}^p(j)-\hat{y}^p(j)}{var(y^p)} \\ &= \tfrac{cov(y^c,y^p-\hat{y}^p(j)}{var(y^p)} + \tfrac{cov(y^c,\hat{y}^p(j)}{var(y^p)} \end{aligned}$$

Define $(y^p - \hat{y}^p(j)) \equiv \varphi(j)$

$$= \frac{cov(y^c, \varphi(j)}{var\,(y^p)} + \frac{var\,(\hat{y}^p)}{var\,(y^p)} \frac{cov(y^c, \hat{y}^p(j)}{var\,(\hat{y}^p)}$$

Define $\frac{cov(y^c,\varphi(j)}{var(y^p)} \equiv \eta(j)$ and $\frac{var(\hat{y}^p)}{var(y^p)} \equiv \kappa(j)$ also notice that $plim\hat{\beta}^{TSTSLS}(j) = \frac{cov(y^c,\hat{y}^p(j)}{var(\hat{y}^p)}$ Then, we have that

$$plim\hat{\beta}^{TSTSLS}(j) = \frac{plim\hat{\beta} - \eta(j)}{\kappa(j)} \tag{A.11}$$

□

Now, if we assume that $E(\epsilon_i|y_i^p) = 0$. Then, $plim\hat{\beta}^{TSTSLS}(j) = \frac{\beta-\eta(j)}{\kappa(j)}$

Let us call $\kappa(j)$ as the lack of variance bias and $\eta(j)$ as the projection bias associated to the projection $\hat{y}^p(j)$. Our administrative records not only allow us to estimate $\hat{\beta}$, but also allow us to mimic a TSTSLS estimation setting for different $y^p(j)$ to measure the magnitudes of $kappa(j)$ and $\eta(j)$. To evaluate the importance of the use of administrative data for IGE estimation, we proceed by simulating the followingb exercise:

Simulated exersice 1:

i From the main sample, we keep only children and their parents, and obtain β from equation (A.6) by OLS.

ii We take a random subsample Σ of 50,000 parents' and children's information from the main sample. We randomly divide this subsample into two sub-subsamples Σ_1 and Σ_2 of 25,000 observations. One sub-subsample (Σ_1) is used to estimate the following Mincer-type equation for pseudo-parents.

$$y_i^{sp} = \gamma' x_i' + v_i, \tag{A.12}$$

where x_i' is composed by age, age squared, occupational sector, education type, and type of contract. We estimate γ by OLS and we use Σ_2 parental information to compute a prediction for $\hat{y}^p$ call it $\hat{y}^p(1)$.

iii We compute $\hat{\beta}^{TSTSLS}(1)$ by regressing y^c on $\hat{y}^p(1)$ from Σ_2

iv We repeat ii)-iii) 1,000 times.

Table 14 shows the distribution of this simulated procedure. As we can see, $\hat{\beta}^{TSTSLS}(1)$ is much larger than $\hat{\beta}$. $\hat{\beta}^{TSTSLS}(1)$ is closer to what Nunez and Miranda (2010, 2011) previously estimate for Chile. They find that IGE is in the ranges of 0.5-0.6.

Table A.12: Results from simulated exercise 1

Coefficient	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\hat{\beta}$	0.287	0.287	0.287	0.287	0.287	0.287
$\hat{\beta}^{TSTSLS}(1)$	0.448	0.479	0.488	0.488	0.498	0.538
$\eta(1)$	0.142	0.155	0.159	0.159	0.162	0.180
$\kappa(1)$	0.238	0.258	0.263	0.263	0.267	0.286
$\frac{\left(\hat{\beta}-\eta(1)\right)}{\kappa(1)}$	0.413	0.473	0.489	0.489	0.504	0.559

