PARIS SCHOOL OF ECONOMICS
ÉCOLE D'ÉCONOMIE DE PARIS

Master Analysis and Policy in Economics

Wealth inequality in Europe and in the United States: estimations from surveys, national accounts and wealth rankings

Thomas Blanchet

June 2016

Supervisor
Thomas Piketty

Referee
Xavier d'Haultfœuille

Abstract

This dissertation studies the distribution of wealth in the United States and six European countries: Austria, France, Germany, Italy, Portugal and Spain. To estimate the top tail of the distribution, I combine survey data with journalist rankings of top wealth holders. I also adjust the distribution for consistency with macroeconomic aggregates. I suggest a method which, unlike previous approaches, does not rely on the Pareto distribution or any other parametric assumption. Instead, I use the properties of order statistics to estimate the quantile function nonparametrically. In the United States, I find that the top 1% owns 40% of the wealth, and the top 0.1% owns 18%. In Europe, wealth inequality is much lower overall, but there are large differences between countries.

JEL codes: D31, C14, C51

Keywords: wealth distribution, inequality, nonparametric estimation, order statistics, Pareto distribution

Acknowledgments

First of all, I wish to thank Thomas Piketty for his support, his trust and his guidance as supervisor of this dissertation, and Xavier d'Haultfœuille for agreeing to be the referee.

I would like to thank Facundo Alvaredo for his availability and help throughout the realization of this project. This also work benefited from the insightful advice of Bertrand Garbinti and Jonathan Goupille. My thanks also go to Wiemer Salverda for providing data for wealth in the Netherlands.

Finally, this dissertation might also never have seen the light of day if it weren't for Muriel Roger and Frédérique Savignac, who initially introduced to the data used in this project, the issue of wealth, and more generally to the world of research.

Contents

List of figures

List of tables

List of abbreviations

BLUE best linear unbiased estimator

CDF cumulative distribution function

ECB European Central Bank

FGLS feasible generalized least squares

GLS generalized least squares

HFCS Household Finance and Consumption Survey

iid independent and identically distributed

IRA individual retirement accounts

MLE maximum likelihood estimation

OLS ordinary least squares

NPISH non-profit institutions serving households

PDF probability density function

PIM perpetual inventory method

SCF Survey of Consumer Finances

SNA system of national accounts

List of notations

$X_1, X_2, \ldots, X_n$ unordered random variables

$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ ordered random variables (order statistics)

f, F, Q PDF, CDF and quantile function

$f_{(r)}$, $F_{(r)}$, $Q_{(r)}$ PDF, CDF and quantile function of $X_{(r)}$

$\mathbb{P}\{A\}$ probability of event A

$\mathbb{E}[X]$ expected value of X

$\mathrm{Var}(X)$ variance of X

$\mathrm{Cov}(X, Y)$ covariance of X and Y

$\mathrm{Med}(X)$ median value of X

$\mu_{(r)}$ expected value of $X_{(r)}$

$\sigma^2_{(r)}$ variance of $X_{(r)}$

$\sigma_{(r)(s)}$ covariance of $X_{(r)}$ and $X_{(s)}$

$m_{(r)}$ median value of $X_{(r)}$

$\mathrm{span(S)}$ linear span of the set of vectors S

$(x)_+$ equivalent to $\max(0, x)$

$\mathrm{B}(a, b)$ Beta function

$I_p(a, b)$ Incomplete regularized Beta function: $I_p(a, b) = \frac{1}{\mathrm{B}(a,b)} \int_0^p t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$

$\lfloor x \rfloor$ biggest integer $\leq x$ (floor function)

Introduction

Wealth, as opposed to income, is not taxed in most countries. That deprives economists of an important data source, which has proved extremely useful in the study of income inequality. When related tax information is available, there are ways around the problem: the estate multiplier method uses inheritance tax data (eg. Kopczuk and Saez, 2004), and the capitalization method uses tax data on capital income (eg. Saez and Zucman, 2016).

Another solution is to use surveys, in which a randomly selected sample of households are asked to tell us about their wealth. The limitation of these sources are well-understood, which is why recent work on inequality (eg. Atkinson and Piketty, 2010) has privileged tax data. Survey data can be subject to misreporting or nonresponse bias. By nature, they are also ill-suited to deal with extreme values. Samples typically include a few thousands observations, which means that the famed top 1% will be represented only by a few dozen data points. Estimates of top shares (or any other statistic that relies on extreme parts of the distribution) may therefore be unreliable. Despite these limitations, surveys have been a useful source to assess income inequalities when no better option is available (eg. Milanovic, 2002).

But when it comes to wealth, survey estimates are even more likely to go wrong. First, because wealth is much more unequally shared than income, so that all the problems associated with the estimation of the shares of small groups of top earners are exacerbated. Saez and Zucman (2016) have shown for example that the increase in wealth inequality in the United States since the 1980s has happened not within the top 1%, but the top 0.1%. Such a tiny fraction of the population cannot, in general, reasonably be captured by a survey. Some countries alleviate those concerns by strongly oversampling the wealthiest. This is desirable, but it can actually create a second problem: since not all countries do it equally well, it hinders comparability. Some countries might report higher top shares just because they capture the top tail of the distribution more accurately than others.

The goal of this dissertation is to use surveys to estimate the distribution of wealth, but improve them using external sources of information. I use the national accounts to correct for systematic misreporting of assets, and I rely on journalist rankings of top wealth holders such as *Forbes* to estimate the top tail.

Until fairly recently, many European countries lacked a proper wealth survey. That has changed

around 2010 with the first wave of the Household Finance and Consumption Survey (HFCS), headed by the European Central Bank (ECB), which established a regular survey of household wealth. I use it alongside its American counterpart, the Survey of Consumer Finances (SCF).

Similar exercises have been done previously in the literature. The annual *Global Wealth Report* from Credit Suisse, which builds upon the work of Davies, Sandström, et al. (2009), also uses similar sources to estimate the worldwide distribution of wealth, although the precise methodology is not made explicit. Vermeulen (2014) combined surveys with the *Forbes* list of billionaires worldwide under the assumption of a Pareto shaped tail to correct raw survey estimates of wealth inequality. Bach, Thiemann, and Zucco (2015) used the same method with more precise rankings published by national magazines. Vermeulen (2016) improved his initial estimates also using data from the national accounts. Using survey data only, Eckerstorfer et al. (2015) estimated a Pareto tail for the wealth distribution in Austria, and then tested their method against a journalist ranking of Austria's wealthiest.

Until now, all these corrections have assumed that the tail of the wealth distribution follows a Pareto law. I innovate by providing a new, rigorous framework which does not rely on any parametric assumption. The gist of the method is to characterize the distribution of wealth by the graph of its quantile function on a logarithmic scale. In the Pareto case, that graph is simply a straight line. In the general case, it is a curve that can be estimated nonparametrically by a quantile regression of order statistics against some well-defined transformation of their rank.

I use the method jointly with the rescaling of assets to match the national accounts, thus providing new estimates of the wealth distribution that should correctly capture the top tail, and be consistent with the national accounts. I apply it to the United States and six European countries for which sufficiently good data was available: Austria, France, Germany, Italy, Portugal and Spain. In Europe, I give estimates for years around 2010, and in the United States I can go back to 1989.

Main findings

In line with the literature, I find that survey data correction increases significantly estimates of inequality, in particular when the survey does not strongly oversample the wealthy.

I find wealth inequality in 2010 to be much stronger in the United States than in Europe: in the six European countries studied, the share of the top 1% is at 22%, and the share of the top 0.1% at 10%. In contrast, the same shares for the United States are 40% and 18%.

The situation in Europe reveals important disparities. Austria, despite important statistical uncertainty surrounding the results, appears about as unequal as the United States, followed by Germany. Spain, on the other hand, has the most equal repartition of wealth, with a top 1% share of 15%, and a top 0.1% share of 5%.

In the United States, I get results consistent with the finding of Saez and Zucman (2016), who used the capitalization method, that wealth inequality has increased significantly over the past decades: between 1989 and 2013, the share of the top 1% went from 31% to 41%, and the share of the top 0.1% went from 12% to 18%.[1]

Organization

This dissertation is divided in three chapters. The first one compares household surveys with national accounts, both in terms of concepts and numbers. I adjust survey data to match the national accounts aggregates, and see how it affects the distribution of wealth.

The second one is purely theoretical. It introduces an important tool — order statistics — to look into the general problem of estimating the tail of a distribution. I start by applying this tool to the Pareto distribution to study the properties of estimators currently used in the literature, and then move to the nonparametric case.

The third chapter applies the method to the actual data. I present the different journalist rankings of top wealth holders that we use, and develop some methodological points that are specific to our setting. Finally, I present the results we obtain with the new method.

[1] Those values are actually very close to what the authors find themselves after adjusting the SCF to the national accounts and adding the *Forbes 400* to the sample. This simple approach works well enough in the United States because the SCF implements strong oversampling while it excludes the *Forbes 400* for confidentiality reasons. But it wouldn't be enough in other countries.

Chapter 1

From national accounts to household surveys

The first chapter of this dissertation looks at household wealth from two perspectives: from the macroeconomic side, using the national accounts, and from the microeconomic side, using household surveys.

National accounts started fairly recently to record the stock of assets and liabilities into national balance sheets (Piketty and Zucman, 2014). Those estimates constitute a good point of reference. They are the product of important efforts poured into the harmonization of concepts to make figures comparable between countries and over time. But, like all national accounts data, they are solely concerned with aggregates.

Household surveys, on the other hand, do carry information on the distribution of wealth, so they can be a useful addition to the national accounts. But to make a coherent whole out of these two sources present some difficulties.

Section 1.1 will deal with the conceptual and methodological differences. Importantly, this is where I give the definition of wealth that will be used throughout this dissertation. Section 1.2 will compare estimates of aggregate wealth between both sources, and section 1.3 will make a simple adjustment to the surveys to make them consistent with the national accounts, and see how this affects the distribution of wealth.

1.1 Sources, methodology and concepts

1.1.1 Definition of wealth

There is no universally agreed definition of wealth. Different concepts coexist, each of which can have its own purpose. Even within the study of inequality, different notions may have some relevance: in the end, the structure of inequality will naturally depend on the definition that is adopted. Empirically, we are also constrained by the availability of the data: some assets may be

too difficult to estimate and thus may be excluded for purely practical reasons. Finally, a good definition should be consistent across time and countries.

National accounts

The system of national accounts (SNA) defines assets as entities "over which ownership rights are enforced by institutional units" and "from which economic benefits may be derived by their owners" (United Nations, 1993, § 10.2).[1] As noted by Piketty and Zucman (2014), that definition excludes human capital, over which property rights cannot be enforced.[2]

The first main item on countries' balance sheets are *financial* assets, which "arise out of contractual relationships between institutional units" (United Nations, 1993, § 10.4). They essentially include deposits (currency, sights and savings accounts), bonds and shares, life insurance and private pension funds. The SNA does not record pay-as-you-go, unfunded pensions as assets: doing so would also call for the inclusion of the net present value of all future taxes and benefits, whose computation raises all sorts of difficulties.

The other assets are called *non-financial*, or *real*. The SNA distinguishes *non-produced* assets (land, essentially) from *produced* assets. They mostly consist in dwellings, but also include other types of buildings, machinery and cultivated resources. More recently, the SNA started including non-tangible assets (intellectual property), but coverage is still erratic, and it essentially pertains to the non-profit sector, so we will ignore it (see section 1.1.2). More importantly, real assets exclude consumer durables: spending in furniture or jewelry by households is considered consumption, not investment, even though it yields a flow of benefits over time. This is done for a purely practical reason: if we consider such goods as wealth, then we must also consider the flow of capital income they produce, which would be difficult to estimate given the lack of rental market for them. Some countries try to do it anyway, but again coverage is erratic so we will ignore them.

Finally, there are *liabilities*, which are financial by nature. The SNA solely distinguishes short-term from long-term loans.[3] In the end, *net wealth* is defined as the sum of all assets, minus the liabilities.

The situation is somewhat different in the United States. The Federal Reserve publishes a very detailed balance sheet, which does not exactly follow the SNA's guidelines. That is not generally a problem because it is possible to rearrange items to match the SNA's definition, but it is

[1] We will rely on the 1993 version of the SNA. In 2008, new guidelines were jointly introduced by the UN, the OECD, the World Bank, the IMF and the European Commission. However, not all countries have moved to the new standard yet, so that the 1993 still provides better country coverage. The differences between both standards are modest and do not impact the results.

[2] Moreover, to include it would require treating education and health spending as an investment; but because those are also services with a consumption value, a basic distinction upon which national accounts are built (consumption vs. investment) would collapse.

[3] A third category exists, "other accounts payable", but represents a very small fraction of total liabilities. It includes late payments and otherwise hard to identify liabilities.

something to be aware of. For example, contrary to the SNA, the United States financial accounts do record defined benefit pensions as assets. I will mention these differences whenever they are relevant.

Household surveys

We will be working with two surveys of household wealth: the SCF for the United States, and the HFCS for European countries. Those surveys do not provide a definition of wealth *per se*: they simply make an inventory of a household assets and liabilities. They do provide a variable for total wealth that is derived from the value of assets and liabilities that the household listed, but this is done solely for convenience.[4] The definition that stem from this variable does not, obviously, carry the same normative weight as the SNA's definition.

In this dissertation, we will use a definition of wealth that is mostly driven by a wish for consistency between macroeconomic and microeconomic data. That is not the case for the wealth variables included in the surveys. The results will therefore sensibly differ from previous studies, including in particular the official publications of the SCF (Bricker et al., 2012) and the HFCS (HFCN, 2013b). I build upon similar exercises that have been realized for the United States (Henriques and Hsu, 2014; Dettling et al., 2015) and Europe (Honkkila and Kavonius, 2013; Andreasch and Lindner, 2014).

Start with the real assets. The SNA's distinction between produced and non-produced assets is absent from the surveys, and understandably so: most homeowners would not be capable of separating the value of their dwelling from the value of the land on which it sits. Real assets in the surveys include consumer durables, which we can remove straightaway since they are usually absent from the national balance sheets. We are then left with real estate assets, and self-employment businesses.

The definition of self-employment businesses is one of the major discrepancies between macro and micro data (HFCN, 2013a, p. 93). The surveys define it as the value of the businesses owned by the household, and where at least one of its members is currently working. That category is absent from the SNA which only considers the legal form of companies. Regardless of the household's involvement in the business, ownership of entities involved in production is considered a financial asset and recorded as either quoted or unquoted equity. Unincorporated businesses are not registered as a separate entity in the national accounts, and so their assets and liabilities are directly attributed to the households. The surveys do provide information on the legal form of businesses, so that it is possible keep defining as real assets the ownership of unincorporated businesses, and recast as financial assets the ownership of corporations. This adjustment should remove most of the discrepancy, although it still doesn't account for the fact

[4] In fact, the SCF does not even provide such a variable in its raw form. But the Federal Reserve Board publishes the SAS programs that are used for the *Bulletin* articles associated with each release of the SCF. Those programs generate a variable for net wealth that researchers often use.

		SNA 93			US NA			HFCS			SCF		
financial assets	deposits	(+)	F.22	Transferable deposits	(+)	B.101 (10)	Deposits	(+)	DA2101	Deposits	(+)	liq	Transactions accounts
		(+)	F.29	Other deposits							(+)	cds	Certificates of deposit
	bonds, stocks and mutual funds	(+)	F.33	Securities	(+)	B.101 (15)	Debt securities	(+)	DA2104	Non self-employment business	(+)	stocks	Stocks
		(+)	F.51	Shares and other equity	(−)	B.101 (17)	Open market paper	(+)	DA2105	Shares, publicly traded	(+)	bonds	Bonds
		(+)	F.52	Mutual funds shares	(+)	B.101 (25)	Corporate equities	(+)	DA2106	Managed accounts	(+)	savbnd	Saving bonds
					(+)	B.101 (26)	Mutual fund shares	(+)	DA2102	Mutual funds	(+)	nmmf	Directly-held mutual funds
					(−)	L.117 (27)	IRAs	(+)	DA2103	Bonds	(+)	othma	Other managed assets
								(+)		Corporate self-employment business	(+)		Corporate self-employment business
	private pensions and life insurance	(+)	F.611	Net equity in life insurance reserves	(+)	B.101 (27)	Life insurance reserves	(+)	DA2109	Voluntary pensions/Whole life insurance	(+)	cashli	Cash value of whole life insurance
		(+)	F.612	Net equity in pension funds	(+)	L.117 (26)	DC pensions				(+)	retqliq	Quasi-liquid retirement accounts
					(+)	L.117 (27)	IRAs						
					(+)	L.117 (28)	Annuities at LICs						
	real assets	(+)	N.111	Tangible fixed assets	(+)	B.101 (4)	Real estate	(+)	DA1110	Main residence	(+)	houses	Primary residence
		(+)	N.211	Land	(+)	B.101 (29)	Noncorporate business	(+)	DA1120	Other real estate	(+)	oresre	Other residential real estate
								(+)	DA1140	Self-employment business	(+)	bus	Business interests
								(−)		Corporate self-employment business	(−)		Corporate self-employment business
	liabilities	(+)	F.41	Short-term loans	(+)	B.101 (34)	Home mortgages	(+)	DL1000	Liabilities	(+)	debt	Debt
		(+)	F.42	Long-term loans	(+)	B.101 (35)	Consumer credit				(−)	odebt	Other debts

The name of some items have been abbreviated. For the United States balance sheet, “B.101 (10)” (for example) refers to the 10th line of the table B.101. SCF variable names refer to the output of the Federal Reserve Board’s SAS programs. There is no variable for corporate self-employment business, but it can be calculated using HD040x and HD080x (for the HFCS), and X3119, X3129, X3219, X3229, X3416 X3420 and X3428 (for the SCF).

Table 1.1: Correspondence table of wealth items

that only the net value of businesses is recorded in the survey, while the national accounts record separately the assets and liability of unincorporated businesses alongside the household's.

Once this adjustment has been done, we get roughly comparable concepts of real wealth by summing the value of land, dwellings and other non-financial assets in the national accounts, and summing the value of real estate and self-employment wealth in the survey. It is not possible in general to further discriminate between real assets in a way that is consistent between surveys and the SNA. The United States are the exception: their balance sheet does include "equity in non-corporate business" as a separate item, so that we can consider separately real estate and self-employment businesses in both the SCF and the national balance sheet.

Next, we move the financial assets. Currency is not recorded in the surveys, so we remove it from the national balance sheets. That is the one place where the United States balance sheet is actually less detailed than in other countries: it does not record currency as a separate item, so we cannot remove it. This is a very minor issue since the value of currency is always negligible. The rest of deposits (sights and savings accounts) can be matched without much difficulty. Retirement products (life insurance and private pension funds) are also present in both the surveys and the national accounts. However, the surveys only ask about whole life insurance, while the national accounts calculate the value of all life insurance products based on the actuarial reserves of life insurance companies. This can at least partly explain why the value of life insurance tends to be lower in the surveys than in the national accounts (by about 30% to 40% in the United States). As I said earlier, "pension entitlements" in the United States balance sheet also include defined benefit pension plans by default, which are not included in the SCF. I therefore remove that item and replace it with the sum of "defined contribution pensions" and "annuities at life insurance companies" (see Dettling et al., 2015). The remaining financial wealth is made up of stocks and bonds, either held directly or by the intermediary of mutual funds. It is hard to precisely match each of these items between the surveys and the national accounts. That is especially true of the United States, where those assets are often held through managed accounts, including individual retirement accounts (IRA). The surveys directly record the value of managed assets, and gives little information regarding what these managed assets are ultimately invested in. The SNA, however, act as if those assets were held directly by the household sector. Therefore, it is preferable to consider them jointly. In the United States, the balance sheet gives the total value of IRAs, which I remove from bond and stock wealth and add to pension wealth.

Finally, we look at liabilities. While the SNA distinguishes liabilities based on terms, the surveys distinguish them based on the presence of a collateral. I therefore consider all liabilities jointly. Otherwise, with a few adjustments (see Dettling et al., 2015), concepts are broadly comparable between sources.

The precise definition of wealth that we will use is finally given in table 1.1. I divide financial assets into three broad categories: deposits, pension products (including life insurance) and the

joint value of stocks, bonds and mutual funds. I consider all real assets jointly, except in the United States where equity in incorporate business is a separate item on the balance sheet. I also consider all liabilities jointly.

1.1.2 Coverage and estimation methods

Estimations of wealth

The national accounts can estimate wealth by two methods. The first one is a census of wealth. Financial assets, for example, are estimated based on the balance sheets and off-balance sheets of individual financial institutions. The second is the perpetual inventory method (PIM), which cumulates past investment flows with suitable price adjustments to approximate the current value of assets. Both methods, including their strength and weaknesses, are discussed in detail in the data appendix of Piketty and Zucman (2014).