As we can see, there is overestimation driven by $\kappa(1)$. The lower value of $\frac{var(\hat{y}^p(1))}{var(y^p)}$ has been deeply discussed in the statistical literature on imputation of missing data (Rubin and Little, 2019, Rubin, 2004; Rubin, 1996). Note that the setting of the TSTSLS estimation is a problem of missing data, where the unseen parental incomes are imputed using regression imputation. However, even under a correctly specified model, regression

imputation does not properly reflect the uncertainty of the missing data. The issue is that the imputed missing parental incomes from the regression model do not include any residual term, not providing enough uncertainty about the missing data. To overcome this problem, one possibility is to estimate the unseen parental incomes by using stochastic regression imputation. Specifically, to correct the lack of an error term in regression imputation, we can introduce error by adding a noise with zero mean and estimated regression variance to the regression imputation, that is, the predicted value from a regression plus a random residual value: $\hat{y}_i^p = \hat{\gamma}' x_i' + N(0, \hat{\sigma}_v^2)$, where $\hat{\sigma}_v^2$ is the estimated variance of the Mincer equation ($y_i^{sp} = ?' x_i' + v_i$). Significantly, to estimate each missing parent's income, the stochastic regression imputation can be repeated several times in the spirit of Multiple Imputation (Rubin, 1978). We now repeat exercise 1 using stochastic regression imputation. The results are in Table 15:

Table A.13: Results from simulated exercise 1 with additional variance

Coefficient	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\hat{\beta}$	0.287	0.287	0.287	0.287	0.287	0.287
$\hat{\beta}^{TSTSLS}(1b)$	0.109	0.123	0.128	0.128	0.132	0.149
$\eta(1b)$	0.135	0.154	0.159	0.159	0.164	0.191
$\kappa(1b)$	0.941	0.989	1	1	1.012	1.06
$\frac{\left(\hat{\beta}-\eta(1)\right)}{\kappa(1)}$	0.097	0.123	0.128	0.128	0.133	0.151

As can be seen in Table A.13, $\kappa(1)$ increases significantly as expected, attaining a median value of 1. However, $\hat{\beta}^{TSTSLS}(1b)$ is a lower bound of $\hat{\beta}$ because the prediction bias remains. We can improve $\hat{y}_p$ by adding an additional predictor to the Mincer's equation: a measure of the parental earning ability. We can estimate that value because of the panel structure on our data. To do this, we use our administrative dataset from 2003 to 2019 to estimate a panel Mincer equation with fixed effects using our main parents sample:

$$y_{it}^{sp} = \alpha_i + \omega z_{it} + \psi_{it} \tag{A.13}$$

where α_i is a parental fixed effect, z_{it} is a vector of time variant observables, and ψ_{it} is a white noise. After adjusting this model by OLS, we recover the estimated fixed effect associated to each parent to be used as measure of the parental earning ability. With this, we repeat our exercise by estimating $\hat{y}^p(2)$. This prediction includes the estimated fixed effect as predictor. We also can estimate $\hat{y}^p(2b)$ which is $\hat{y}^p(2)$ but after correcting for the lack of variance.

Table A.14: Results from simulated exercise 1 with additional variance and more prediction

Coefficient	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\hat{\beta}$	0.287	0.287	0.287	0.287	0.287	0.287
$\hat{\beta}^{TSTSLS}(2)$	0.317	0.335	0.34	0.34	0.345	0.363
$\hat{\beta}^{TSTSLS}(2b)$	0.253	0.272	0.276	0.276	0.28	0.298
$\kappa(2)$	0.787	0.805	0.81	0.81	0.815	0.833
$\kappa(2b)$	0.972	0.993	1	1	1.006	1.033
$\eta(2)$	0	0.009	0.011	0.011	0.013	0.020
$\eta(2b)$	-0.003	0.008	0.011	0.011	0.014	0.023
$\frac{\hat{\beta}-\eta(2)}{\kappa(2)}$	0.327	0.338	0.34	0.34	0.343	0.355
$\frac{\hat{\beta}-\eta(2b)}{\kappa(2b)}$	0.262	0.273	0.276	0.276	0.279	0.293

From this exercise, we understand that the reasons that the traditional $\hat{\beta}^{TSTSLS}$ does not work to estimate β are insufficient prediction power for y_p and the lack of variance of that prediction. Typically, predictions for y_p cannot be improved by information on cross-sectional household surveys, and even in the case of having the right Mincer's equation, TSTSLS can be still biased for the lack of variance of the imputed parental earnings. Of course, the main challenge is to build a good model to impute y_p, especially in developing countries where earnings/income are usually determined by unobservable covariates such as social capital, non-cognitive skills, or neighborhood. Thus, this exercise shows the importance of using administrative information to estimate intergenerational mobility.[2]

[2]As Chetty et al. (2014) pointed out, income/earnings intergeneration elasticity is highly sensitive

to the inclusion of parents with 0 earnings. We find the same for Chile, a result that is consistent with Corak and Heisz (1999), who also found that IGE is highly non-linear.

Appendix B

Appendix for Chapter 2

B.1 A stochastic model for income process

Following Díaz et al. (2020), we assume that time is continuous and there is a continuum of agents indexed by i. Workers are heterogeneous in their total income y_i. Total income is equal to

$$y_i \equiv w_i + x_i, \tag{B.1}$$

where w_i is the part of the income that does not depend on the level of previous or actual income (the additive part, e.g., labour income or social security income), and x_i is the piece of income that depends on the level of previous income or actual income (the multiplicative part, e.g., dividends, interest payments, real estate income). Assume that an agent needs some level of income $\bar{w}$ (threshold income) to buy assets that generate some income x_i. If $w_i < \bar{w}$ then $x_i = 0$.[1] Suppose that the dynamics of w is given by the following reduced-form model.[2]

[1] One might think that a retired agent has $w_i = 0$ but $x > 0$. However, w_i is not only labour income; it is all income that does not depend on the previous income. Assuming that a retired agent gets a constant flow of money from pension plans (such as RPP, RRSP, or social security) then this is an additive source of income instead of a multiplicative source. Besides, on average we can argue that a retired agent also needs a minimum income to start saving and we assume that on average is $\bar{w}$.

[2] See Champernowne (1953), Aoki and Nirei (2016), Kim (2015) and Gabaix et al. (2016). Those papers use models with this reduced form. In particular, Gabaix et al. show that this equation could be derived from a general equilibrium model with individual optimization.

$$\frac{dw}{dt} = \mu + \sigma\epsilon(t), \tag{B.2}$$

where μ is the drift term, σ is the amplitude term and $\epsilon(t)$ is white noise.[3] Because dw does not depend on w itself, neither μ nor σ depend on w. In the case of x, we can describe its dynamics using the following stochastic process:

$$\frac{dx}{dt} = \mu(x) + \sigma(x)\epsilon(t). \tag{B.3}$$

Here, $\mu(x)$ and $\sigma(x)$ depend on x. Thus, y has a reflecting barrier at $\bar{w}$ but for incomes that are greater than $\bar{w}$.[4] To solve those equations, we need to find a stationary distribution, one method to do this is using the Kolmogorov forward equation, which is given by:

$$\frac{\partial f(m,t)}{\partial t} = -\mu(m)\frac{\partial f(m,t)}{\partial m} + \frac{\sigma(m)^2}{2}\frac{\partial^2 f(m,t)}{\partial m^2}, \tag{B.4}$$

where $f(m,t)$ is the probability density function implicit in the stochastic process described by (B.4).

In order to find a parametric solution, we solve for the stationary distribution, that is, we impose $\frac{\partial f(m,t)}{\partial t} = 0$. Call this pdf $\bar{f}(m)$. Solving equation (B.2) we have the following stationary distribution:[5]

$$\bar{f}(w) = \lambda e^{-\lambda w}. \tag{B.5}$$

That is, the stationary distribution of w is an exponential distribution with parameter λ.

[3]Some literature estimates a reduced-form labour income process using equation (B.2). For example, Hearthcote et al. (2010) and Meghir and Pistaferri (2011).

[4]To ensure the existence of a stationary distribution, we need to add some "stabilizing force" (Gabaix, 2009). In this case, we add a reflecting barrier at $\bar{w}$; that is, the income cannot be below $\bar{w}$ after the level $\bar{w}$ is achieved.