In surveys, wealth is measured based on the households' best assessment of the market value of their assets and liabilities. It mirrors the census-based estimates in the national accounts because a household's asset is some other entity's liability, and *vice versa*. It should also match PIM based estimates because present wealth must reflect past investments. In accounting terms, both should be equal. In practice, that is not the case. Discrepancies can be due to remaining inconsistencies between definitions, which cannot be fully eliminated. But that is probably not sufficient to explain what we observe. The rest of the disparities need to be explained by incomplete population coverage (especially of the wealthiest), or systematic errors of valuation. Section 1.2.2 compares both sources in detail.

Survey estimates are, of course, subject to sampling variability, and therefore have standard errors associated to them. But because of the complex design of the HFCS and the SCF, standard formulas do not apply. Instead, both surveys provide a set of a thousand so-called replicate weights, following the rescaling bootstrap procedure of Rao and Wu (1988). Those weights can be used to perform bootstrap replications of an estimate, which provide an accurate evaluation of uncertainty.

Another issue needs to be dealt with: partial non-response. All surveys suffer from it, but it is particularly critical here. We are indeed interested in variables (such as net wealth) that are derived from many others. If we were to drop an observation as soon as one of its subcomponents was missing, we would in general be left with very few and very unrepresentative households. To solve the problem, both the HFCS and the SCF impute missing values. Because the imputation process is not deterministic, it introduces additional uncertainty that needs to be taken into account when we report estimates. Both surveys have adopted the multiple imputation procedure of Rubin (1987): all imputations are performed five times, yielding five slightly different data sets. We then report average estimates over the five data sets, with a specific adjustment of standard errors to take into consideration the additional uncertainty due to the imputation. The HFCS methodological report (HFCN, 2013a) explains in detail how to combine both the

bootstrap replicate weights and the multiple imputation.

Institutional units

We are interested in household wealth. However, the national accounts do not distinguish in a systematic way households from non-profit institutions serving households (NPISH). The frontier between both sectors is indeed fuzzy: for example, tax deductions on charitable givings can provide incentives for rich households to create shell foundations to shelter their assets and avoid taxes (Fack and Landais, 2016). In that case, merging both sectors is the right thing to do (Piketty and Zucman, 2014). In practice, the question of their inclusion is not a primary concern: whenever they can be distinguished from households, NPISH represent less than 10% of private wealth.

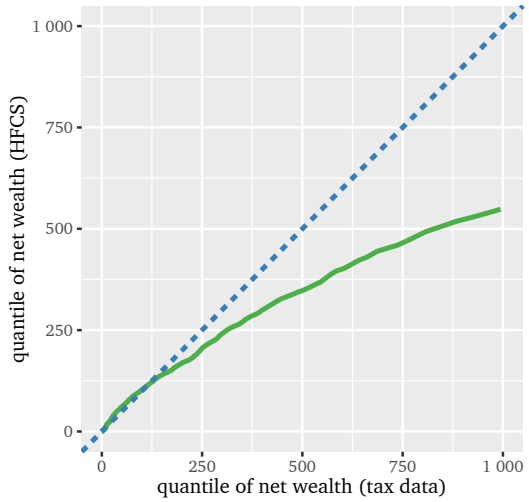
In the surveys, all households are part of the population of reference, with a few exceptions. The homeless, the prisoners, and the institutionalized population (including retirement homes) are typically excluded from the survey design (see HFCN, 2013a, p. 33, for details). Retirement homes, in particular, may represent a sizable chunk of the population with a relatively high wealth, and thus sensibly affect the results. In the United States, the SCF also excludes very high net worth individuals from the *Forbes* rankings for confidentiality reasons.

We also need to decide on the proper unit of statistical analysis. Wealth is always estimated at the household level, which is why most publications use the household as the unit of analysis. That can create problems: because it makes the distribution of wealth depend on household structure, it may hinder cross-country comparability (Bover, 2010; Fessler and Schürz, 2013). To avoid that problem, I will use the individual as the statistical unit, and individualize wealth by splitting it equally between spouses.

Coverage of countries

Regarding the surveys, the SCF and the HFCS cover a total of 16 countries: Austria, Belgium, Cyprus, Finland, France, Germany, Greece, Italy, Luxembourg, Malta, the Netherlands, Portugal, Slovakia, Slovenia, Spain, and the United States. First, I remove Finland from the sample because it uses register data that does not fully match the definitions of the survey (HFCN, 2013a, p. 75). Then, I also remove Greece and the Netherlands because of concerns over data quality. Of course, all wealth surveys potentially have weaknesses, and this is precisely what this dissertation is trying to address. But those two countries seem to be beyond the scope of what we can reasonably correct using national accounts or wealth rankings. Their surveys suggest implausibly low levels of inequality, with the top 1% owning less than 10% of the wealth.

The case of Netherlands is easy to observe by comparing the HFCS with administrative wealth data from Statistics Netherlands, provided by Wiemer Salverda. It is worth noting that this administrative data does not include life insurance or private pension funds, and uses households instead of individuals as the statistical unit. Therefore, it gives results that are not directly



Each point of the solid green line corresponds to the same quantile in both data sources. If the two distributions were the same, that line would be identical to the 45° line (dashed blue).

Amounts are expressed in 2010 euros. Households are the statistical unit of analysis.

Source: author calculations using the HFCS and wealth data from Statistics Netherlands, provided by Wiemer Salverda.

Figure 1.1: QQ-plot comparing net wealth in tax data and in the HFCS for the Netherlands

comparable to the ours. However, we can compare both distributions if we change the definition of wealth in the HFCS to match the concepts. In figure 1.1, we can see the QQ-plot for net wealth, which draws the quantile of net wealth in the tax data against the same quantile in the HFCS. Starting at about €150 000, the wealth associated with the same quantile becomes much lower in the HFCS than in the tax data. There is a large underestimation of wealth in the top half of the distribution because even moderately wealthy households are severely underrepresented. This is confirmed in table 1.2: the HFCS dramatically underestimates the mean and the top shares. The Dutch survey is also hard to reconcile with the *Quote 500*, the national ranking of top wealth holders. The fact that Netherlands is the only country where the survey was conducted solely through computer-assisted web interviews may at least partially explain the problem.

	mean	median	top 10% share	top 1% share
tax data	166 100	33 000	59.6%	22.4%
HFCS	127 300	54 100	46.2%	9.7%
	(5 300)	(7 600)	(1.5%)	(1.1%)

Bootstrapped standard errors in parentheses. 2010 euros. Households are the statistical unit of analysis. *Source:* author calculations using the HFCS and wealth data from Statistics Netherlands, provided by Wiemer Salverda.

Table 1.2: Summary statistics of the wealth distribution in the Netherlands

The Greek survey appears to suffer from the same kind of weaknesses, although there are no administrative data available to make a proper comparison. The problems for Greece may come from the survey design, which excludes small villages representing about 7% of the population.

Regarding the national accounts, most countries publish a financial balance sheet, but nonfinancial balance sheets are less widespread. That also reduces the number of countries we can study. Finally, we also need to consider journalist ranking of top wealth holders for chapter 3.

Countries with insufficient information on that front are also removed from the sample. In the end, we are left with seven countries for which the analysis can be carried out entirely: Austria, France, Germany, Italy, Portugal, Spain and the United States.

Oversampling

A major methodological difference between surveys concerns the oversampling of the wealthy, which means setting a higher probability of inclusion in the survey for richer households. Those wealthy households are subsequently given lower weights, to keep the survey representative of the whole population. But doing so can improve the quality of estimates, given that wealth is fat tailed and highly skewed.

In practice, oversampling requires some prior knowledge of how wealthy a potential respondent is before they are interviewed: the problem, of course, is that knowing the wealth of a respondent is precisely what the interview is for. The solution is to use some auxiliary information that can both be known in advance and is correlated with wealth. That information then serves as the basis for attributing different probabilities of inclusion. But not all countries have access to the same information, as shown in table 1.3. France and Spain use household-specific information on taxable wealth, so they can very precisely target the wealthy. The United States do the same with taxable income information. Germany uses relatively small-scale regional information on taxable income, while Austria and Portugal simply oversample their largest cities. Italy, finally, does not oversample at all.

	basis for oversampling	effective rate of oversampling		number of houesholds
		top 10%	top 1%	
Austria	Vienna oversampled	1%	1%	2380
Germany	Taxable income of region	115%	152%	3565
Spain	Taxable wealth	193%	880%	6197
France	Taxable wealth	126%	439%	15006
Italy	No oversampling	3%	-17%	7951
Portugal	Lisbon and Porto oversampled	14%	24%	4404
United States	Taxable income	126%	903%	6482

The effective oversampling rate of the top p% is $100(100S_p/p-1)$%, where S_p is the share of observations in the top p%. *Source:* author calculation, HFCN (2013a), and Kennickell (2009).

Table 1.3: Comparison of oversampling in different surveys

Those different strategies lead to different effective rates of oversampling. France, Spain and the United States exhibit extremely high rates of oversampling, followed by Germany. Austria, Portugal and Italy have much lower — or even negative — rates of oversampling. Unsurprisingly, oversampling is much more efficient when based on individual information.

Oversampling can decrease the sampling error: it lowers the variance of estimators, and also lowers finite sample biases (of extreme quantiles and top shares, in particular). It does not, in

itself, solves the problem of nonresponse bias. But we should still expect surveys with stronger oversampling to perform better on that front. To see why, we need to understand where nonresponse bias comes from.

Households refusing to answer do not necessarily create bias. Problems only arise if nonresponse is both unobserved and endogenous to the variable of interest (i.e. wealth). Statisticians can, and do correct for any nonresponse that is explained by observable factors through an *ex post* adjustment of survey weights. And if nonresponse is independent from wealth, it will not bias estimates of it. The real worry is that wealthier households have a lower response rate which is explained by unobserved factors. For countries that base their survey design on taxable wealth information, that set of potential unobserved factors is dramatically reduced. Bias will not arise there as long as nonresponse conditional on taxable wealth is independent from survey wealth, which is a fairly weak assumption.

In practice, it means that when countries have prior information on taxable wealth, they can observe how response rates vary with it, and correct the survey accordingly. Technically, such a correction could be made whether the country oversampled or not. In practice, when a country does not oversample, that is because it did not have access to that information in the first place. That is why countries with weaker oversampling schemes are also more subject to nonresponse bias.

There is very little that can convincingly be done on a sound statistical basis to correct for nonresponse bias, if it exists. There are, however, ways to limit sampling error, which will be the object of chapters 2 and 3. Sampling bias, in particular, can be a major concern when dealing with inequality statistics such as top shares (Taleb and Douady, 2015), and strongly limits the comparability of raw survey estimates.

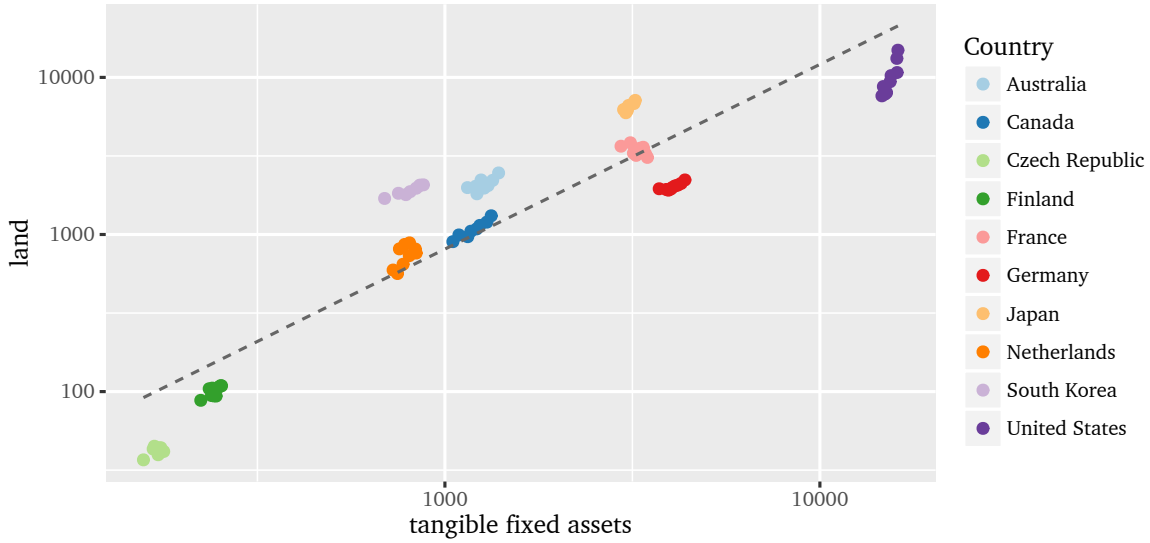
1.2 Aggregate wealth

1.2.1 Imputation of missing land values

Two countries (Austria and Portugal) do have a non-financial balance sheet, but only for produced assets (dwellings, essentially). In particular, they give no estimate for the value of land. This is a problem for comparisons, since in the surveys the value of dwellings and land are combined. To still include them in the analysis, I perform an imputation based on the fact that the value of tangible fixed assets is a strong predictor of the value of land (see figure 1.2). I estimate the following model:

$$\log(\text{land}) = \beta_0 + \beta_1 \log(\text{tangible fixed assets}) + \varepsilon$$

on all OECD countries for which data was available in the period 2006-2014. Table 1.4 shows OLS estimates with standard errors adjusted for country clusters. The HFCS methodological report (HFCN, 2013a) does a similar exercise, but simply assumes a constant ratio, which amounts to setting $\beta_1 = 1$. I allow for more flexibility, although the estimates I get are not significantly



Billions of constant 2010 euros at market exchange rates (log-log scale).
Source: author computations from OECD data.

Figure 1.2: Value of land and fixed tangible assets in some OECD countries, 2006–2014

different from the simpler model. Based on estimated coefficients, I impute the value of land as:

$$\text{land} = e^{\beta_0 + \frac{\text{Var}(\varepsilon)}{2}} (\text{tangible fixed assets})^{\beta_1}$$

	log(value of land)
log(value of fixed tangible assets)	1.176***
	(0.184)
constant	-2.643
	(2.624)
Observations	88
Clusters	10
R^2	0.843

Robust standard errors adjusted for country clusters in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table 1.4: Imputation model for the value of land

For Portugal, in addition to the absence of land value, there is no data for tangible fixed assets before 2012, whereas the survey was realized in 2010. Since dwellings make up the most part of that category (> 90%), I calculate a rough estimate for the year 2010 by deflating the 2012 amount using Eurostat's House Price Index.[5] In practice, I divide the 2012 value by 0.88.

[5] See `http://ec.europa.eu/eurostat/web/hicp/methodology/housing-price-statistics/house-price-index`.

1.2.2 Household surveys and national accounts

Table 1.5 gives the aggregate value of net wealth and its components according to the national balance sheet. To ease comparisons between countries, I give the figures on a per capita basis. Over the seven countries we study, net wealth per capita ranges from €70 000 in Portugal to €155 000 in Spain. The composition of wealth also varies: in particular, the United States stands out as the only country where financial assets (€80 000 per capita) exceed real assets (€50 000 per capita). In Spain, on the contrary, real assets are four times as valuable as financial assets. That heterogeneity is driven by multiple factors. For Spain, results may partially reflect the housing bubble, which had not entirely deflated by the year 2009. For the United States, it reflects both a low value of real assets, combined with a high value of private pension funds, bonds, stocks and mutual funds.

			real assets			financial assets				
		net wealth	all real	housing	business	all financial	deposits	bond, stocks and funds	life insurance and private pension funds	liabilities
Austria†	2010	131 500	94 800			56 200	25 100	21 000	10 000	19 400
France	2010	148 200	108 900			56 400	17 400	16 600	22 400	17 100
Germany	2011	110 300	75 800			53 100	20 800	12 700	19 600	18 600
Italy	2011	139 600	98 400			52 700	16 100	26 300	10 300	11 500
Portugal†	2010	68 600	51 100			32 800	13 600	12 300	6 900	15 400
Spain	2009	156 600	141 300			35 200	16 200	13 800	5 200	19 900
United States	2010	106 200	56 800	40 300	16 500	80 200	19 500	33 100	27 600	30 800

† Value of land imputed. Constant 2010 euros at market exchange rates. Values rounded to the nearest hundred. *Source:* author calculations from the Federal Reserve Financial Accounts for the United States, the W2ID for the real assets of Italy and Spain, and the OECD.

Table 1.5: Private wealth per capita (national accounts)

Table 1.6 estimates the same quantities, but using surveys. The orders of magnitude are similar, but the precise numbers can be quite different. Remaining differences between definitions can explain some of the discrepancies, but certainly not all of them.

			real assets			financial assets				
		net wealth	all real	housing	business	all financial	deposits	bond, stocks and funds	life insurance and private pension funds	liabilities
Austria	2010	112 900	95 100	69 400	25 700	25 400	13 200	10 300	1 900	7 600
France	2010	94 800	78 800	77 200	1 600	27 000	7 500	10 900	8 600	11 000
Germany	2011	87 200	71 400	65 500	5 900	28 700	10 000	12 600	6 100	13 000
Italy	2011	100 000	91 700	84 700	7 100	12 800	5 100	6 700	1 000	4 500
Portugal	2010	53 300	46 000	44 800	1 200	13 800	5 600	7 400	800	6 500
Spain	2009	104 300	98 700	93 900	4 800	17 800	6 400	9 500	1 900	12 200
United States	2010	126 700	80 400	62 300	18 100	74 400	11 100	37 000	26 300	27 700

Constant 2010 euros at market exchange rates. Values rounded to the nearest hundred. *Source:* author calculations from the HFCS and the SCF (for wealth data) and the OECD (for population data).

Table 1.6: Private wealth per capita (household surveys)

Table 1.7 compares both measures by giving the value of aggregate wealth in the survey as a percentage of aggregate wealth in the national accounts. In the United States, the SCF is in general quite close to the national accounts, with two exceptions: housing and deposits (see

		net wealth	real assets			financial assets				liabilities
			all real	housing	business	all financial	deposits	bond, stocks and funds	life insurance and private pension funds	
Austria†	2010	86%	100%			45%	53%	49%	19%	39%
France	2010	64%	72%			48%	43%	66%	39%	65%
Germany	2011	79%	94%			54%	48%	100%	31%	70%
Italy	2011	72%	93%			24%	32%	26%	9%	39%
Portugal†	2010	78%	90%			42%	41%	60%	12%	42%
Spain	2009	67%	70%			50%	40%	69%	36%	61%
United States	2010	123%	141%	155%	112%	93%	57%	112%	95%	90%

† Value of land imputed. The percentages are the ratios of aggregate wealth in the surveys to the value of aggregate wealth in the national accounts. *Source:* author's calculations from the HFCS, the SCF and the national accounts.

Table 1.7: Comparison of aggregate wealth in the survey and in the national account

also Henriques and Hsu, 2014; Dettling et al., 2015). Deposits are consistently lower in the SCF than in the national accounts. Part of the problem may be that currency has to be included in the national accounts total because the United States balance sheet does not separate it from other deposits. However, currency is usually a negligible part so it should not affect the result that much. Henriques and Hsu (2014) refer to Avery, Elliehausen, and Kennickell (1988) for other explanations: check float and the holdings of nonprofits like churches could explain some of the difference. Housing is the opposite: the SCF total is higher than in the national accounts, which can partially be attributed to homeowners overvaluing their house, especially at a time of price reversal.[6]

The other countries are covered by the HFCS, which generally finds lower levels of aggregate wealth than the national accounts. Different methodologies between countries can explain the heterogeneity of the results. Underestimation of net wealth is most severe in France and Spain, despite being the two countries with the strongest oversampling of the wealthy. That would suggest that underreporting, not just insufficient coverage of the top tail, is a problem.

Financial assets are the most severely underestimated. It is especially true of life insurance and private pension funds, which partially reflects a more restrictive definition (only whole life insurance is collected in the survey). The value of bonds, stocks and funds in Germany is an outlier: the survey estimate matches the national account, a result driven by a remarkably high value of stocks in the survey. For real assets, the survey estimates are closer to the national accounts. For Austria, in particular, both values are almost equal. That, however, reflects an exceptionally high value of business assets which is subject to a great uncertainty (see table 1.8). Moreover, the value of land was imputed in Austria, so it should be interpreted with caution.