[5]In order to build the imputation method, it is necessary to work in terms of levels. However, one may think that equation (B.2) should be in terms of logs instead of levels, but there are two arguments to address this with this. First, a reflecting barrier around 0 could be added to address the fact that $\epsilon(t)$ could be highly negative. Second, the stationary distribution is an exponential which works only with positive numbers.

By equation (B.3), assuming that $\mu(x) = \mu x$ and $\sigma(x) = \sigma\sqrt{x}$ we have:

$$\bar{f}(x) = \eta_x \frac{\bar{w}^{\eta_x}}{x^{\eta_x+1}}. \tag{B.6}$$

That is, the stationary distribution of x is a Pareto distribution with parameter η_x and $\bar{w}$ threshold parameter.

Now, following Gabaix (2009), if w and x are independent power law processes with η_w and η_x as exponents.[6] then, the process composed by the sum of w and x follows:

$$\eta_y \equiv \eta_{w+x} = min(\eta_w, \eta_x), \tag{B.7}$$

where η_y is equal to the power law exponent of y. When one combines two independent power law processes, the fattest (the one with the smallest exponent) power law dominates.[7] In simple words, in the tail of the distribution, the unequal process dominates. This result is also derived by Clemens, Gottlieb, Hémous and Olsen (2017); they use a different framework to arrive at the conclusion that the wage distribution of doctors is dominated by the shape parameter of the distribution of practitioners which is a distribution that generates higher inequality.[8] In our case, where w is distributed as an exponential, we can say that $\eta_w = \infty$ and given that x is Pareto distributed, then $\eta_x < \infty$. That is, $\eta_{w+x} = \eta_x$. Given the assumption that the income generated through assets (multiplicative income) is more unequal to the additive income (labour income and social security), the distribution in the tail is dominated by the shape of the (x) process.

[6] A power law is a relation of the type $Y = kX^{\alpha}$. A distribution that satisfies at least in the upper tail $P(y > x) = k \cdot x^{-\eta}$ where η is the power law exponent and k is a constant.

[7] Given that in the model x is different from 0 if $w \geq \bar{w}$,one may think that w and x are not independent. However, for the application of this property is the tail of the distributions that matters. Thus, conditional on $w \geq \bar{w}$ it is not crazy to think that the return of the income generated though assets are independent of the labour income.

[8] In the context of Clemens et al.ś work:"suppose that there are two cities with physicians and one city has more inequality than the other. If we observe the aggregate income distribution, we see the shape of the city with the highest inequality".

With this in mind, the cumulative distribution function (cdf) of total income is:[9]

$$F(y) = \begin{cases} 1 - e^{-\lambda y} & y \leq \bar{w} \\ 1 - e^{-\lambda \bar{w}} + e^{-\lambda \bar{w}} \left(1 - \left(\frac{\bar{w}}{y}\right)^{\eta_x}\right) & y > \bar{w} \end{cases}. \tag{B.8}$$

This income distribution depends on three key parameters: λ, $\bar{w}$ and η_x. The first of those parameters is the exponential parameter which in this context gives the shape of the additive income process. The second parameter $\bar{w}$ is the core parameter of interest; it represents the minimum income that an individual will require to start generating income through his assets. This means that if the market income of some agent is less than $\bar{w}$, this agent does not hold any retained earnings. Thus, the imputation procedure starts to assign values different from 0 only if $y_i \geq \bar{w}$. Finally, η_x is the shape parameter of the Pareto tail; a lower parameter means an unequal economy.

B.1.1 Effect of the inclusion of retained earnings in the stochastic income model

In order to have ownership over retained earnings, it is required to own a part of a firm, that is, it is necessary to hold an asset (for instance, the entire firm in the case of sole-ownership firms or some amount of stock in a publicly traded firm). Those assets could be bought today or have been bought in the past, that is, the income generated through retained earnings depends on the current or previous income. Thus, the income generated through retained earnings is a multiplicative part of the income. In particular, we assume that retained earnings has the same structure as x. That is, retained earnings are only different from 0 if $w_i < \bar{w}$. Thus, the inclusion of retained earnings does not change $\bar{w}$.

Now, define z_i as the income generated through retained earnings; then, we have the following identity for the income process:

$$y_i^* \equiv w_i + x_i + z_i, \tag{B.9}$$

[9]Notice that a normalization restriction $\int_0^\infty f(y)dy = 1$ was imposed.

where y_i^* is the total income including retained earnings, w_i and x_i are defined as before. Because x_i and z_i is income that comes from assets, we define $h_i \equiv x_i + z_i$. Given that h_i is a multiplicative part, it has the same structure as equation $(B.3)$

$$\frac{dh}{dt} = \mu(h) + \sigma(h)\epsilon(t). \tag{B.10}$$

As we can see, $\mu(h)$ and $\sigma(h)$ depend on h. Then, using equation (B.4) in (B.10) and solving for the stationary distribution, we see that the stationary distribution for h is a Pareto process with η_h parameter and $\bar{w}$ as a threshold parameter. Then, the stationary distribution for y^* is

$$F(y^*) = \begin{cases} 1 - e^{-\lambda y} & y^* \leq \bar{w} \\ 1 - e^{-\lambda \bar{w}} + e^{-\lambda \bar{w}} \left(1 - \left(\frac{\bar{w}}{y}\right)^{\eta_{w+h}}\right) & y^* > \bar{w} \end{cases}. \tag{B.11}$$

If retained earnings are more concentrated than other assets (e.g., housing, bonds) it is expected that $\eta_h \leq \eta_x$, which implies that the tail of the income distribution is driven by the shape of the distribution of retained earnings. However, it is quite difficult to determine the exact distribution of h; we need administrative data about income and ownership of firms for each individual, data which are quite difficult to obtain for longer periods of time and different countries. For this reason, in the next section, we develop a procedure to estimate η_h without knowing the whole distribution.

B.2 Standard error estimation

B.2.1 Estimation of standard errors

To provide confidence intervals we need to estimate standard errors for the parameters. In this context, there are two possibilities, the first one is to compute the asymptotic distribution of the estimators, and the second procedure is to use a bootstrap procedure. As Cowell and Van Kerm (2015) suggested we use the bootstrap procedure.[10]

[10]Hansen (2000) in the context of Threshold autoregressive models (TAR) found that the distribution of the threshold estimate is non-standard. So the computation of an asymptotic distribution is far from

Bootstrap and threshold models

It is known that the standard bootstrap under mild conditions can provide an estimate of the exact variance of the estimator.[11] However, following Kapetanios (2009) in the case of the estimator of the threshold parameter, there are doubts of the consistency of the bootstrap.[12] In particular, an essential condition for bootstrap validity for an estimator is that the mapping between the joint distribution function of the sample and the estimator is continuous. It is not clear whether this continuity assumption is satisfied in a threshold model.[13] Moreover, as Yu (2014) showed, the non-parametric bootstrap fails in discontinuous threshold regression, demonstrating the inconsistency of the non-parametric bootstrap for inference on the threshold point.[14]

In this context, Davidson and Flachaire (2007) show that a semi-parametric model helps to address inference. Constructing bootstrap inference using a semi-parametric model enhances the precision of confidence intervals and tests. The idea is to build bootstrap samples by resampling from the survey data, that is by drawing observations with replacement from the bottom of the sample with probability p^* and taking observations simulated from the Pareto distribution with probability $1 - p^*$. Note that in this procedure, point estimates are still calculated on the basis of the full non-parametric sample (both in the full sample and in the resamples) and the semi-parametric bootstrap does not involve re-estimation of the parameter in each bootstrap sample.

B.3 Proof of proposition 2

Proof. we have:

the scope of this paper.

[11] Under the assumption of asymptotic pivotalness (independence of the asymptotic distribution from nuisance parameters), the bootstrap estimator may converge more quickly to the true distribution compared to the asymptotic approximation

[12] Diebold and Chen (1996) provide simulation evidence without theory that the parametric bootstrap works well for structural change tests applied to AR(1) processes.

[13] This is true either when a fixed threshold is assumed as in Chan (1993) or when a small threshold assumption is made, such as Assumption 3, which is used by Caner and Hansen (2004).