[6]Other explanations have been suggested, having to do with the way national accounts estimate housing wealth: see Henriques (2013).

		net wealth		real assets						financial assets								liabilities	
				all real		housing		business		all financial		deposits		bonds, stocks and funds		life insurance and private pension funds			
		< 0	mean	> 0	mean	> 0	mean	> 0	mean	> 0	mean	> 0	mean	> 0	mean	> 0	mean	> 0	mean
Austria	2010	7.6%	165 000	57.7%	241 000	56.9%	178 200	8.4%	446 500	97.6%	38 000	97.3%	19 900	18.4%	81 800	19.3%	14 100	37.7%	29 300
		(1.1%)	(31 700)	(1.3%)	(48 800)	(1.2%)	(11 000)	(0.7%)	(303 300)	(0.3%)	(7 100)	(0.3%)	(1 700)	(1.2%)	(37 500)	(1.1%)	(2 600)	(1.3%)	(7 000)
France	2010	7.4%	138 000	65.4%	175 600	64.8%	173 600	3.2%	71 700	99%	39 700	98.8%	11 000	27.7%	57 000	39.5%	31 800	51%	31 500
		(0.4%)	(3 900)	(0.5%)	(2 900)	(0.5%)	(2 600)	(0.2%)	(26 200)	(0.1%)	(3 100)	(0.1%)	(300)	(0.5%)	(11 000)	(0.7%)	(1 300)	(0.6%)	(800)
Germany	2011	9.6%	116 800	55.9%	171 200	53.9%	163 000	6.3%	124 900	96.6%	39 900	95.2%	14 100	25.6%	66 200	49.4%	16 500	50.5%	34 400
		(0.8%)	(8 200)	(0.7%)	(11 500)	(0.7%)	(8 400)	(0.5%)	(43 900)	(0.5%)	(4 200)	(0.7%)	(600)	(1.2%)	(15 000)	(1.3%)	(800)	(1.2%)	(1 600)
Italy	2011	2.8%	155 800	76%	187 900	74%	178 100	11.2%	98 300	85%	23 500	83%	9 700	23.1%	45 200	19.4%	7 700	27.4%	25 600
		(0.3%)	(4 900)	(0.7%)	(5 500)	(0.8%)	(4 400)	(0.6%)	(17 500)	(0.5%)	(1 300)	(0.6%)	(400)	(0.6%)	(3 900)	(0.6%)	(300)	(0.8%)	(1 200)
Portugal	2010	4.2%	84 400	77.1%	94 600	76.7%	92 600	4.5%	41 600	95.4%	22 800	95.4%	9 300	10.4%	111 900	15.1%	8 600	40.2%	25 500
		(0.5%)	(4 800)	(1.3%)	(3 200)	(1.3%)	(3 100)	(0.5%)	(7 000)	(0.4%)	(3 700)	(0.4%)	(500)	(0.8%)	(32 500)	(0.7%)	(800)	(1.1%)	(1 000)
Spain	2009	5.5%	170 500	87.6%	184 100	87%	176 400	8.1%	96 000	93.7%	31 000	92.7%	11 300	19.5%	79 600	25.3%	12 200	53.5%	37 300
		(0.6%)	(5 700)	(0.8%)	(4 900)	(0.9%)	(4 400)	(0.7%)	(12 600)	(0.6%)	(2 400)	(0.7%)	(700)	(0.9%)	(10 300)	(1%)	(1 000)	(1.3%)	(1 900)
United States	2010	18.8%	210 900	73.7%	181 600	72.5%	143 200	12.1%	249 600	94.6%	130 900	93.6%	19 700	32.9%	186 900	61%	71 800	76.7%	60 200
		(0.4%)	(6 900)	(0.6%)	(4 900)	(0.6%)	(3 800)	(0.4%)	(12 200)	(0.3%)	(4 900)	(0.3%)	(900)	(0.7%)	(7 900)	(0.7%)	(3 400)	(0.5%)	(1 300)

Bootstrapped standard errors on parentheses. Constant 2010 euros at market exchange rates. Values rounded to the nearest hundred. Mean values are given conditionally on the participation in each asset category. *Source:* author calculations from the HFCS and the SCF.

Table 1.8: Ownership and conditional mean value of assets and liabilities in the surveys

		net wealth		real assets						financial assets								liabilities	
				all real		housing		business		all financial		deposits		bonds, stocks and funds		life insurance and private pension funds			
		top 10% share	top 1% share	top 10% share	top 1% share	top 10% share	top 1% share	top 10% share	top 1% share	top 10% share	top 1% share	top 10% share	top 1% share	top 10% share	top 1% share	top 10% share	top 1% share	top 10% share	top 1% share
Austria	2010	60.6%	22.6%	49.8%	18.6%	39.1%	12.5%	56%	10.5%	65.4%	28%	50.4%	15.1%	78.3%	26.3%	56.6%	12.9%	51%	11.6%
		(6.4%)	(7.1%)	(7.1%)	(6.1%)	(2.8%)	(3.4%)	(9.9%)	(10.7%)	(6.1%)	(9.7%)	(1.9%)	(2%)	(6.5%)	(13.5%)	(6.4%)	(6%)	(4%)	(3.2%)
France	2010	48.5%	17.3%	33.8%	9.6%	33%	8.9%	65.8%	34.6%	71.4%	35.2%	50.6%	14.4%	78.5%	45%	67.9%	28.4%	42.7%	10%
		(1.3%)	(1.9%)	(0.8%)	(1%)	(0.7%)	(0.8%)	(11.2%)	(21%)	(2.1%)	(4.5%)	(0.9%)	(1.3%)	(3.7%)	(8.6%)	(1.1%)	(2.1%)	(1%)	(0.8%)
Germany	2011	58.9%	23.8%	43.8%	16.3%	39.7%	13.1%	78.7%	0%	63.3%	31.5%	50.1%	13.2%	77.5%	37%	47.2%	12.9%	48.9%	10.4%
		(2.5%)	(3.7%)	(3.6%)	(4.2%)	(2.6%)	(3%)	(7.6%)	(8.4%)	(3.9%)	(6.4%)	(1.4%)	(1.1%)	(5.5%)	(12.4%)	(2.6%)	(3.1%)	(1.7%)	(1.3%)
Italy	2011	45.4%	14%	38.5%	11.5%	34.1%	9.1%	67%	28.5%	62.8%	26.5%	50.9%	19%	66.5%	18.3%	36.5%	9.2%	46.9%	10.1%
		(1.1%)	(1%)	(1.3%)	(1.1%)	(1.2%)	(0.7%)	(5.7%)	(7.7%)	(1.8%)	(2%)	(1.7%)	(2.2%)	(2%)	(4.5%)	(1.8%)	(1.5%)	(2.2%)	(1.5%)
Portugal	2010	54.1%	21.8%	40.4%	11.9%	39%	11.9%	55.6%	13.8%	78.8%	48.7%	63.7%	21.9%	81.5%	50.9%	51.3%	11.1%	35%	6.3%
		(2%)	(2.8%)	(1.3%)	(1.2%)	(1.4%)	(1.2%)	(6.5%)	(7.1%)	(3.2%)	(7.3%)	(1.6%)	(2.4%)	(5.2%)	(11.8%)	(2.5%)	(2%)	(1.4%)	(0.5%)
Spain	2009	44.4%	15%	35.4%	9.8%	34.2%	9%	49.7%	15.3%	75.4%	40%	64.8%	25.2%	78.1%	43.9%	54.6%	11.9%	40.9%	9.4%
		(1.1%)	(1.4%)	(0.8%)	(0.8%)	(0.9%)	(0.9%)	(5.6%)	(4.4%)	(1.7%)	(3.5%)	(2.3%)	(3.2%)	(2.8%)	(6.2%)	(2.8%)	(3.9%)	(1.6%)	(1.6%)
United States	2010	74.3%	33%	53%	21.6%	43.6%	12.8%	76.1%	36.1%	78.6%	36.6%	78.8%	36.8%	82.7%	41.1%	59.8%	18%	42.5%	10.9%
		(0.7%)	(0.9%)	(0.9%)	(0.8%)	(0.9%)	(0.5%)	(1.4%)	(2.7%)	(0.8%)	(1.1%)	(1%)	(2%)	(0.7%)	(1.6%)	(1.4%)	(1.5%)	(0.7%)	(0.5%)

Bootstrapped standard errors on parentheses. Top shares are given conditionally on the participation in each asset category. *Source:* author calculations from the HFCS and the SCF.

Table 1.9: Top 10% and 1% shares of assets and liabilities in the surveys

1.3 The distribution of wealth

1.3.1 Before adjustment

Table 1.8 shows mean net wealth, the fraction of people with negative net wealth, and for each subcomponent the participation rate and mean value conditional on participation. Table 1.9 show the top 10% and 1% shares for net wealth and each of its subcomponents. All these estimates come solely from the surveys, so they should be interpreted with caution. The top 1% share, in particular, is primarily given as a point of reference, but certainly goes beyond what we can ask of a survey, at least in countries without a strong oversampling of the wealthy (Austria, Germany, Italy, Portugal).

In spite of these limitations, some observations can be made. Estimates for Austria are characterized by large standard errors. This reflects a large variability between imputations in the top tail, and it is an issue that will be reflected in all the future estimates. The United States has a few distinctive features that separates it from European countries. First, the share of people with negative net worth (19%) is far higher than in any European country of the sample. It has a higher participation in financial assets (except deposits). It also has the highest participation rate and mean value of business assets (if we set Austria aside, whose estimate is meaningless given the standard error).

When it comes to inequality, real assets are more equally distributed than financial ones. That is a well-known fact, but it actually conceals significant heterogeneity. Inequality is indeed quite low for housing assets, but can be very high for business assets. As for the unequal repartition of financial assets, it is largely driven by bonds, stocks and mutual funds, more than deposits, pension funds or life insurance.

As I said, cross-country comparisons should be made with care given the methodological differences that have not yet been dealt with. But we can say that the United States seem to have the highest level of wealth inequality, with a top 10% estimated at 75%. For France and Spain, which have comparable oversampling, the same indicator is at 48% and 44% respectively. Despite low oversampling, Germany and Austria have much higher shares, around 60% each.

1.3.2 After adjustment

If we compare table 1.9 (which shows inequality for each subcomponent of net wealth) with table 1.7 (which shows by how much the surveys underestimate each of these subcomponents), we can see why surveys may underestimate inequality. Financial assets are the most unequal, and at the same time are relatively less important in surveys than in the national accounts. As a consequence, total wealth is likely to appear more equal in surveys than it really is.

We can easily deal with that issue under the assumption that misreporting of assets reflects systematic valuation mistakes, which are constant along the distribution of each component of wealth. All we need to do is to rescale the value of each assets and liabilities to match the

national accounts totals. In practice, because financial assets are a bigger share of net wealth at the top, that assumption means richer households do undervalue their wealth more severely, which is often suspected. But that effect is purely driven by the composition of their wealth.

		before adjustment			after adjustment		
		mean	top 10% share	top 1% share	mean	top 10% share	top 1% share
Austria	2010	165 000	60.6%	22.6%	192 200	63.5%	22.6%
		(31 700)	(6.4%)	(7.1%)	(35 800)	(7%)	(7.1%)
France	2010	138 000	48.5%	17.3%	215 800	49.7%	18.2%
		(3 900)	(1.3%)	(1.9%)	(5 800)	(1.2%)	(1.9%)
Germany	2011	116 800	58.9%	23.8%	147 900	56.2%	20.7%
		(8 200)	(2.5%)	(3.7%)	(9 200)	(2.3%)	(3.2%)
Italy	2011	155 800	45.4%	14%	217 300	49.5%	16.3%
		(4 900)	(1.1%)	(1%)	(7 300)	(1.3%)	(0.9%)
Portugal	2010	84 400	54.1%	21.8%	108 700	60.9%	25.6%
		(4 800)	(2%)	(2.8%)	(7 600)	(2.2%)	(3.6%)
Spain	2009	170 500	44.4%	15%	256 000	44.9%	15.1%
		(5 700)	(1.1%)	(1.4%)	(8 600)	(1.1%)	(1.4%)
United States	2010	210 900	74.3%	33%	176 700	80.7%	36.2%
		(6 900)	(0.7%)	(0.9%)	(6 300)	(0.8%)	(1%)

Constant 2010 euros at market exchange rates. *Source:* author calculations from the HFCS, the SCF and the national accounts.

Table 1.10: The distribution of wealth before and after adjustment to the national accounts

Table 1.10 compares the distribution of wealth before and after the adjustment to the national accounts. In general, the adjustment increases wealth inequality, but there are exceptions. In Austria, the top 1% share is not affected by the adjustment, and in Germany it actually decreases. That is because the adjustment also increases the importance of deposits, life insurance and pension funds, which are less unequal. In Germany, in particular, bonds, stocks and funds are the same in the survey and in the national accounts, so they lose importance compared to other assets. As I explained earlier, that effect is largely driven by an exceptionally high value of stocks in the surveys, for reasons that are not entirely clear.

Wealth inequality also increases in the United States. Housing and deposits are the only assets that are strongly affected by the adjustment. Both assets are quite equally shared. Thus, after the adjustment, the value of deposits is higher, which decreases inequality, and the value of houses is lower, which increases it. Overall, the effect on houses dominates, and overall inequality is higher. The reason why housing assets in the United States are so much higher in the survey that in the national accounts is not entirely clear either, but as we can see it is relevant to wealth inequality.

Chapter 2

Pareto and beyond

The first chapter of this dissertation looked at what household surveys could teach us on the distribution of wealth. I emphasized the limitations of these sources and the need to combine them with external information on top wealth holders. It is usually not enough, however, to just add those individuals to the sample. A true correction requires statistical modeling, which is the object of this second chapter.

The first statistical model for the distribution of income and wealth famously dates back to Vilfredo Pareto who, in the 1890s, noticed a striking fact: when plotted on logarithmic paper, the relationship between a given level of income x and the number $N(x)$ of people above that level looked linear. Pareto therefore conjectured the following relationship:

$$\forall x \geq \omega \qquad \log\left[\frac{N(x)}{N(\omega)}\right] = -\alpha \log\left(\frac{x}{\omega}\right) \tag{2.1}$$

and the associated probability distribution became known as the Pareto distribution. To Pareto, the relation (2.1) was not a mere empirical regularity: it was a universal law, one by which all human societies must abide. This had the far-reaching implication that nothing could — or for that matter should — be done about the unequal repartition of wealth and power.

That dogmatic interpretation has fallen out of fashion, but the model remains. Davies and Shorrocks (2000) call it a stylized fact. The Pareto distribution underlies many of the discussions on income and wealth: in particular, it is used by all the papers trying to address the measurement of wealth inequality from survey data (Vermeulen, 2014; Eckerstorfer et al., 2015; Bach, Thiemann, and Zucco, 2015; Westermeier and Grabka, 2015).

When it comes to the estimation of the model, however, the literature seems stuck in a dilemma. The first solution was adopted by Pareto himself, and involves fitting equation (2.1) using, say, ordinary least squares (OLS). The second solution is to use maximum likelihood estimation (MLE). The first approach is natural, intuitive, and has a clear graphical interpretation. Because it is simple to understand, it can easily be adapted to nonstandard configurations, which is why Vermeulen (2014) or Bach, Thiemann, and Zucco (2015) use it when combining survey

data and wealth rankings. But, to the best of my knowledge, there is no rigorous analysis of its statistical properties. The second approach is the standard of parametric statistics, and as such it has the stamp of approval of modern statistical theory. But, to the practitioner, it may seem to work in mysterious ways.

That dichotomy is, in fact, a false one. I will show that the regression approach has two well-identifiable flaws. Upon correction, it yields a new estimator with excellent properties that turns out to be exactly equal to the maximum likelihood estimator. It is therefore possible to recast the MLE procedure in a way that has the same graphical and intuitive interpretation as the regression approach.

Then, I will deal with the possibility that the Pareto distribution may in fact be too restrictive a model. I will introduce a new concept, the tail function, as a nonparametric generalization of (2.1). It builds on Pareto's original insight that the relationship between the logarithm of wealth and the logarithm of the rank behaves well, but allows for more flexibility. The parametric framework can be adapted to estimate the tail function nonparametrically, yielding a new method which also works for distributions that are not strictly Paretian.

Section 2.1 introduces the basic tools used to derive the estimators. Section 2.2 analyzes the parametric problem, and section 2.3 extends the results to the nonparametric case.

2.1 Preliminaries

This section start by introducing the so-called order statistics, which will prove central in the analysis. I give a few of their properties, most of which can be found in David and Nagaraja (2005). I then present the Pareto distribution in detail, and most importantly its relation with the exponential distribution.

2.1.1 Order statistics

Let $(X_1, X_2, \dots, X_n)$ be n independent and identically distributed (iid) absolutely continuous random variables. Denote F their cumulative distribution function (CDF), and f their probability density function (PDF). If the variables are sorted in increasing order and written as:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

then $X_{(r)}$ is called the r-th order statistic. Although the original sample $(X_1, X_2, \dots, X_n)$ is iid, the sorted sample $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ obviously isn't: different order statistics cannot have the same distribution, and the inequality between them implies at least some dependence.

Increasing transformation

Let g be an increasing function. For all $i \in \{1, 2, \dots, n\}$, define $Z_i = g(X_i)$. Then, for all $r \in \{1, 2, \dots, n\}$:

$$Z_{(r)} = g(X_{(r)}) \tag{2.2}$$

That property simply states that an increasing transformation does not change the ordering of variables.

Distribution of a single order statistic

Let $f_{(r)}$ be the PDF of $X_{(r)}$:

$$f_{(r)}(x) = \lim_{\mathrm{d}x \to 0} \frac{\mathbb{P}\{X_{(r)} \in [x, x+\mathrm{d}x]\}}{\mathrm{d}x} \tag{2.3}$$

The event $X_{(r)} \in [x, x+\mathrm{d}x]$ can be represented as in figure 2.1. Out of the entire set of n variables, it requires $r-1$ variables in $]-\infty, x]$, one variable in $[x, x+\mathrm{d}x]$, and $n-r$ variables in $[x+\mathrm{d}x, +\infty[$.

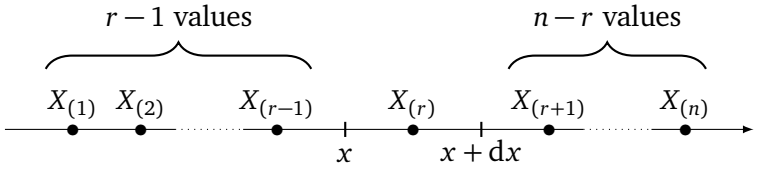


Figure 2.1: Configuration of the event $X_{(r)} \in [x, x+\mathrm{d}x]$

It amounts to choosing $r-1$ variables from the entire set of n variables, then one variable among the remaining $n-r+1$; the rest will necessarily fall in the last category. Define $C_{r,n}$ the number of such groupings. Recall that the number of ways of choosing k values among n is given by the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Therefore:

$$C_{r,n} = \underbrace{\binom{n}{r-1}}_{\substack{\text{choose } r-1 \\ \text{values among } n}} \times \underbrace{\binom{n-r+1}{1}}_{\substack{\text{choose 1 value among} \\ \text{the remaining } n-r+1}}$$

which simplifies to:

$$C_{r,n} = \frac{n!}{(r-1)!(n-r+1)!} \times \frac{(n-r+1)!}{1!(n-r)!} = \frac{n!}{(r-1)!(n-r)!}$$

F being the CDF of X, each of these repartitions happens with probability:

$$[F(x)]^{r-1} \times [F(x+\mathrm{d}x) - F(x)] \times [1 - F(x+\mathrm{d}x)]^{n-r}$$

Therefore, the probability of the event $X_{(r)} \in [x, x+\mathrm{d}x]$ is:

$$\begin{aligned}\mathbb{P}\{X_{(r)} \in [x, x+\mathrm{d}x]\} &= C_{r,n} \times [F(x)]^{r-1} \times [F(x+\mathrm{d}x) - F(x)] \times [1 - F(x+\mathrm{d}x)]^{n-r} \\ &= C_{r,n} \times [F(x)]^{r-1} \times f(x)\mathrm{d}x \times [1 - F(x+\mathrm{d}x)]^{n-r} + O(\mathrm{d}x)^2\end{aligned}$$

Replacing in (2.3) gives:

$$f_{(r)}(x) = C_{r,n} f(x) [F(x)]^{r-1} [1 - F(x)]^{n-r} \tag{2.4}$$

Joint distribution of two order statistics

Similar arguments can be used to derive the joint PDF of two order statistics $X_{(r)}$ and $X_{(s)}$ $(1 \leq r < s \leq n)$. Consider the event $X_{(r)} \in [x, x+\mathrm{d}x]$ and $X_{(s)} \in [y, y+\mathrm{d}y]$, which is realized by the configuration of figure 2.2.

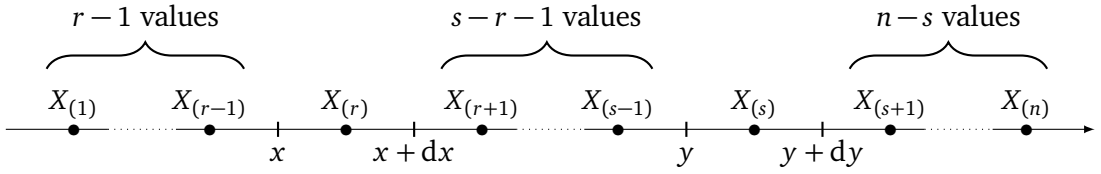


Figure 2.2: Configuration of the event $X_{(r)} \in [x, x+\mathrm{d}x]$ and $X_{(s)} \in [y, y+\mathrm{d}y]$

Here, it requires $r-1$ variables in $]-\infty, x]$, 1 in $[x, x+\mathrm{d}x]$, $s-r-1$ in $[x+\mathrm{d}x, y]$, 1 in $[y, y+\mathrm{d}y]$, and finally $n-s$ in $[y+\mathrm{d}y, +\infty[$. The number of such groupings is:

$$C_{r,s,n} = \underbrace{\binom{n}{r-1}}_{\substack{\text{choose } r-1 \\ \text{values among } n}} \times \underbrace{\binom{n-r+1}{1}}_{\substack{\text{choose 1 value among} \\ \text{the remaining } n-r+1}} \times \underbrace{\binom{n-r}{s-r-1}}_{\substack{\text{choose } s-r-1 \text{ values} \\ \text{among the remaining } n-r}} \times \underbrace{\binom{n-s+1}{1}}_{\substack{\text{choose 1 value among} \\ \text{the remaining } n-s+1}}$$

which simplifies to:

$$C_{r,s,n} = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}$$

Each of these configurations happens with probability:

$$[F(x)]^{r-1} \times [F(x+\mathrm{d}x) - F(x)] \times [F(y) - F(x)]^{s-r-1} \times [F(y+\mathrm{d}y) - F(y)] \times [1-F(y)]^{n-s}$$

Using the same argument as before, it follows that for $y > x$:

$$f_{(r)(s)}(x,y) = C_{r,s,n}\, f(x) f(y) [F(x)]^{r-1} [F(y) - F(x)]^{s-r-1} [1-F(y)]^{n-s} \tag{2.5}$$

and $f_{(r)(s)}(x,y) = 0$ otherwise.

Expected value

Define $\mu_{(r)}$ the expected value of $X_{(r)}$, if it exists.[1] Equation (2.4) implies:

$$\begin{aligned} \mu_{(r)} &= \int_{-\infty}^{+\infty} x f_{(r)}(x)\, \mathrm{d}x \\ &= C_{r,n} \int_{-\infty}^{+\infty} x f(x) [F(x)]^{r-1} [1-F(x)]^{n-r}\, \mathrm{d}x \end{aligned}$$

With the change of variable $u = F(x)$, we get:

$$\mu_{(r)} = C_{r,n} \int_0^1 Q(u)\, u^{r-1} (1-u)^{n-r}\, \mathrm{d}u \tag{2.6}$$

where $Q = F^{-1}$ is the quantile function of X.

[1] $\mu_{(r)}$ will always be finite if $\mathbb{E}|X| < +\infty$. This condition is sufficient but not necessary, since for example all the order statistics of a Cauchy distribution except the first and the last one have a finite expected value. Analogous conditions hold for all moments.

Variance

Define $\sigma^2_{(r)}$ the variance of $X_{(r)}$. As before, equation (2.4) implies:

$$\begin{aligned}\mu_{(r)} &= \int_{-\infty}^{+\infty} (x-\mu_{(r)})^2 f_{(r)}(x)\,\mathrm{d}x \\ &= C_{r,n}\int_{-\infty}^{+\infty} (x-\mu_{(r)})^2 f(x)[F(x)]^{r-1}[1-F(x)]^{n-r}\,\mathrm{d}x\end{aligned}$$

With the change of variable $u = F(x)$, we get:

$$\mu_{(r)} = C_{r,n}\int_0^1 [Q(u)-\mu_{(r)}]^2\, u^{r-1}(1-u)^{n-r}\,\mathrm{d}u \tag{2.7}$$

Covariance

The covariance $\sigma_{(r)(s)}$ of $X_{(r)}$ and $X_{(s)}$ can be derived similarly using equation (2.5):

$$\begin{aligned}\sigma_{(r)(s)} &= \int_{-\infty}^{+\infty}\int_{-\infty}^{y} (x-\mu_{(r)})(y-\mu_{(s)}) f_{(r)(s)}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= C_{r,s,n}\int_0^1\int_0^v [Q(u)-\mu_{(r)}][Q(v)-\mu_{(s)}]\, u^{r-1}(v-u)^{s-r-1}(1-v)^{n-s}\,\mathrm{d}u\,\mathrm{d}v\end{aligned} \tag{2.8}$$

Median

We finish this overview of order statistics by looking at the median. We know that $X_{(r)}$ is absolutely continuous because we previously derived its PDF: see formula (2.4). Hence, a median value $m_{(r)}$ of $X_{(r)}$ is any solution to the equation:

$$F_{(r)}[m_{(r)}] = \frac{1}{2} \tag{2.9}$$

where $F_{(r)}$ is the CDF of $X_{(r)}$. An expression for $F_{(r)}$ can be obtained by integration of (2.4), or more simply by a direct argument:

$$\begin{aligned}F_{(r)}(x) &= \mathbb{P}\{X_{(r)} \le x\} \\ &= \mathbb{P}\{\text{at least } r \text{ of the } X_1, X_2, \ldots, X_n \text{ are less than or equal to } x\} \\ &= \sum_{i=r}^{n} \mathbb{P}\{\text{exactly } i \text{ of the } X_1, X_2, \ldots, X_n \text{ are less than or equal to } x\} \\ &= \sum_{i=r}^{n} \binom{n}{i} F^i(x)[1-F(x)]^{n-i}\end{aligned} \tag{2.10}$$

Then, define the regularized incomplete Beta function as:

$$I_p(a,b) = \frac{1}{\mathrm{B}(a,b)}\int_0^p t^{a-1}(1-t)^{b-1}\,\mathrm{d}t \tag{2.11}$$

where $\mathrm{B}(a,b)$ is the Beta function. A well-known relationship (Walck, 1996, p. 163) states that (2.10) can be rewritten as:

$$F_{(r)}(x) = I_{F(x)}(r, n-r+1)$$

Equation (2.11) implies that $p \mapsto I_p(a,b)$ is continuous and strictly increasing from 0 to 1 over $[0,1]$. If, moreover, F is also continuous and strictly increasing, then equation (2.9) admits a unique solution, and the median of $X_{(r)}$ is uniquely defined. Solving the equation yields:

$$m_{(r)} = Q[I^{-1}_{1/2}(r, n-r+1)] \tag{2.12}$$

where $p \mapsto I^{-1}_p$ is the inverse regularized Beta function.

2.1.2 The Pareto distribution

Let $\omega > 0$ and $\beta > 0$. We say that X follows a Pareto distribution, and write $X \sim \mathscr{P}(\omega, \beta)$ if for all $x \geq \omega$:

$$\mathbb{P}\{X \geq x\} = \left(\frac{\omega}{x}\right)^{1/\beta} \tag{2.13}$$

and $\mathbb{P}(X \geq x) = 1$ for $x < \omega$. Therefore, the CDF and the PDF of a Pareto distribution are, for $x \geq \omega$:

$$f(x) = \frac{\omega^{1/\beta}}{\beta x^{1/\beta+1}} \qquad \text{and} \qquad F(x) = 1 - \left(\frac{\omega}{x}\right)^{1/\beta}$$

Comments on the parameters

Most authors — including Pareto himself, as we saw in introduction — work with a different parameterization, namely $\alpha = 1/\beta$, where α is called the Pareto coefficient. The parameterization we use will prove more practical for our purpose. It is also consistent with extreme value theory where β is called the tail index.

The Pareto distribution has a strictly positive lower bound ω. For that reason, it is obviously not a suitable model for the lower end of the wealth distribution. In practice, the researcher chooses *ex ante* a value for ω above which the Pareto model is deemed adequate, and discards all observations below that threshold. The choice of ω is not always obvious and results may be sensitive to it. The nonparametric approach of section 2.3 will be a way to circumvent the problem. From now on, I will assume that ω is known and focus on the estimation of β.

Relation to Pareto's original formulation

The link between Pareto's original statement (2.1) and the definition (2.13) appears if we consider $\mathbb{P}\{X \geq x\} = \frac{N(x)}{N(\omega)}$, where $N(x)$ is the size of the population with wealth above x. Take the exponential on both sides of (2.1) and we get:

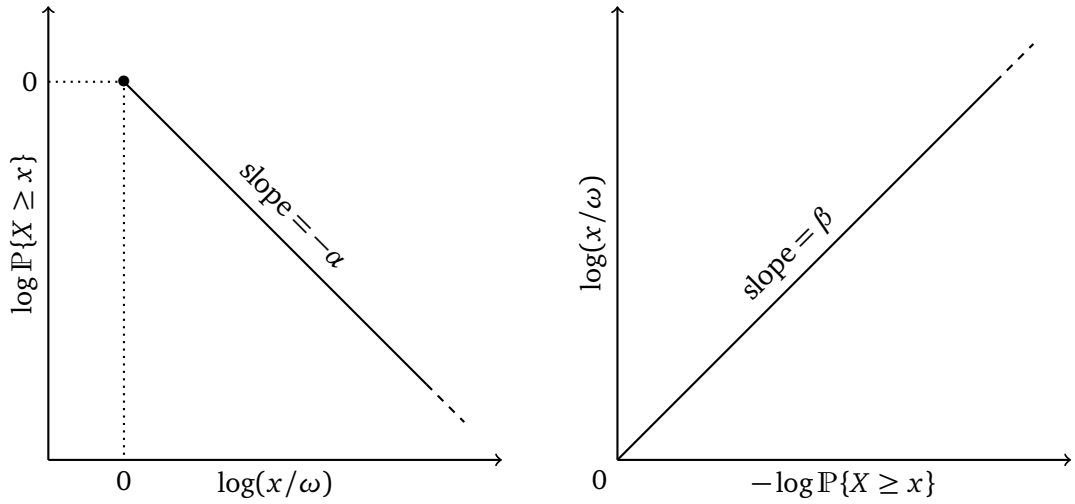
$$\begin{aligned}\mathbb{P}\{X \geq x\} &= \left(\frac{x}{\omega}\right)^{-\alpha} \\ &= \left(\frac{\omega}{x}\right)^{1/\beta}\end{aligned}$$

with $\beta = 1/\alpha$, which is the definition (2.13).

The reversed Pareto law

In addition to the change of parameterization, we will write Pareto's original law (2.1) in a different way: we will put $\log(x/\omega)$ on the left-hand side of the equality, and $\log \mathbb{P}\{X \geq x\}$ on the right-hand side. Using β as the parameter, it means:

$$\log\left(\frac{x}{\omega}\right) = \beta \times (-\log \mathbb{P}\{X \geq x\}) \tag{2.14}$$



(a) Pareto's original diagram

(b) The new diagram

Figure (a) correspond to Pareto's original decreasing relationship between wealth x and the probability of being above x. α is the so-called Pareto coefficient from equation (2.1). Figure (b) is more suited to the alternative parameterization used in this dissertation: the slope $\beta = 1/\alpha$ is called the tail index.

Figure 2.3: The Pareto diagrams for two parameterizations

Figure 2.3 compares the diagram initially used by Pareto with the one implied by (2.14). The distinction may seem superfluous, yet there are good reasons to use (2.14) rather than (2.1). Indeed, in the estimation, we will interpret (2.14) as a regression equation where $\log(x/\omega)$ is random and $\log \mathbb{P}\{X \geq x\}$ is fixed. Thus, the algebra will be more natural if we view the former as the "outcome variable" and the latter as the "explanatory variable".

Expected value and variance

Table 2.1 gives the expected value and the variance of a Pareto distribution for different values of the parameter. Obviously, $\beta \geq 1$ is not acceptable since it implies an infinite mean, which would translate into an infinite amount of aggregate wealth. Any other value $0 < \beta < 1$ is possible, although evidence strongly suggests $1/2 < \beta < 1$.

The distribution of wealth would therefore have an infinite variance. That fact is not innocuous for statistical analysis: with infinite variance, many staples of statistical theory break down.

tail index β	$0<\beta<1/2$	$1/2\leq\beta<1$	$1\leq\beta<+\infty$
Pareto coefficient $\alpha=1/\beta$ inverted Pareto coefficient $b=\frac{1}{1-\beta}$	$2<\alpha<+\infty$ $1<b<2$	$1<\alpha\leq 2$ $2\leq b<+\infty$	$0<\alpha\leq 1$ $+\infty$
expected value	$\frac{\omega}{1-\beta}$		$+\infty$
variance	$\frac{\omega^2\beta^2}{(1-\beta)^2(1-2\beta)}$	$+\infty$	

Table 2.1: Moments of a Pareto distribution

Although the law of large numbers still applies, convergence may be very slow, and the absence of central limit theorem prevents an easy assessment of uncertainty.

The logarithm of a Pareto law: the exponential distribution

Let $\theta>0$. We say that X follows an exponential distribution, and write $X\sim\mathscr{E}(\theta)$ if for all $x\geq 0$:

$$\mathbb{P}\{X\geq x\}=\mathrm{e}^{-x/\theta}$$

and $\mathbb{P}(X\geq x)=1$ for $x<0$. Therefore, the CDF and the PDF of an exponential distribution are, for $x\geq 0$:

$$f(x)=\frac{1}{\theta}\mathrm{e}^{-x/\theta}\qquad\text{and}\qquad F(x)=1-\mathrm{e}^{-x/\theta}$$

The exponential distribution is tamer than the Pareto distribution since it admits finite moments of every order. Interestingly, there is a simple relation between both distributions. Assume $X\sim\mathscr{P}(\omega,\beta)$ and define $Z=\log(X/\omega)$. We have, for all $z>0$:

$$\begin{aligned}\mathbb{P}\{Z\geq z\}&=\mathbb{P}\{\log(X/\omega)\geq z\}\\&=\mathbb{P}\{X\geq\omega\mathrm{e}^{z}\}\\&=\left(\frac{\omega}{\omega\mathrm{e}^{z}}\right)^{1/\beta}\\&=\mathrm{e}^{-z/\beta}\end{aligned}$$

Hence $Z\sim\mathscr{E}(\beta)$. By working with the logarithm of wealth, we can circumvent the infinite variance problem while still retaining all the relevant information on the parameter of interest.

Order statistics of the exponential distribution

Let $Z\sim\mathscr{E}(\theta)$. Its quantile function is given by:

$$Q(x)=-\theta\log(1-x)$$

From (2.6), the expected value of $Z_{(r)}$ is therefore:

$$\mu_{(r)}=\theta C_{r,n}\int_0^1-\log(1-u)\,u^{r-1}(1-u)^{n-r}\,\mathrm{d}u$$

Solving the integral yields (Balakrishnan and Basu, 1995, p. 19):

$$\mu_{(r)} = \theta(H_n - H_{n-r}) \tag{2.15}$$

where H_m is called the m-th harmonic number and is defined as:

$$H_m = \sum_{k=1}^{m} \frac{1}{k}$$

Similarly, equations (2.7) and (2.8) yield the formula for variance and covariance ($1 \leq r < s \leq n$):

$$\sigma^2_{(r)} = \sigma_{(r)(s)} = \theta^2(H^{(2)}_n - H^{(2)}_{n-r}) \tag{2.16}$$

where $H^{(2)}_m$ is called the second order m-th harmonic number and is defined as:

$$H^{(2)}_m = \sum_{k=1}^{m} \frac{1}{k^2}$$

Finally, equation (2.12) gives the median value of $Z_{(r)}$:

$$m_{(r)} = -\theta \log[1 - I^{-1}_{1/2}(r, n-r+1)] \tag{2.17}$$

2.2 Parametric estimation

Let $\omega > 0$ and $\beta > 0$. Let $(X_1, X_2, \ldots, X_n) \overset{\text{iid}}{\sim} \mathscr{P}(\omega, \beta)$. ω is known and we seek to estimate the tail index β. In section 2.2.1, I consider what I call the simple estimator, which is akin to Pareto's original approach, and is still used in practice. In section 2.2.2, I address the flaws of the simple estimator and define the generalized least squares (GLS) estimator of the tail index, which will turn out to be identical to MLE.

None of the methods developed in this section will be directly applied in this dissertation, yet I mention them for two reasons. First, they provide results on the estimation of Pareto distributions which may be of interest and seem currently missing from the literature. Second, they lay the groundwork for the nonparametric approach of section 2.3.1, which will be used afterwards in chapter 3.

2.2.1 The simple estimator

Recall from section 2.1.2 that the Pareto distribution obeys:

$$\log\left(\frac{x}{\omega}\right) = -\beta \log \mathbb{P}\{X \geq x\}$$

For $k \in \{1, 2, \ldots, n\}$, consider the following empirical counterpart, which corresponds to the diagram 2.3b:

$$\log\left(\frac{X_{(k)}}{\omega}\right) \approx -\beta \log\left(\frac{n-k+1}{n+1}\right) \tag{2.18}$$

To get an estimate of β, we apply OLS fitting to the n equations (2.18), which leads to the following minimization problem:

$$\hat{\beta}^{\text{simple}} = \underset{\beta}{\operatorname{argmin}} \sum_{k=1}^{n} \left[\log\left(\frac{X_{(k)}}{\omega}\right) - \beta \log\left(\frac{n+1}{n-k+1}\right) \right]^2$$

The first order condition for that minimization problem is:

$$-2\sum_{k=1}^{n} \log\left(\frac{n-k+1}{n+1}\right)\left[\log\left(\frac{X_{(k)}}{\omega}\right) - \hat{\beta}^{\text{simple}} \log\left(\frac{n+1}{n-k+1}\right)\right] = 0$$

Therefore:

$$\hat{\beta}^{\text{simple}} = \frac{\sum_{k=1}^{n} \log\left(\frac{X_{(k)}}{\omega}\right) \log\left(\frac{n+1}{n-k+1}\right)}{\sum_{k=1}^{n} \left[\log\left(\frac{n+1}{n-k+1}\right)\right]^2}$$

The simple estimator applies OLS fitting to a problem that does not strictly satisfy the standard assumptions of the method: indeed, the approximation (2.18) cannot be written in terms of conditional expectations, and the observations are not iid. Although that will not affect the asymptotic convergence of the estimator, that will make it biased and inefficient.

Bias

Since $\log(X/\omega) \sim \mathscr{E}(\beta)$, equation (2.15) yields:

$$\mathbb{E}\left[\log\left(\frac{X_{(k)}}{\omega}\right)\right] = \beta(H_n - H_{n-k}) \tag{2.19}$$

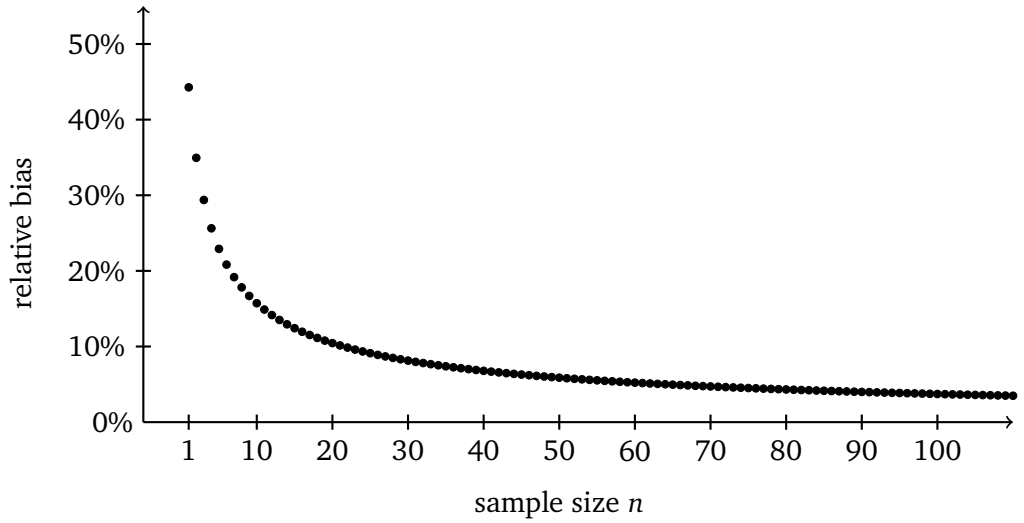
Therefore:

$$\mathbb{E}[\hat{\beta}^{\text{simple}}] = \beta \frac{\sum_{k=1}^{n} (H_n - H_{n-k}) \log\left(\frac{n+1}{n-k+1}\right)}{\sum_{k=1}^{n} \left[\log\left(\frac{n+1}{n-k+1}\right)\right]^2} \neq \beta \tag{2.20}$$

As we can see, the simple estimator is biased, because the sum at the numerator and the sum at the denominator in (2.20) do not cancel out. Figure 2.4 shows that the bias can be large for small n, but diminishes quickly and eventually becomes very little for large enough sample sizes.