[14] The main reason for the non-parametric bootstrap failure is the discreteness generated from the bootstrap sampling on the local data around $\bar{w}$. To break such discreteness, we can smooth the data in the neighborhood of $q = \bar{w}$. Such a procedure is termed as the smoothed bootstrap by Efron (1979)

$$Pr(h \leq u) = Pr(x + \psi(x) \leq u)$$

Define $r(k) = h^{-1}(k)$, because of assumption 1 and 2 this function is unique, then we have:

$$\begin{aligned} Pr(h \leq u | x > \bar{w}) &= Pr\left(x \leq r(u) | x > \bar{w}\right) \\ &= \int_{\bar{w}}^{r(u)} f(x)dx \end{aligned}$$

Then, taking derivatives with respect to u and using the Leibniz rule we will have:

$$f_h(u) = r'(u) f_x\left(r(u)\right)$$

where $f_h(\cdot)$ is the conditional pdf of h and $f_x(\cdot)$ is the conditional pdf of x. Now, using assumption 3 we have:

$$\frac{\eta_h \bar{w}^{\eta_h}}{u^{1+\eta_h}} = r(u)' \frac{\eta_x \bar{w}^{\eta_x}}{r(u)^{1+\eta_x}}$$

Using some algebra we get:

$$\frac{r'(u)}{r(u)} = \frac{\eta_h}{\eta_x} \frac{\bar{w}^{\eta_h}}{\bar{w}^{\eta_x}} \frac{r(u)^{\eta_x}}{u^{1+\eta_h}}$$

Try the following educated guess $r(u) = Au^{\frac{\eta_h}{\eta_x}}$ and integrating both sides, we have:

$$\log(r(u)) = \frac{\eta_h}{\eta_x}\frac{\bar{w}^{\eta_h}}{\bar{w}^{\eta_x}} \cdot A \cdot \log(u) + C$$

Given the educated guess we have that $A = \frac{\bar{w}^{\eta_x}}{\bar{w}^{\eta_h}}$ and $C = A$. Then we have that:

$$r(u) = \frac{\bar{w}^{\eta_x}}{\bar{w}^{\eta_h}} u^{\frac{\eta_h}{\eta_x}}$$

then $r^{-1}(x) = h$ is equal to

$$hequivx + \psi^*(x) = x^{\frac{\eta_x}{\eta_h}} \frac{\bar{w}}{\bar{w}^{\frac{\eta_x}{\eta_h}}}$$

Then

$$\psi^*(x) = x^{\frac{\eta_x}{\eta_h}} \frac{\bar{w}}{\bar{w}^{\frac{\eta_x}{\eta_h}}} - x$$

□

Appendix C

Appendix for chapter 3

C.1 Tables for empirical analysis

Table C.1: Data for Canada. μ in current Canadian Dollars

Year	μ	G^*	Gini 1%	z	$G_u r$	p	$G_n r$
1999	41966	0.4497	0.1521	0.0241	0.4630	0.0010	0.4615
2005	52466	0.4753	0.1971	0.1025	0.5291	0.0025	0.5247
2012	65853	0.4816	0.2549	0.0488	0.5069	0.0010	0.5051
2016	73760	0.4846	0.3103	0.0080	0.4888	0.0002	0.4885

Table C.2: Data for Chile. μ in current Chilean Pesos

Year	μ	G^*	Gini 1%	z	$G_u r$	p	$G_n r$
2003	301528	0.5679	0.2987	0.0395	0.5849	0.0001	0.5846
2006	328722	0.5524	0.2771	0.0859	0.5908	0.0003	0.5897
2009	419562	0.5515	0.2574	0.1108	0.6012	0.0015	0.5979
2011	436479	0.5391	0.2320	0.1134	0.5914	0.0009	0.5891
2013	414181	0.5098	0.2255	0.1086	0.5630	0.0011	0.5605
2015	473438	0.4970	0.2196	0.1604	0.5777	0.0006	0.5746
2017	526054	0.5022	0.2674	0.1281	0.5660	0.0007	0.5634