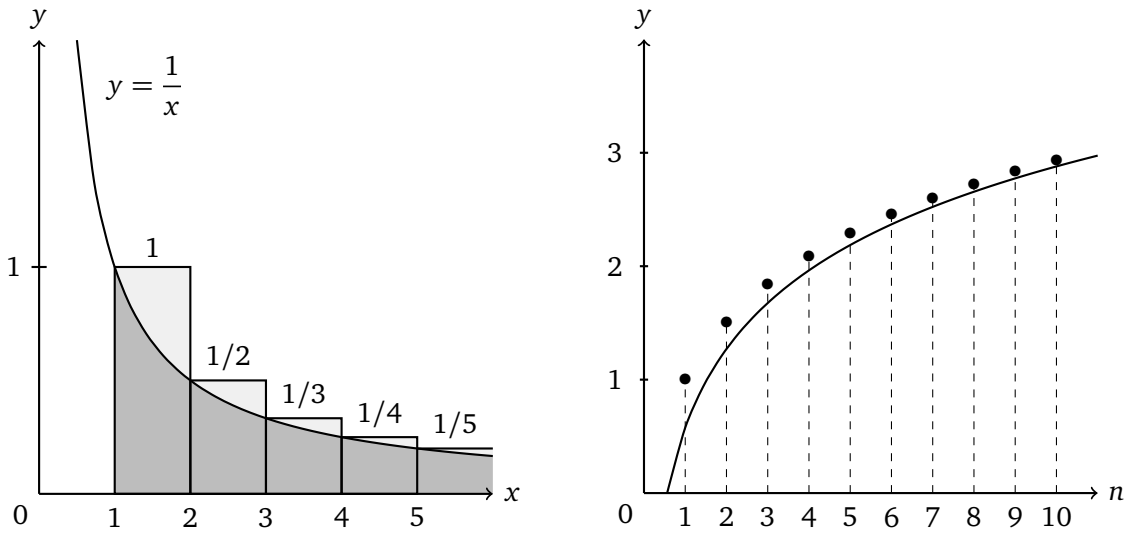
However small, a deeper analysis of that bias carries some insights that are relevant to the pertinence of Pareto's diagram, especially for the nonparametric approach of section 2.3. The bias would disappear if we were to replace $\log\left(\frac{n+1}{n-k+1}\right)$ by $H_n - H_{n-k}$ in (2.20). Such a modification would require putting observations along the x-axis according to their harmonic rank $H_n - H_{n-k}$, and not, as Pareto did, their logarithmic rank $\log\left(\frac{n+1}{n-k+1}\right)$. Truth is, the difference is subtle since, as figure 2.5 shows, harmonic numbers are a Riemann sum approximation of the logarithm. More precisely, Euler proved in 1734:

$$\lim_{n \to +\infty} H_n - \log n = \gamma$$



The relative bias is defined as $\mathbb{E}[\hat{\beta}^{\text{simple}}]/\beta - 1$.

Figure 2.4: Relative bias of the simple estimator of the tail index



(a) Integral and Riemann sum

(b) H_n and its asymptotic limit $\gamma + \log n$

In figure (a), the logarithm is the dark-gray area under the curve (i.e. the integral) while the harmonic number is the dark and light-gray area under the rectangles (i.e. the Riemann sum). The area of the difference (light-gray) converges towards the Euler-Mascheroni constant $\gamma \approx 0.577$.

Figure 2.5: Comparison of harmonic numbers and the natural logarithm

where $\gamma \approx 0.577$ is the Euler-Mascheroni constant.

That result has different consequences depending on which order statistic we consider at the limit. For middle range order statistics, i.e. $X_{(k)}$ with $k/n \to 0$, we have:

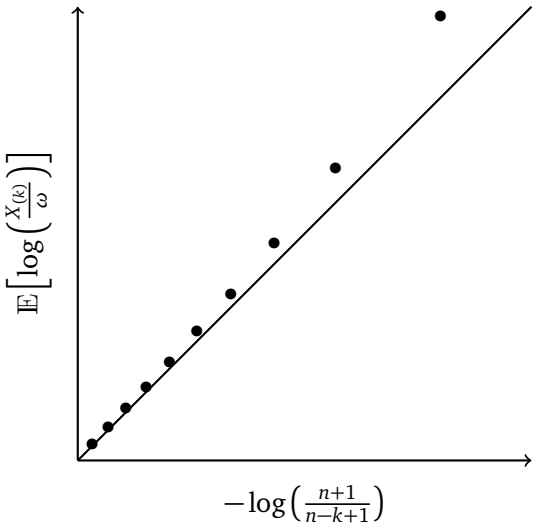
$$\lim_{n \to +\infty} H_n - H_{n-k} - \log\left(\frac{n+1}{n-k+1}\right) = 0$$

Asymptotically, they are therefore correctly positioned on Pareto's diagram. That is not true,

however, of extreme order statistics. Take for example $X_{(k)}$ with $k = n - r + 1$ (the r-th highest order statistic, r being fixed):

$$\lim_{n\to+\infty} H_n - H_{n-k} - \log\left(\frac{n+1}{n-k+1}\right) = \gamma - H_{r-1} + \log r$$

So the last data points on Pareto's diagram are systematically misplaced, even in very large samples. Figure 2.6 shows graphically what this misplacing looks like.



The simple estimator assumes that, on average, observations lie on the straight line.
The bullets show the true average positions of observations for $n = 10$.

Figure 2.6: Bias of the simple estimator

The bias of the simple estimator disappears asymptotically because, as opposed to the extreme order statistics, the middle range order statistics get more numerous when the sample size increases. Thus, the non-biased part of the estimator dominates eventually. Still, the bias can be costlessly avoided by choosing the proper abscissa for $X_{(k)}$ on Pareto's diagram.

Another way to see the problem is that for OLS to operate properly here, we would need to be able to write:

$$\mathbb{E}\left[\log\left(\frac{X_{(k)}}{\omega}\right)\right] = -\beta \log\left(\frac{n-k+1}{n+1}\right)$$

But this equality is false. The real relationship is:

$$\mathbb{E}\left[\log\left(\frac{X_{(k)}}{\omega}\right)\right] = -\beta(H_n - H_{n-k})$$

To understand where the incorrect equality came from, consider n uniform random variables $(U_1, \ldots, U_n) \overset{\text{iid}}{\sim} \mathscr{U}([0,1])$. We then have (David and Nagaraja, 2005, p. 35):

$$\mathbb{E}[U_{(k)}] = \frac{k}{n+1}$$

Apply the quantile function $Q : x \mapsto -\beta \log(1-x)$ of $\log(X/\omega)$ to that equality and we get:

$$Q(\mathbb{E}[U_{(k)}]) = -\beta \log\left(\frac{n-k+1}{n+1}\right)$$

Moreover, $Q(U)$ has the same distribution as $\log(X/\omega)$. Hence:

$$\mathbb{E}[Q(U_{(k)})] = \mathbb{E}\left[\log\left(\frac{X_{(k)}}{\omega}\right)\right]$$

Therefore, the simple estimator implicitly equates $\mathbb{E}[Q(U_{(k)})]$ and $Q(\mathbb{E}[U_{(k)}])$. But this would only be true if the quantile function were linear, which would mean that $\log(X/\omega)$ followed a uniform distribution in the first place.

In a nutshell, because of the functional non-invariance of the expected value, Pareto's diagram does not, without adjustments, lend itself to regression analysis. In the strict Pareto case, the solution is to use the harmonic ranks of order statistics instead of their logarithmic ranks. Section 2.3 will explore a more general solution to the problem that works in the nonparametric case.

Inefficiency

Even if the bias problem were solved, the simple estimator would still be inefficient because order statistics are correlated and their variances differ: see formula (2.16). Figure 2.7 shows how the variance of $\log(X_{(k)}/\omega)$ increases with k. In particular, there is a dramatic increase for the very last order statistics.

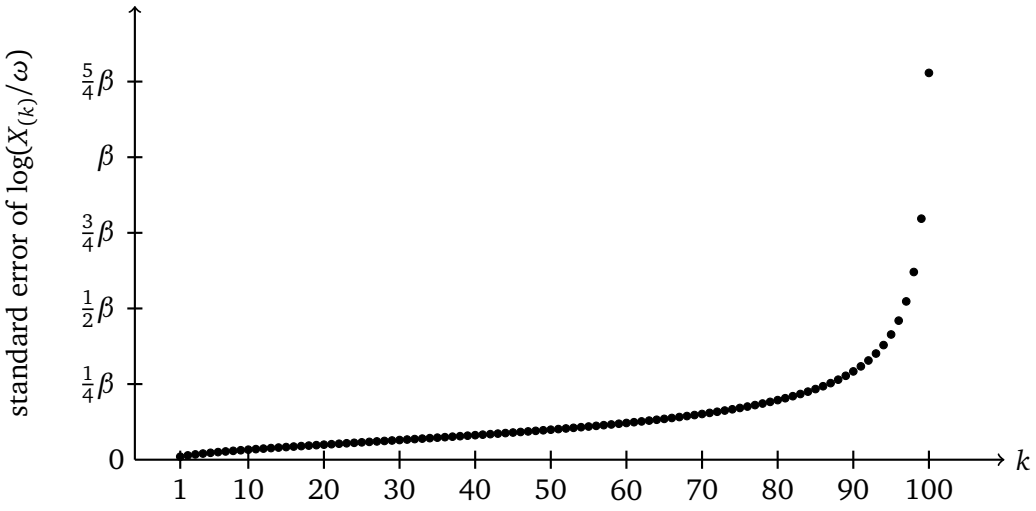


Figure 2.7: Standard error of $\log(X_{(k)}/\omega)$ for $n = 100$

The intuition behind that result is the following: $X_{(1)}$ cannot be much higher than its expected value, because by definition, it must be lower than $X_{(2)}$. In other words, for $X_{(1)}$ to be significantly above average, so must $X_{(2)}$, $X_{(3)}$ and so on: the entire sample needs to be unexpectedly high, which makes it very unlikely. That constraint is very progressively relaxed as we consider higher order statistics, until the last one for which no constraint is present.

Heteroscedasticity is well-known to make the OLS estimator inefficient in regressions. The intuition here is exactly the same: order statistics with low variance are more precisely located

on Pareto's diagram, and as such they should be given more weight. Ideally, the weight of observations should be inversely proportional to their standard error. The dependence between order statistics creates the same kinds of issues.

2.2.2 The Generalized Least Squares estimator

I now introduce the GLS estimator of the tail index, which addresses both flaws of the simple estimator. We will regress $\log(X_{(k)}/\omega)$ against $H_n - H_{n-k}$ instead of $\log\left(\frac{n+1}{n-k+1}\right)$ to remove the bias, and we will use GLS instead of OLS to make the estimator efficient.

For simplicity, we move to matrix notation. Consider the column vectors $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$\boldsymbol{y} = \begin{bmatrix} \log(X_{(1)}/\omega) \\ \log(X_{(2)}/\omega) \\ \vdots \\ \log(X_{(n)}/\omega) \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} H_n - H_{n-1} \\ H_n - H_{n-2} \\ \vdots \\ H_n - H_0 \end{bmatrix}$$

and the matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} H_n^{(2)} - H_{n-1}^{(2)} & H_n^{(2)} - H_{n-1}^{(2)} & \cdots & H_n^{(2)} - H_{n-1}^{(2)} \\ H_n^{(2)} - H_{n-1}^{(2)} & H_n^{(2)} - H_{n-2}^{(2)} & \cdots & H_n^{(2)} - H_{n-2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ H_n^{(2)} - H_{n-1}^{(2)} & H_n^{(2)} - H_{n-2}^{(2)} & \cdots & H_n^{(2)} - H_0^{(2)} \end{bmatrix}$$

Equations (2.15) and (2.16) imply:

$$\mathbb{E}[\boldsymbol{y}] = \beta\boldsymbol{x} \qquad \text{and} \qquad \text{Var}(\boldsymbol{y}) = \mathbb{E}[(\boldsymbol{y} - \beta\boldsymbol{x})(\boldsymbol{x} - \beta\boldsymbol{y})'] = \beta^2\boldsymbol{\Sigma}$$

The GLS estimor of β is:

$$\hat{\beta}^{\text{GLS}} = \underset{\beta}{\text{argmin}}\,(\boldsymbol{y} - \beta\boldsymbol{x})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \beta\boldsymbol{x}) \tag{2.21}$$

The GLS procedure was introduced by Aitken (1936) as a way to derive an optimal estimator for linear regression models. It is, however, quite unfashionable among econometricians because the covariance matrix $\boldsymbol{\Sigma}$ is generally unknown, so it has to be estimated. That procedure gives rise to the so-called feasible generalized least squares (FGLS) estimator, and although it is asymptotically equivalent to GLS, it may have worse finite sample properties than OLS, and needs stronger assumptions to ensure consistency (Wooldridge, 2010, p. 176). Here, however, we are in one of the rare cases where $\boldsymbol{\Sigma}$ is known up to a multiplicative constant, so that the improvement in efficiency comes at no cost.

Like OLS, the GLS estimator minimizes the distance between $\boldsymbol{y}$ and $\beta\boldsymbol{x}$. But instead of using the Euclidean distance, it uses the so-called Mahalanobis distance, which is defined as the

quadratic form $\boldsymbol{v} \mapsto \boldsymbol{v}'\boldsymbol{\Sigma}^{-1}\boldsymbol{v}$. This choice can be motivated as follows. First, notice that $\boldsymbol{\Sigma}$, like any covariance matrix, is symmetric positive definite. We can therefore factorize $\boldsymbol{\Sigma}$ as:

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$$

where $\boldsymbol{\Sigma}^{1/2}$, the square root of $\boldsymbol{\Sigma}$, is also a symmetric positive definite matrix. Then, rewrite (2.21) as:

$$\hat{\beta}^{\text{GLS}} = \underset{\beta}{\operatorname{argmin}} [\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \beta \boldsymbol{x})]'[\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{y} - \beta \boldsymbol{x})]$$

The GLS thus amounts to multiplying the data by $\boldsymbol{\Sigma}^{-1/2}$ before using OLS. Interestingly, the covariance matrix of the new, transformed data is given by:

$$\begin{aligned}
\text{Var}(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}) &= \boldsymbol{\Sigma}^{-1/2}\,\text{Var}(\boldsymbol{y})\,\boldsymbol{\Sigma}^{-1/2} \\
&= \beta^2\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2} \\
&= \beta^2\boldsymbol{I}_n
\end{aligned}$$

which, up to a multiplicative constant, is the identity matrix. Under those conditions, the Gauss-Markov theorem applies and states that $\hat{\beta}^{\text{GLS}}$ is the best linear unbiased estimator (BLUE).

Here, the GLS estimator eventually simplifies to (see appendix C for proof):

$$\hat{\beta}^{\text{GLS}} = \frac{1}{n}\sum_{k=1}^{n} \log\left(\frac{X_k}{\omega}\right)$$

This estimator is exactly what we get using MLE (see appendix C). Therefore, for Pareto distributions, MLE does not just have good asymptotic properties, it also inherits the excellent finite sample properties of GLS estimation.

2.3 Non-parametric estimation

The previous sections showed how to estimate the tail index of a Pareto distribution. We will now extend the approach developed above to the nonparametric case.

Why is such an extension desirable? Because the Pareto model rarely holds exactly. We usually think of it as the limit distribution (up to some renormalization) of $X|X > x$ when x goes to infinity. It is, therefore, an approximation that gets better and better as we consider higher order statistics. Now, remember that the case for the GLS/MLE estimator of the tail index relied on the idea that higher order statistics should be weighted less in the regression because of their higher variance. That argument ignores that higher order statistics may follow the Pareto distribution more closely. Hence, by weighting them less, we decrease the variance, but at the cost of increasing the misspecification bias.

Because it gives "too much" weight to extreme order statistics, the simple estimator may, as a collateral benefit, have better statistical properties when the Pareto model holds only as a

limiting distribution of the tail. Of course, if we make that argument, then we should find what weighting scheme that optimally balances the trade-off between bias and variance. That is implicitly the road taken by Csorgo, Deheuvels, and Mason (1985) with their kernel estimator of the tail index.

I will take another direction, more suited to our setting, namely to estimate the distribution nonparametrically. To estimate the tail of a distribution nonparametrically might seem like a hopeless endeavor, given that there are by definition very few data points in that part of the distribution. But we can do it here because we have access to the wealth rankings (see chapter 3), which give information on the very last order statistics, beyond the survey sample. Hence, we can "interpolate" between the survey and the rankings to impute the wealth of the missing rich: those too poor to be in the rankings, but too rich to be in the survey.

Chapter 3 will explain in detail how to actually combine the survey samples and the rankings. For now, we still assume and iid sample.

2.3.1 The tail function

Let X be an absolutely continuous random variable with support $[\omega, +\infty[$ and a strictly increasing CDF. Recall from (2.14) that if it followed a Pareto distribution with tail index β, we would have:

$$\log\left(\frac{x}{\omega}\right) = \beta \times (-\log \mathbb{P}\{X \geq x\})$$

In a nonparametric context, we can always rewrite the relationship above as:

$$\log\left(\frac{x}{\omega}\right) = \phi(-\log \mathbb{P}\{X \geq x\}) \tag{2.22}$$

where $\phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is a strictly increasing, differentiable function such that:

$$\phi(0) = 0 \qquad \text{and} \qquad \lim_{x \to +\infty} \phi(x) = +\infty$$

Put another way, we replaced the linear relationship between $\log(x/\omega)$ and $-\log \mathbb{P}\{X \geq x\})$ by an arbitrary one. By analogy with the tail index β, I will call ϕ the tail function. The Pareto distribution corresponds to the special case where $\phi : x \mapsto \beta x$.

The tail function as a characterization tool

The tail function characterizes any absolutely continuous random variable with support $[\omega, +\infty[$ and a strictly increasing CDF. Indeed, by construction, ϕ is invertible, so that we can write the CDF using (2.22):

$$F(x) = 1 - \exp\left[-\phi^{-1}\left(\log\left(\frac{x}{\omega}\right)\right)\right] \tag{2.23}$$

and conversely, we can define ϕ using the quantile function $Q = F^{-1}$:

$$\phi(x) = \log\left[\frac{Q(1 - e^{-x})}{\omega}\right] \tag{2.24}$$

We will therefore use $\mathscr{D}(\omega, \phi)$ to denote the distribution characterized by the support $[\omega, +\infty[$ and the tail function ϕ. We get the PDF by differentiating (2.23):

$$f(x) = \frac{\exp\left[-\phi^{-1}\left(\log\left(\frac{x}{\omega}\right)\right)\right]}{x\phi'\left(\phi^{-1}\left(\log\left(\frac{x}{\omega}\right)\right)\right)}$$

and the quantile function by inverting (2.24):

$$Q(x) = \omega \exp\left[\phi(-\log(1-x))\right]$$

The point of the tail function is to get a characterization of a distribution that behaves more nicely in the tail than the PDF, the CDF or the quantile function. For any "Paretoesque" distribution, the tail function will be close to, if not exactly, a linear function.

The tail function and the exponential distribution

Any random variable $X \sim \mathscr{D}(\omega, \phi)$, can easily be linked to the exponential distribution:

$$X \sim \mathscr{D}(\omega, \phi) \Leftrightarrow \phi^{-1}\left[\log\left(\frac{X}{\omega}\right)\right] \sim \mathscr{E}(1) \tag{2.25}$$

Indeed, we have for all $x \geq 0$:

$$\begin{aligned}
\mathbb{P}\left\{\phi^{-1}\left[\log\left(\frac{X}{\omega}\right)\right] \geq x\right\} &= \mathbb{P}\left\{X \geq \omega \mathrm{e}^{\phi(x)}\right\} \\
&= \exp\left[-\phi^{-1}\left(\log\left(\frac{\omega \mathrm{e}^{\phi(x)}}{\omega}\right)\right)\right] \\
&= \mathrm{e}^{-x}
\end{aligned}$$

which characterizes the exponential distribution. In the strict Pareto case, that property collapses to:

$$X \sim \mathscr{P}(\omega, \beta) \Leftrightarrow \frac{1}{\beta}\log\left(\frac{X}{\omega}\right) \sim \mathscr{E}(1)$$

which means, as we already saw, that the logarithm of a Pareto distribution is an exponential distribution.

2.3.2 Estimation

Let $(X_1, X_2, \ldots, X_n) \overset{\text{iid}}{\sim} \mathscr{D}(\omega, \phi)$. The definition (2.22) of the tail function strongly suggests using Pareto's diagram for its estimation. The idea would be to estimate ϕ using a nonparametric regression in the same way that we used linear regression to estimate the tail index of a Pareto distribution.

Actually, that generalization is not as immediate as we might hope, for the same reasons that led to the bias of the simple estimator (section 2.2.1). To properly identify ϕ as a regression function, the problem must be written as:

$$\mathbb{E}\left[\log\left(\frac{X_{(k)}}{\omega}\right)\right] = \phi(r_k)$$

where r_k is a value to be determined. In the strict Pareto case ($\phi : x \mapsto \beta x$), the simple estimator was biased because it assumed $r_k = \log\left(\frac{n-k+1}{n+1}\right)$ instead of the correct answer $r_k = H_n - H_{n-k}$.

Here, however, $r_k = H_n - H_{n-k}$ will not work in general. Let $(Z_1, Z_2, \ldots, Z_n) \overset{\text{iid}}{\sim} \mathcal{E}(1)$. On the one hand, we get from property (2.25):

$$\mathbb{E}\left[\log\left(\frac{X_{(k)}}{\omega}\right)\right] = \mathbb{E}[\phi(Z_{(k)})]$$

On the other hand:

$$\phi(\mathbb{E}[Z_{(k)}]) = \phi(H_n - H_{n-k})$$

Because of the functional non-invariance of the expected value, $\mathbb{E}[\phi(Z_{(k)})] \neq \phi(\mathbb{E}[Z_{(k)}])$ in general, so that $r_k = H_n - H_{n-k}$ does not work (except for the strict Pareto case where ϕ is linear). More generally, r_k will always depend on the distribution of X, so we cannot use it in practice in order to estimate ϕ.

We could argue that the bias created by choosing the wrong r_k might be small enough to be tolerable: after all, that was true of the simple estimator, whose the bias disappeared asymptotically. As discussed in David and Nagaraja (2005, p. 85), and as we observed for the Pareto case in section 2.2.1, the approximation we would make is asymptotically good for middle range order statistics, but not for extreme ones. In the parametric setting, the simple estimator was consistent because middle range order statistics carried increasingly more weight in the estimator as the sample size increased. That kind of argument does not work in a nonparametric setting since we will always find ourselves estimating part of the tail function based solely on extreme order statistics. In that part of the distribution the estimator would not just be biased, it would be inconsistent.

One way to solve the problem is to drop the expected value operator in favor of an operator that satisfies a functional invariance property, like the quantile. Technically, any quantile will do, but the median is the most natural choice. Because ϕ is strictly increasing, we get:

$$\begin{aligned}\text{Med}\left[\log\left(\frac{X_{(k)}}{\omega}\right)\right] &= \text{Med}[\phi(Z_{(k)})] \\ &= \phi(\text{Med}[Z_{(k)}]) \\ &= \phi(-\log[1 - I^{-1}_{1/2}(k, n-k+1)])\end{aligned} \tag{2.26}$$

using equation (2.17). Hence we can estimate ϕ with a nonparametric median regression.

Nonparametric quantile regression

In order to estimate ϕ in practice, we still need to specify a way to perform a nonparametric median regression. Two solutions exist (see Tsybakov (2009), for example). The first one is called local polynomial fitting.[2] At each point x, it locally approximates ϕ using a polynomial,

[2] Of which the Nadaraya-Watson estimator and local linear fitting are special cases.

which corresponds to the truncated Taylor expansion around x. That polynomial is estimated like any linear model, but giving more weights to the observations that are closer to x. This approach is feasible: it has been tested on the data and gives correct results. But it has significant drawbacks. First, it can be computationally intensive because it requires performing a new regression every time we need to evaluate $\phi(x)$. Second, it makes it fairly difficult to impose *a priori* constraints on ϕ, such as derivability, monotonicity or $\phi(0) = 0$. Third, it can exhibit erratic behavior near the boundaries of the data, and offers no reasonable fallback solution outside of their range.

The other solution, which I will use, is called projection estimation. It approximates ϕ globally by its projection over a sufficiently flexible function space of finite dimension. That is, if we consider a family $\mathscr{B} = \{f_1, f_2, \dots, f_K\}$ of functions, it approaches ϕ by an element of:

$$\operatorname{span}(\mathscr{B}) = \left\{ \sum_{k=1}^{K} \theta_k f_k : (\theta_1, \theta_2, \dots, \theta_K) \in \mathbb{R}^K \right\}$$

where the coefficients θ_k can be estimated by a multiple linear regression. This solution requires performing only one regression, and the properties of the estimate can easily be controlled by adjusting $\{f_1, f_2, \dots, f_K\}$.

For $\mathscr{B}$, I will use natural cubic splines expressed in the restricted power basis. Let $M = K - 1$. They are defined for $x \geq 0$ by $f_1(x) = x$ and:

$$\forall m \in \{0, 1, \dots, M-1\} \qquad f_{m+2}(x) = d_m(x) - d_M(x)$$

where:

$$\forall m \in \{0, 1, \dots, M\} \qquad d_m(x) = \frac{(x - \xi_m)_+^3 - (x - \xi_{M+1})_+^3}{\xi_{M+1} - \xi_m}$$

and $\xi_0 = 0 < \xi_1 < \xi_2 < \cdots < \xi_{M+1}$ are called the knots of the spline. Any function g that can be written as a linear combination of those functions satisfy the following properties:

(i) $g(0) = 0$

(ii) g is a cubic polynomial over $[\xi_m, \xi_{m+1}]$ for $m \in \{0, 1, \dots, M\}$

(iii) g is a linear polynomial over $[\xi_{M+1}, +\infty[$

(iv) g is twice continuously differentiable over $\mathbb{R}^+$

For simplicity, we do not explicitly impose the constraint that ϕ is strictly increasing. Indeed, order statistics are already sorted by construction, so that the function we estimate spontaneously turns out to be strictly increasing in virtually all cases. Explicitly imposing $\phi' > 0$ would be superfluous because such constraint would almost never be binding.

The natural cubic spline provides a functional form that is flexible, smooth and numerically stable. Beyond the last knot, the linear behavior implies that the data follow a Pareto distribution, which seems to be the most reasonable option.

One question remains, namely the choice of the knots $\xi_1, \xi_2, \ldots, \xi_{M+1}$ (ξ_0 is always set at 0). On the one hand, if we choose too few knots, the function ϕ will be poorly approximated, resulting in biased estimations (oversmoothing).[3] On the other hand, if we choose too many knots, the estimate of ϕ will have a large variance (undersmoothing). We face bias/variance trade-off, and the appropriate choice of knots should optimally balance both objectives.

Ideally, that choice should itself be the result of an estimation procedure. But so far, there is no proper way of doing so in the present context. In the simple case of iid data, some solutions exist for optimal smoothing in quantile regressions (Yu and Jones, 1998; Abberger, 1998; Abberger, 2001). But they do not apply here because order statistics exhibit positive serial correlation, which leads to undersmoothing if not taken into account. Indeed, if one observation is significantly above its true mean, then the serial dependence will make the next observations above average too, and a naive nonparametric estimator would wrongly interpret those successive deviations as a sign that the true mean has changed (Opsomer, Wang, and Yang, 2001).

In the absence of appropriate data-driven method, the choice of the knots is left to our discretion. In practice, it rarely seems necessary to go beyond five or six knots in total. I will give more details about this choice in chapter 3.

Given a set of knots $(\xi_0, \xi_1, \ldots, \xi_{M+1})$, we can proceed to the estimation of ϕ. Define $\hat{\phi}$ the estimator. We have:

$$\hat{\phi}(x) = \sum_{k=1}^{K} \hat{\theta}_k f_k(x) \tag{2.27}$$

The coefficients $\hat{\theta}_k$ can be estimated by a linear quantile regression. Based on (2.26) and (2.27), we solve the following ℓ_1 minimization problem:

$$(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K) = \underset{(\theta_1,\theta_2,\ldots,\theta_K)\in\mathbb{R}^K}{\operatorname{argmin}} \sum_{r=1}^{n} \left| \log\left(\frac{X_{(r)}}{\omega}\right) - \sum_{k=1}^{K} \hat{\theta}_k f_k(-\log[1 - I^{-1}_{1/2}(r, n-r+1)]) \right|$$

That problem can be recast as a linear one, solved using either simplex or interior point methods. That procedure is implemented in most statistical software packages, and yields the desired estimates for $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K)$. We finally obtain an estimate for ϕ using (2.27).

[3] Choosing zero knots would put us back in the strict Pareto case.

Chapter 3

Wealth inequality in Europe and in the United States

This last chapter combines the data of the chapter 1 with the methodology of chapter 2 to provide new, improved estimates of wealth inequality. These new estimates take better account of the top tail of the distribution, while remaining consistent with the national accounts.

Section 3.1 will explore the different public rankings of top wealth holders that are available. They will provide precious information on the very top of the tail of the wealth distribution, thus allowing us to use the methodology of chapter 2 effectively. Section 3.2 will show how that methodology can be applied to our specific setting, and section 3.3 will finally give the results.

3.1 Wealth rankings

In the last decades, a number of newspapers have started to regularly publish lists of top wealth holders. Journalists compile them based on a mix of publicly available information on stock holding, insider knowledge, and educated guesses. They are a very valuable source of information, because they go precisely where surveys do not: the very top of the distribution.

At the same time, they have issues. They are, to a large extent, black boxes. Journalists do write a few words on how they proceed, but the details are kept secret. The methodology may also not be fully comparable from one list to the next. They may also have biases because they tend to privilege the most visible types of wealth, and ignore the rest.

Those limitations must be acknowledged. But the rankings, for all their imperfections, are better than no information at all. In practice, they appear to convincingly extend the survey data. Importantly, our estimates will not overly rely on the precise distribution of wealth in the rankings. They rather serve as an anchor point, a piece of information that tells us where the tail of the distribution is heading.

3.1.1 Overview of rankings

The most famous ranking of wealth is probably the one published by *Forbes*, with covers the entire world. Because it exclusively covers billionaires, it often includes very few individuals from a specific country. Rankings published by national magazines can offer a better coverage because they go further down the distribution.

Forbes (worldwide)

Forbes started to publish a list of the 400 wealthiest Americans in 1983, and called it the *Forbes 400*. In 1987, they extended their efforts to the entire world, and started a census of all billionaires worldwide (as measured in US dollars), called the *The World's Billionaires*.

More than 50 reporters in 16 countries work on these lists. They interview sources close to candidates (employees, attorneys, rivals), they keep track of major transactions and charitable donations. Some people even cooperate directly. The figures try to account for financial assets (public and private companies), real assets (real estate, yachts, paintings, etc.) and debt (Dolan, 2012).

For the United States, the *Forbes* rankings are the only ones at our disposal.[1] They actually provide excellent coverage of the top tail of the distribution. We also use *The World's Billionaires* for Italy and Spain despite much smaller samples sizes, because no better choices were available.

Challenges (France)

The weekly French magazine *Challenges* has published a list of France's 500 biggest fortunes since 1996. The list focuses on gross financial wealth, which at this point of the distribution should arguably be very close to net wealth. Journalists look first a stock market data, but fortunes built on public company stocks only represent a third of the ranking. They track the rest based on an examination of professional publications, seminars, award ceremonies and similar events. They value private companies by comparing them with publicly traded ones, based on their balance sheets. Finally, they send letters to potential members of the list asking if they want to suggest improvements or clarifications, and some do cooperate (Treguier, 2012).

Manager Magazin (Germany)

Manager Magazin has compiled a list Germany's 300 richest people since 2000, and has extended it to the 500 richest since 2010. They estimate net wealth using information from archives, registers, stock markets, lawyers, asset managers and the people themselves. Some asked to be removed for privacy or security reasons (Bach, Thiemann, and Zucco, 2015).

[1] A competing list exists, *Bloomberg Billionaires*. But it only started in 2012, after the period we are studying. See `http://www.bloomberg.com/billionaires/`.

Trend (Austria)

The business magazine *Trend* publishes a list of Austria's 100 wealthiest people. They do not appear to give any information on the methodology.

Exame (Portugal)

The Portuguese magazine *Exame* publishes a list of the 25 wealthiest people in Portugal. They value the financial assets based on market value for listed companies, and using "conservative" estimates for unlisted ones (Exame, 2009).

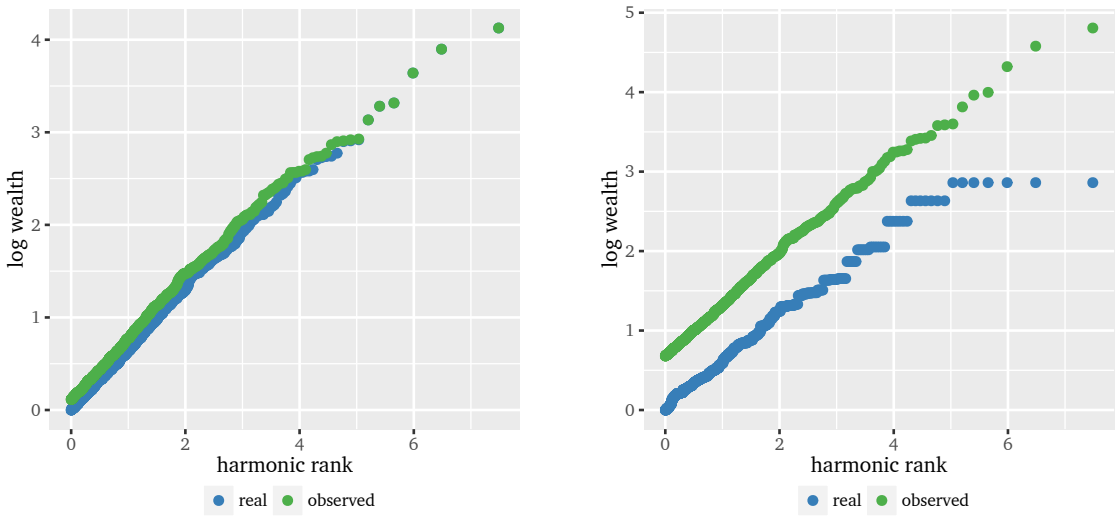
3.1.2 Comparability and corrections

Not all rankings of wealth are directly comparable. In general, national rankings find higher levels of wealth than the *Forbes* list. In countries where both are available, if we were to run estimates with the data as they are published, we would find higher levels of wealth inequality with national lists than with *Forbes*. That can be attributed to two problems. First, *Forbes* (like the surveys) considers the country of residence of individuals, while national lists also include citizens living abroad. It seems to be the case of about 20% of Challenges' ranking, for example. Second, most entries in *Forbes* refer to individuals. Some refer to families, but they are the exception, not the rule. National lists have a much bigger tendency to aggregate the wealth of families. The first entry of the *Trend* list is a spectacular example: it corresponds to the Porsche and Piëch families, for a total wealth of €34bn in 2010. That fortune, however, is spread among about a hundred family members.

The first problem has actually a minor impact on the results. To understand why, consider the stylized model plotted in figure 3.1a. Wealth follows a Pareto distribution with tail index 2/3. I simulate $n = 1000$ observations. I add 20% of observations to the sample to simulate the presence of individuals that are unwanted because they reside in another country. Then I plot a Pareto diagram of the last n observations for both the original sample and the sample with additional observations. As we can see, the difference is very barely noticeable.[2] That can be compared to figure 3.1b, which simulates in a similar way a distribution of wealth with families of seven members who share their wealth equally. The impact is much stronger and can thus change estimates significantly. For that reason, I will overlook the first issue and focus on the second. As a robustness check, I provide estimates of top shares for all countries with 20% of the wealth rankings removed in table 3.5. As expected, it changes very little to the results.

Wealth rankings specify whether an entry refers to an individual or a family. But they never say how many members they are in the family. However, *Forbes* publishes another list, *America's Richest Families*, which lists wealthy families whose individual members do not necessarily make it to the *Forbes 400*. That list does give the number of people who share the family's wealth. I

[2] It also gets smaller as inequality increases: at the limit, where one individual gets all the wealth, the impact vanishes entirely.



(a) impact of country of residence

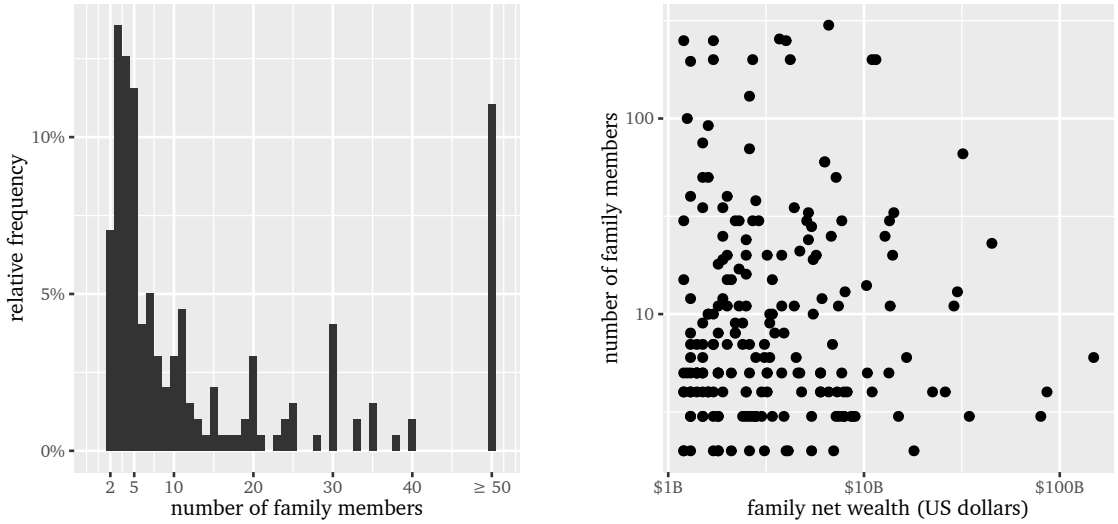
(b) impact of families

I simulated a Pareto sample of size $n = 1000$ with tail index 2/3. In figure (a), I added $0.2n$ observations with the same distribution to model citizens residing in another country. In figure (b), I split each data point in 7 points with equal value to model families. In each case, I plotted observed against their harmonic rank (see chapter 2) as if the sample size was n, and compared the "real" distribution with the "observed" one. Samples are normalized so that the "real" distribution starts at one.

Figure 3.1: Impact of country of residence and families on the wealth rankings

assume that this list is representative of wealthy families in all countries, and use it to impute a number of family members when needed. Figure 3.2a shows the distribution of family sizes. With a median of 7, and a mean of 25, it is highly skewed to the right. I exclude the DuPont family, which is a clear outlier with 3500 members. The resulting sample has a maximum value of 300, and little more than 10% of observations are above 50. Figure 3.2b shows on a scatter plot the absence of relationship between the size of the family and its wealth, so that we need not worry about larger families being richer (a linear regression gives a p-value of 0.934). Every time a ranking marks an entry as referring to a family, I draw with replacement from the distribution of figure 3.2a, and divide wealth randomly between members following a uniform distribution. That procedure introduces additional variability, which remains modest. I take it into account in the standard errors using the multiple imputation procedure of Rubin (1987), following in that regards a practice already in place in the surveys (see section 1.1.2): I generate five different corrected rankings and associate each of them to an implicate of the survey.

The imputation procedure can induce the presence of silently missing observations towards the bottom the ranking, because of some people who might have made it to the list as a family while being individually poorer than others who did not make it as persons. To avoid the problem, I systematically discard all observations that are below the last entry originally marked as referring to an individual.



(a) Distribution of the size of families

(b) Size of families and their wealth (log-log scale)

Source: Forbes (2015).

Figure 3.2: *America's Richest Families* list

Additionally, the rankings get less reliable when we look near the bottom, because journalists are more likely to miss people with relatively lower wealth. Given that the method does not require a large amount of data points, I only keep the top 50 of individuals. I make an exception with *Forbes* for two reasons. First, because it only tracks billionaires, who should arguably all be rich enough for their wealth to be traceable (as opposed to national rankings which need to go further down the distribution). Second, because we need to consider the lower part of their list to get information on the billionaires in Italy or Spain, who can be near the bottom globally even though they are at the top in their country. It would be inconsistent to just keep the top 50 in the United States while still using observations in lower parts of the ranking for other countries. Table 3.1 summarizes the sample sizes we get in the end.

country	year	source	observations
Austria	2010	Trend	32
France	2010	Challenges	50
Germany	2011	Manager Magazin	50
Italy	2011	Forbes	10
Portugal	2010	Exame	10
Spain	2009	Forbes	12
United States	2010	Forbes	373

Table 3.1: Final sample sizes for wealth rankings

The correction for families does not impact all rankings equally: for example, *Forbes* is mostly left unaffected, but *Challenges* goes down significantly. Appendix B shows graphically, for each country, the impact of the correction alongside the tail of the survey in Pareto diagrams.

3.2 Estimation procedure

Now that I have introduced all of the data, I will explain how the estimation procedure works in our specific setting. Chapter 2 explored theoretical considerations, but all within the context of a single, iid sample. Moreover, it paid no attention to standard errors. But here, not only do we have two samples (the survey and the ranking), none of which are iid, we also need to get some sense of the precision of the estimates.

3.2.1 Point estimation

Let N be the true size of the population of reference in a country. We view the distribution of wealth over that population as an iid sample of size N, and write it $(X_1, X_2, \dots, X_N)$. The wealth ranking simply correspond to the last m order statistics of that sample:

$$(X_{(N-m+1)}, X_{(N-m+2)}, \dots, X_{(N)})$$

Matters are somewhat more complicated for the survey, because we do not know exactly the true population rank of survey observations. But we can estimate it, using the fact that order statistics can be viewed as sample estimates of the quantile.

Let n be the size of the survey sample. We consider standard asymptotics for that kind of problem, namely $n \to +\infty$, $N \to +\infty$, and $n/N \to 0$. Let $\hat{F}_N$ be the empirical CDF over the true population. Denote f the PDF and Q the quantile function of the underlying distribution. With $r = \lfloor Np \rfloor + 1$, the so-called Bahadur (1966) representation of the quantile states:

$$X_{(r)} = Q(p) + \frac{\hat{F}_N(Q(p)) - p}{f(Q(p))} + o(N^{-1/2}) \tag{3.1}$$

Francisco and Fuller (1991) extended that result to complex survey designs. Denote $(Z_1, \dots, Z_n) \subset (X_1, \dots, X_N)$ the survey sample, and $(w_1, \dots, w_n)$ the associated survey weights.[3] Consider:

$$s = \min \left\{ k \in \{1, \dots, n\}; \sum_{i=1}^{k} w_{(i)} \geq Np \right\}$$

And let $\tilde{F}_n$ be the empirical CDF over the survey data. Under some technical conditions, we have:

$$Z_{(s)} = Q(p) + \frac{\tilde{F}_n(Q(p)) - p}{f(Q(p))} + o(n^{-1/2}) \tag{3.2}$$

Using both formulas, we can find a true population rank relevant to the survey observation $Z_{(s)}$. Let $\lambda \in]0, 1]$, a parameter offering some leeway reflecting the fact that the quantile function is not unique in finite samples. Define:

$$p = \frac{1}{N} \left[\sum_{i=1}^{s-1} w_{(i)} + \lambda w_{(s)} \right]$$

[3] I assume, as it is always the case in practice, that the survey weights are calibrated so that $\sum_{i=1}^{n} w_i = N$.

and let $r = \lfloor Np \rfloor + 1$. With those definitions, formulas (3.1) and (3.2) both apply. Since $n/N \to 0$, the term $o(N^{-1/2})$ is negligible compared to $o(n^{-1/2})$, so we get:

$$Z_{(s)} - X_{(r)} = \frac{\tilde{F}_n(Q(p)) - p}{f(Q(p))} + \frac{\hat{F}_N(Q(p)) - p}{f(Q(p))} + o(n^{-1/2})$$

Moreover, $\hat{F}_N(Q(p)) - p = O(N^{-1/2})$, so the second term is negligible too:

$$Z_{(s)} - X_{(r)} = \frac{\tilde{F}_n(Q(p)) - p}{f(Q(p))} + o(n^{-1/2}) \tag{3.3}$$

Francisco and Fuller (1991) also showed $\tilde{F}_n(Q(p)) - p = O(n^{-1/2})$, hence we get the following convergence:

$$Z_{(s)} - X_{(r)} = O(n^{-1/2})$$

Additionally, it follows from (3.3) that the asymptotic distribution of $n^{1/2}(Z_{(s)} - X_{(r)})$ is the same as that of $n^{1/2}(\tilde{F}_n(Q(p)) - p)/f(Q(p))$. Therefore:

$$n^{1/2}(Z_{(s)} - X_{(r)}) \to \mathcal{N}(0, \sigma^2)$$

where σ^2 is a complex variance term that depends on the survey design, and whose expression is given by Francisco and Fuller (1991). $Z_{(s)}$ is a consistent and asymptotically normal estimator of $X_{(r)}$. For some $\lambda \in]0, 1]$, we will therefore replace $X_{(r)}$ by $Z_{(s)}$ in the estimation, where:

$$r = \left\lfloor \sum_{i=1}^{s-1} w_{(i)} + \lambda w_{(s)} \right\rfloor + 1 \tag{3.4}$$

The choice of λ is asymptotically irrelevant. I use $\lambda = 1/2$ because it correspond to the middle of the interval. Through intuitive considerations, Vermeulen (2014) and Bach, Thiemann, and Zucco (2015) chose $\lambda = 0$ and replace $x \mapsto \lfloor x \rfloor + 1$ by the identity function. That has no perceptible impact on the results.

There is one catch: the result holds for s fixed, and $n \to +\infty$. The approximation may therefore be bad for the last observations. This is not a major concern as the whole point of the approach is to rely on the ranking for extreme values, not the survey.

In the end, we can view both samples as subsets of the true population order statistics. The rankings correspond to the last ones, and the survey to the middle ones, whose order can be approximated by formula 3.4.

Having only a selection of all order statistics changes nothing to the method of chapter 2: we just ignore missing observations in the regression. To apply it, we must first decide on a threshold above which to model the distribution. Because the estimate is nonparametric, that has no consequences on the results. I always model the top 20% of the distribution.

Then we must choose the knots of the spline. The first one is always set at zero. I place the last one at the very beginning of the ranking. It means that in practice, wealth is forced to follow a

Pareto law once we arrive in that part of the distribution. There is no better choice since the number of observations there is generally too low to estimate a more flexible model. It is also reasonable since the ranking represents at most the last few hundreds observations out of a population of tens of millions. If the Pareto distribution is the limiting distribution of the tail, then we should at least expect it to be a reasonable model here. In the survey sample, I set knots as indicated in table 3.2, to provide a good fit and sufficiently precise estimates. In Austria, given the imprecision of the data, I used the sparsest model possible, with only three knots in total.

country	quantile for the last knot in the survey tail	number of knots in the survey tail	total number of knots
Austria	95%	2	3
France	99%	5	6
Germany	95%	3	4
Italy	95%	3	4
Portugal	95%	3	4
Spain	95%	3	4
United States	95%	4	5

In the survey part of the tail, I place the last knot at the unweighted quantile indicated in the second column. The other ones are equally spaced between that knot and the first one. I place the very last one where the ranking starts.

Table 3.2: Position of the knots of the spline

A final concern relates to the adjustment of data to the national accounts, which was performed in chapter 1. When we model the tail, we typically increase mean wealth. With adjusted data, it implies that total wealth will be above the national accounts: in other words, we will overcorrect survey data. To ensure that totals are still consistent with the national accounts, I use an iterative procedure similar to Vermeulen (2016). I estimate the model on the adjusted data, and calculate the value of aggregate wealth, which should be higher than the national accounts. I scale down wealth in the survey to match again the national accounts, and re-estimate the model on the new data. I repeat the process until convergence of the estimated total towards the national accounts total. In theory, there is no guarantee that it will converge, and as a matter of fact, it sometimes doesn't. However, the final relative difference is always below 0.5% in practice, and most of the time below 0.1%.

3.2.2 Standard errors and hypothesis testing

Both the SCF and the HFCS have complex survey designs, to which researchers don't have access for confidentiality reasons. As a consequence, there is no way to get explicit formulas for the variance of even the simplest estimators. That is *a fortiori* the case for the present estimator, which is fairly complex.

Bootstrap is a good solution to that kind of problem. Standard bootstrap fails with complex survey designs (Kolenikov, 2010), but this is why both the SCF anf the HFCS implement the

rescaling bootstrap procedure of Rao and Wu (1988), and provide a set of so-called replicate weights that enable proper inference (see section 1.1.2).

We also need to take the rankings into account. Standard bootstrap will also fail in that situation because the last m order statistics of the population is not an iid sample either. Instead, for each bootstrap replication, I simulate an entirely new ranking as follows. Let $\hat{\phi}$ be the estimated tail function, and consider the last m order statistics from a uniform sample of size N: $(U_{(N-m+1)}, \dots, U_{(N)})$. Define for all $k \in \{N-m+1, \dots, N\}$:

$$X^*_{(k)} = \omega \exp\left[\hat{\phi}(-\log(1-U_{(k)}))\right]$$

Given properties (2.2) and (2.25), the sample $(X^*_{(N-m+1)}, \dots, X^*_{(N)})$ correspond to the last m order statistics of the estimated distribution of wealth, and can thus be used as a bootstrap replication sample.

That method requires simulating the last order statistics from a uniform sample. A naive solution would be to simulate a sample of size N, and then discard all observations but the last m. Given that $N \gg m$, it would a huge waste of computing power. A much better approach is to simulate those last order statistics directly, with the method of Schucany (1972). Generate a uniform iid sample of size m, $(V_1, \dots, V_m)$, set $U_{(N)} = V_1^{1/N}$, and then recursively set $(U_{(N-1)}, \dots, U_{(N-m+1)})$ for all $k \in \{1, \dots, m-1\}$ using:

$$U_{(N-k)} = U_{(N-k+1)} V_k^{1/(N-k)}$$

The resulting sample corresponds to a draw of the last m order statistics out of a sample of size N.

Interestingly, once we know the variance of the estimator, we can devise a simple test of the Pareto shape. Remember that $\hat{\phi}$ is estimated as a spline, which is a linear combination of the functions $\{f_1, \dots, f_K\}$ with coefficients $\{\theta_1, \dots, \theta_K\}$. By construction, f_1 is the linear component of the spline, and $f_2, \dots, f_K$ represent deviations from linearity. For a Pareto distribution, the tail function is a straight line. That shape can be tested through the null hypothesis:

$$H_0 : \theta_2 = 0, \dots, \theta_K = 0$$

Under asymptotic normality, we have:

$$\sqrt{n}(\hat{\theta}_2 - \theta_2, \dots, \hat{\theta}_K - \theta_K) \sim \mathcal{N}(0, \mathbf{\Sigma})$$

where $\mathbf{\Sigma}$ can be estimated via bootstrap.[4] If H_0 is verified, then asymptotically:

$$T = \frac{1}{n}(\hat{\theta}_2, \dots, \hat{\theta}_K)\hat{\mathbf{\Sigma}}^{-1}(\hat{\theta}_2, \dots, \hat{\theta}_K)' \sim \chi^2_{K-1}$$

To test the Pareto shape, we compare the test statistic T with typical values of a χ^2 distribution with $K-1$ degrees of freedom. That test has some interesting features compared to other, more

[4] $\mathbf{\Sigma}$ also needs to take multiple imputation into account.

standard goodness of fit tests. The Kolmogorov-Smirnov test, in particular, has been used in the past to disprove that wealth at the top followed a Pareto law based on *Forbes* rankings (e.g. Clauset, Rohilla, and Newman, 2007). But *Forbes* only publishes rounded figures, to which the Kolmogorov-Smirnov test is known to be very sensitive: rejection could simply a product of that fact. The new test is arguably less subject to that kind of problem because it relies on the global fitting of a curve via quantile regression, which makes it less sensitive to specific observations.

	without adjustment		with adjustment	
	threshold	p-value	threshold	p-value
Austria	194 100	0.482	225 000	0.420
France	200 700	< 0.001	302 500	0.032
Germany	153 600	0.304	200 500	0.070
Italy	211 100	0.076	288 200	0.325
Portugal	100 400	0.042	129 900	0.008
Spain	227 300	0.013	349 100	0.109
United States	185 300	< 0.001	144 200	< 0.001

Constant 2010 euros at market exchange rate. The threshold corresponds to the 80% quantile. "adjustment" refers to the rescaling of assets to match the national accounts.

Table 3.3: Test of the Pareto shape for the top 20% of the distribution

Table 3.3 shows the results of that test on the data, both adjusted and non-adjusted to the national accounts. The cases where the Pareto hypothesis is not rejected (at the traditional confidence levels) are Austria (although it largely reflects lack of power given the imprecision of the survey), Germany (for unadjusted data), Italy (for adjusted data) and, to a lesser extent, Spain (for adjusted data). Otherwise, the hypothesis is always rejected at the 10% level, and sometimes much more strongly, like in the United States.

3.3 Results

I estimate the top tail of the distribution for each country separately: appendix A provides details. In particular, I plot the tail function alongside the survey data and the wealth rankings to give a visual assessment of the quality of the fit.

Once the tail is estimated, the full distribution can be characterized piecewise by a mixture model: the bottom 80% of the distribution is represented by raw survey observations, and the top 20% by the estimated tail function. Using the tail function, we get the quantile function Q from formula (2.24). Then, I calculate expected values by numerical integration of the quantile function. The top p% share is given by:

$$\frac{\int_h^1 Q(x)\,\mathrm{d}x}{\int_0^1 Q(x)\,\mathrm{d}x}$$

where $h = 1 - p/100$.

3.3.1 Individual countries

Table 3.4 shows top wealth shares for each country. The first three columns correspond to the original survey data, which I give as a reference point. The next three columns also give standard survey estimates, but after adjustment to the national accounts. That first half of the table is essentially a repeat of table 1.10, with the top 0.1% added to it. The three columns after that include the tail correction with the rankings, but no adjustment to match national account totals. The last three columns give the final estimates, which include tail correction and are consistent with the national accounts.

Final estimates show significantly higher inequalities than the original surveys. The increase is mostly due to the estimation of the tail, not to the adjustment to the national accounts, with the exception of the United States, where both corrections contribute equally. In Germany the adjustment to national accounts actually decreases inequality.

The upward revision of top shares is particularly strong for the top 0.1%. That is not surprising, as extreme top shares estimates can suffer from a strong downward bias even in reasonably large samples (Taleb and Douady, 2015), and the number of observations in the top 0.1% of surveys is generally very low. In Austria, the top 0.1% share goes from 4.4% to 21.1%, and in Germany from 7.7% to 14.7%.

The method has virtually no impact on estimates for France and Spain, which strongly oversample their surveys. However, it does change them for the United States, which has even stronger oversampling (but excludes the *Forbes 400*).

Austria and the United States have the highest inequalities of wealth. Their top 1% and 0.1% shares are comparable, although standard errors are quite high for Austria given its imprecise survey data. Adding the wealth ranking, however, mitigates the problem. They are followed by Germany, Portugal, then Italy. France and Spain have the lowest levels of inequality, with top 1% shares of 18% and 15% respectively.

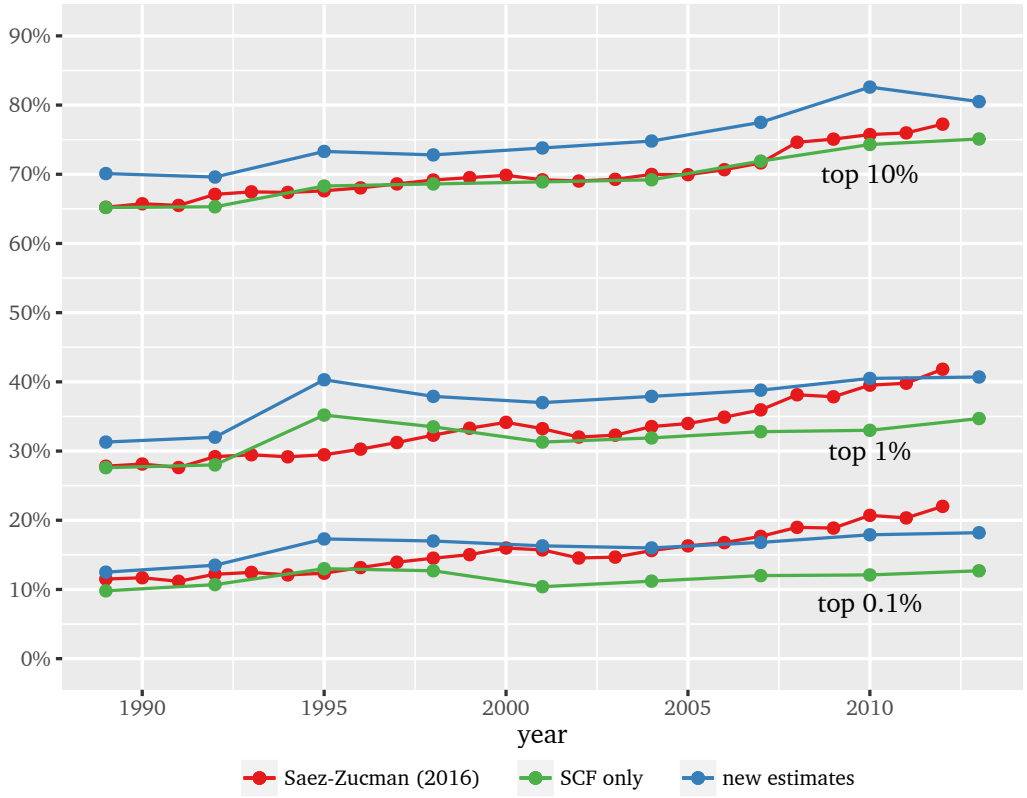
The top 1% share in Austria is within the upper range of Vermeulen (2014), and also close to the 38% estimate of Eckerstorfer et al. (2015). In Germany, I find lower inequality than Vermeulen (2014), due to the equalizing impact of the adjustment to the national accounts. Non-adjusted estimates, however, are very similar. Other estimates are also close to what we find in the literature.

In the United States, it is possible to use the previous waves of the SCF to look at the evolution of wealth inequality since 1989. Figure 3.3 compares this evolution from three sources: the new method, the direct estimates from the SCF, and the capitalization method from Saez and Zucman (2016). Like Saez and Zucman (2016), and unlike direct SCF estimates, I find that the top 1% and 0.1% shares have sharply increased since 1989. Today's top shares are similar

		survey only			survey + national accounts			survey + tail correction			final estimates		
		top 10% share	top 1% share	top 0.1% share	top 10% share	top 1% share	top 0.1% share	top 10% share	top 1% share	top 0.1% share	top 10% share	top 1% share	top 0.1% share
Austria	2010	60.6% (6.4%)	22.6% (7.1%)	4.4% (2.4%)	63.5% (7%)	22.6% (7.1%)	4% (2.4%)	68.3% (4.1%)	37.7% (3.4%)	19.5% (1.6%)	72% (5.3%)	40.7% (4.6%)	21.1% (2.3%)
France	2010	48.5% (1.3%)	17.3% (1.9%)	7.1% (1.9%)	49.7% (1.2%)	18.2% (1.9%)	7.4% (1.9%)	49% (0.9%)	18% (1.2%)	7.9% (1%)	49.9% (0.9%)	18.4% (1.2%)	7.5% (1%)
Germany	2011	58.9% (2.5%)	23.8% (3.7%)	7.7% (4.4%)	56.2% (2.3%)	20.7% (3.2%)	6.5% (3.7%)	63.1% (2%)	31% (2.9%)	17.2% (2.7%)	59.8% (2.3%)	27% (3.6%)	14.7% (3.9%)
Italy	2011	45.4% (1.1%)	14% (1%)	3.1% (1.1%)	49.5% (1.3%)	16.3% (0.9%)	3.8% (1.3%)	47.8% (1.6%)	17.7% (1.6%)	7.1% (0.9%)	52.5% (1.9%)	22% (2.1%)	8.6% (1%)
Portugal	2010	54.1% (2%)	21.8% (2.8%)	9.4% (2.4%)	60.9% (2.2%)	25.6% (3.6%)	11.8% (3.3%)	56.4% (2.6%)	26% (4.1%)	13.9% (4.5%)	61.4% (1.7%)	26.8% (2.6%)	13.1% (2.7%)
Spain	2009	44.4% (1.1%)	15% (1.4%)	5.5% (1.1%)	44.9% (1.1%)	15.1% (1.4%)	5.5% (1.1%)	44.6% (4.6%)	15.7% (4.3%)	6% (4.4%)	44.5% (1.8%)	15.1% (2%)	5.2% (1.8%)
United States	2010	74.3% (0.7%)	33% (0.9%)	12.1% (0.6%)	80.7% (0.8%)	36.2% (1%)	13.1% (0.7%)	76% (0.7%)	36.2% (1%)	15.5% (0.7%)	82.6% (0.7%)	40.5% (1.1%)	17.9% (0.9%)

Bootstrapped standard errors in parentheses. The last three columns, “final estimates”, includes both the tail correction and the adjustment to the national accounts.

Table 3.4: Estimates of wealth inequality



Note that Saez and Zucman (2016) use a different statistical unit. *Source:* author calculations from the SCF, Forbes and the Financial Accounts of the United States; Saez and Zucman (2016).

Figure 3.3: Top wealth shares in the United States, 1989–2013

to those of Saez and Zucman (2016). The increase, however, has been less steady. Moreover, top shares now seem to have stabilized, whereas they are still increasing according to Saez and Zucman (2016). Despite those differences, the results corroborate their finding that wealth inequality in the United States is higher than previously thought.

Table 3.5 provides a robustness check against the possibility that some rankings of wealth mistakenly include non-residents. It provides the same estimates as before, but with 20% of the people in the rankings removed at random. The top shares are slightly lower, but not enough to change the conclusions.

3.3.2 Europe and the United States

We can finally combine the distributions of wealth estimated for the six European countries, in order to compare the overall distribution of wealth in Europe and in the United States. Let $F_1, \ldots, F_K$ be the CDF for the distribution of wealth in K countries, and let $N_1, \ldots, N_K$ the population of each of these countries. Define the total population $N = \sum_{k=1}^{K} N_k$. The distribution

		no adjustment			with adjustment		
		top 10% share	top 1% share	top 0.1% share	top 10% share	top 1% share	top 0.1% share
Austria	2010	67.7% (4.7%)	36.6% (4.4%)	18.3% (2.2%)	71.4% (5.9%)	39.5% (5.6%)	19.8% (2.8%)
France	2010	48.8% (0.9%)	17.7% (1.2%)	7.6% (1%)	49.9% (0.9%)	18.2% (1.2%)	7.3% (1%)
Germany	2011	62.5% (2.1%)	29.8% (2.8%)	15.8% (1.8%)	59.1% (2%)	25.9% (2.6%)	13.3% (1.7%)
Italy	2011	47.7% (1.6%)	17.5% (1.6%)	6.9% (0.9%)	52.5% (1.9%)	21.9% (2.1%)	8.6% (1.1%)
Portugal	2010	55.7% (2.4%)	24.7% (3.7%)	12.4% (3.9%)	60.9% (1.8%)	25.9% (2.5%)	12% (2.3%)
Spain	2009	44.2% (3%)	14.7% (3.7%)	5.5% (4%)	44.5% (1.5%)	14.6% (1.5%)	5.1% (1.1%)
United States	2010	75.7% (0.7%)	35.5% (0.9%)	14.6% (0.6%)	82.4% (0.7%)	39.6% (1%)	16.5% (0.7%)

Bootstrapped standard errors in parentheses. “adjustment” refers to the rescaling of assets to match the national accounts. Removing 20% of people in the wealth rankings accounts for the possibility that some individuals in the list may not reside in the country they are attributed to, and serves as a robustness check.

Table 3.5: Estimates with 20% of the wealth rankings removed

of wealth over the K countries can be represented as a mixture model, with the following CDF:

$$F(x) = \sum_{k=1}^{K} \frac{N_k}{N} F_k(x)$$

We can calculate quantiles by numerical inversion of F, and expected values as weighted sums of the same expected values in each country. Hence, we can estimate global top shares.

	population	GDP	wealth	top 10% share	top 1% share	top 0.1% share
Europe (6 countries)	270 millions	\$9.8tn	\$47.6tn	53.6% (0.9%)	21.8% (1.1%)	9.8% (1%)
United States	310 millions	\$15tn	\$57.1tn	83.6% (0.7%)	40.5% (1.1%)	17.9% (0.9%)

Bootstrapped standard errors in parentheses. The six European countries are Austria, France, Germany, Italy, Portugal and Spain. Figures for the population, GDP and aggregate wealth correspond to the year 2010. GDP is calculated at purchasing power parity, and wealth at market exchange rates. *Source:* United Nations Population Division; OECD, WID and Financial Accounts of the United States.

Table 3.6: Wealth inequality in Europe and in the United States

Table 3.6 shows the result of such estimates. For context, it compares the two areas in terms of

population, GDP and wealth. The six European countries represent 87% of the population, 65% of the GDP (at PPP) and 83% of the aggregate wealth of the United States. Despite significant heterogeneity, the six European countries still have a much lower level of wealth inequality than the United States: their shares of the top 1% and 0.1% are 22% and 10%, nearly twice as low as the same figures for the United States: 40% and 18%.

Conclusion

This dissertation provides new estimates of wealth inequality that are based, mostly, on survey data. It uses the Survey of Consumer Finances (SCF) for the United States, and the Household Finance and Consumption Survey (HFCS) for six European countries: Austria, France, Germany, Italy, Portugal and Spain. Survey data, however, is known to suffer from misreporting of assets, and can be bad at capturing the top tail of the distribution given the size of samples.

I address the first problem by comparing survey estimates of aggregate household wealth with comparable concepts in the national accounts. Financial assets suffer from underreporting more than real ones, and at the same are more unequal: as a consequence, survey estimates might underestimate wealth inequality. Under the assumption that misreporting reflects systematic valuation errors, I rescale the value of each asset to match the national accounts totals. That, alone, can lead to significant upward revision of top shares estimates.

Then, I address the second problem by combining the survey data with journalist rankings of top wealth holders to estimate the top tail of the distribution. As opposed to previous work on the subject, I do not rely on the Pareto distribution, or any other parametric family of distribution. Instead, I characterize a distribution by the graph of its quantile function on a logarithmic scale. That function behaves very well for fat-tailed distribution, and can be estimated nonparametrically via a quantile regression of order statistics against a certain transform of their rank.

The new method provides better estimates of top wealth shares than raw survey data, in particular for the top 0.1%. Using this method, I estimate that the top 1% own 18% of the wealth in the United States, as opposed to 10% in the six European countries analyzed. The situation varies between European countries: in Austria, the top 0.1% own about 21% of the wealth, against just 5% in Spain.

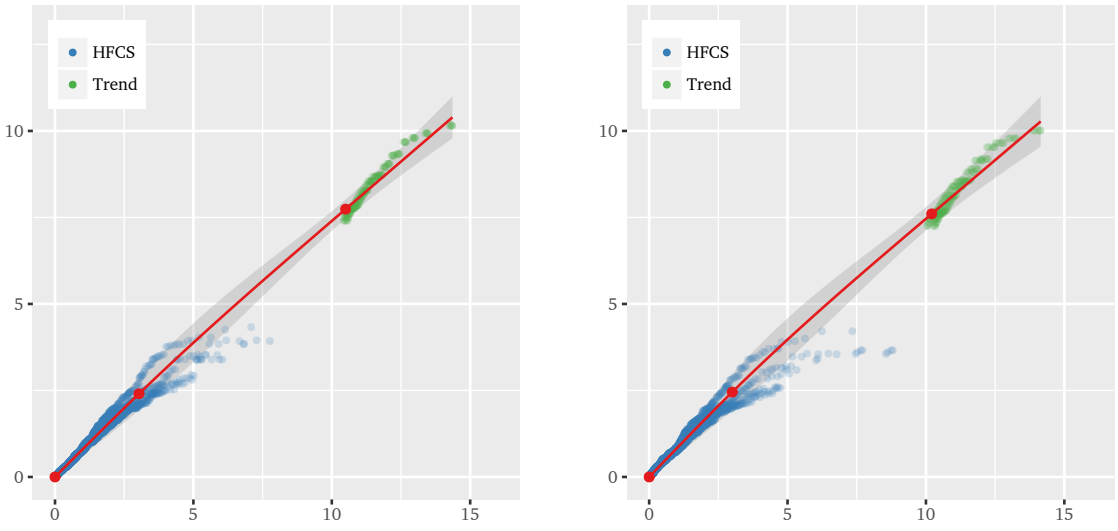
The method may still suffer from problems: in particular, it does not explicitly correct for nonresponse bias, so that actual inequality may be even higher. However, it shows that a proper estimation of the top tail of the distribution can go a long way towards a realistic assessment of wealth inequality. It opens interesting perspectives for the estimation of wealth inequality worldwide, given that many countries do not provide sufficient administrative data on wealth.

Appendix A

Detailed country results

This appendix shows the detailed results of the estimation of the tail function, country by country, for both adjusted and non-adjusted survey data. The first set of graphs shows the tail function alongside the survey and the wealth ranking to evaluate the quality of the fit. The five implicates are drawn on top of each other to better visualize uncertainty. The second set of graphs shows the full quantile function on a logarithmic scale. All amounts are in constant 2010 euros.

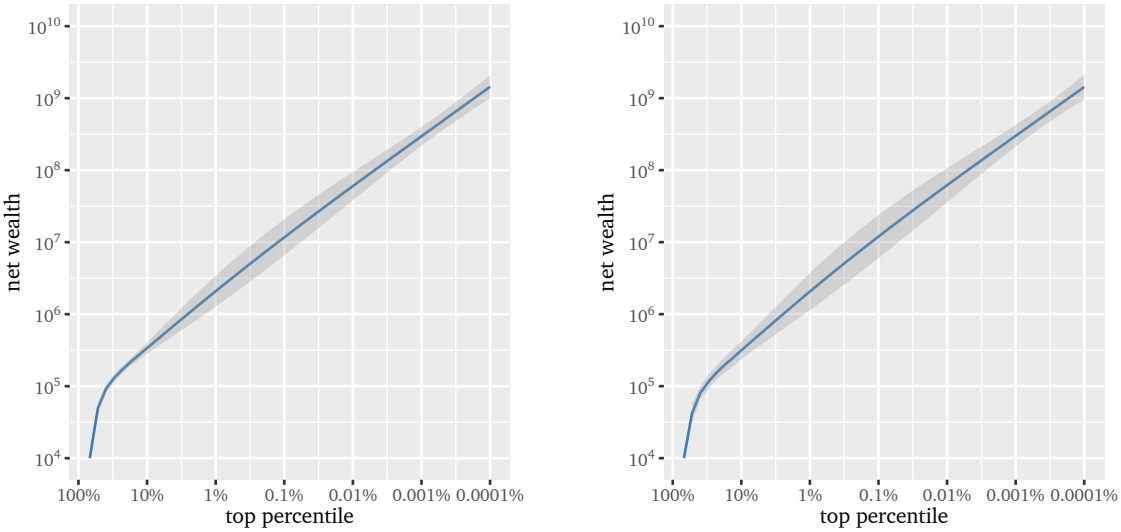
A.1 Austria



(a) non adjusted (b) adjusted

The five imputed data sets are drawn on top of each other. 95% confidence interval in grey. Bullets indicate the knots of the spline.

Figure A.1: Tail function (Austria)

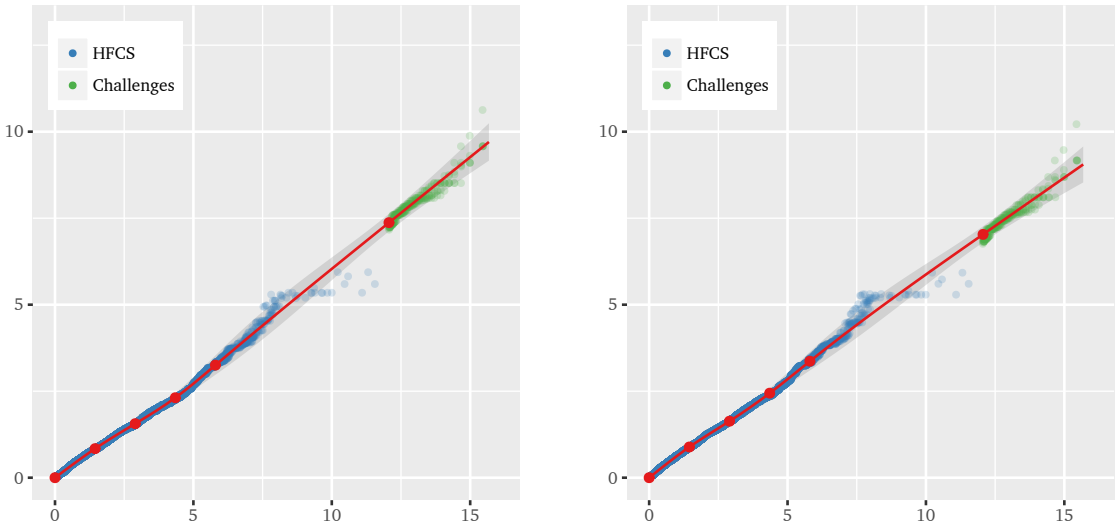


(a) non adjusted (b) adjusted

Constant 2010 euros. 95% confidence interval in grey.

Figure A.2: Quantile function (Austria)

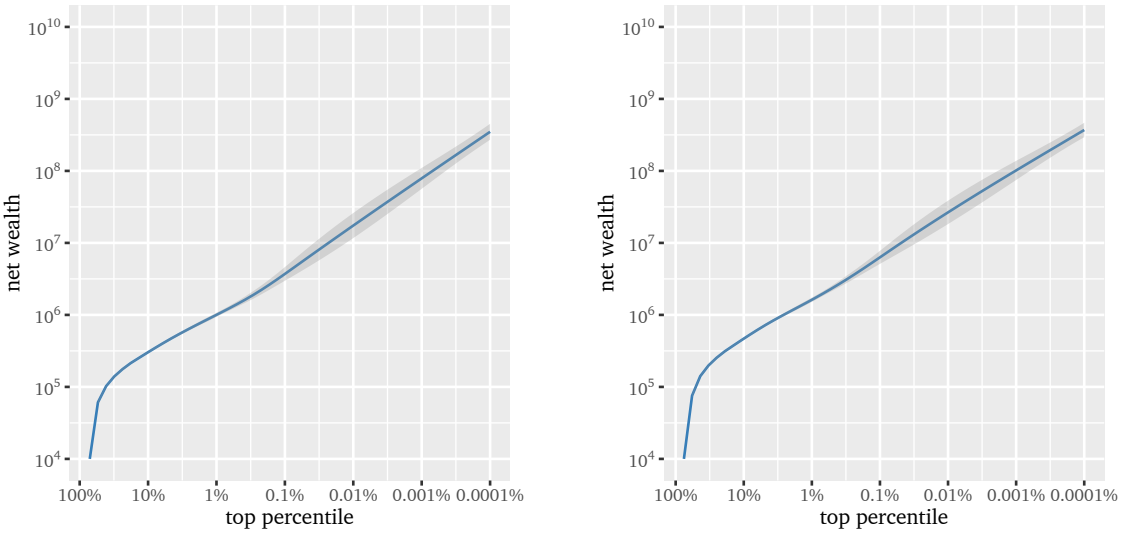
A.2 France



(a) non adjusted (b) adjusted

The five imputed data sets are drawn on top of each other. 95% confidence interval in grey. Bullets indicate the knots of the spline.

Figure A.3: Tail function (France)

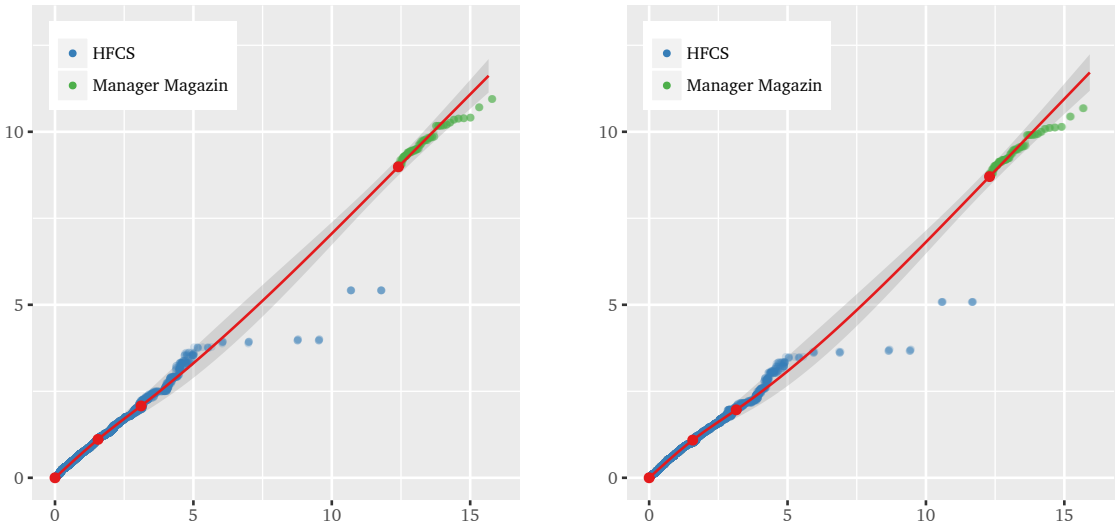


(a) non adjusted (b) adjusted

Constant 2010 euros. 95% confidence interval in grey.

Figure A.4: Quantile function (France)

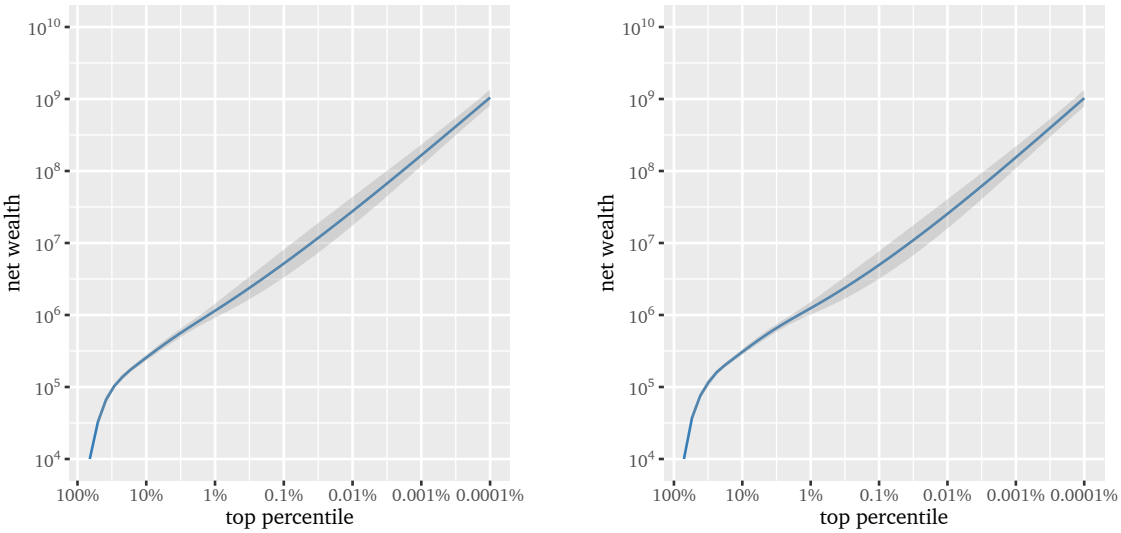
A.3 Germany



(a) non adjusted (b) adjusted

The five imputed data sets are drawn on top of each other. 95% confidence interval in grey. Bullets indicate the knots of the spline.

Figure A.5: Tail function (Germany)

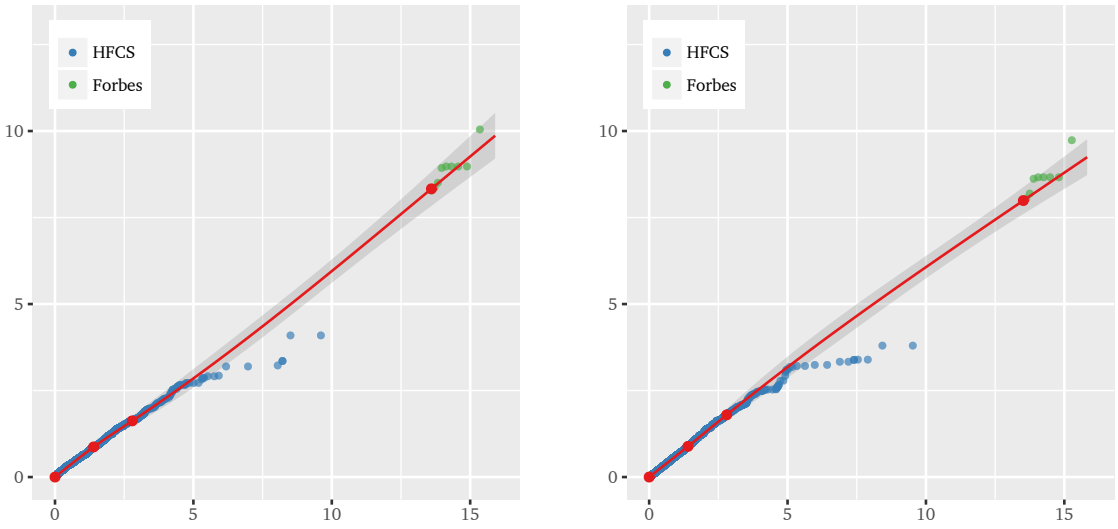


(a) non adjusted (b) adjusted

Constant 2010 euros. 95% confidence interval in grey.

Figure A.6: Quantile function (Germany)

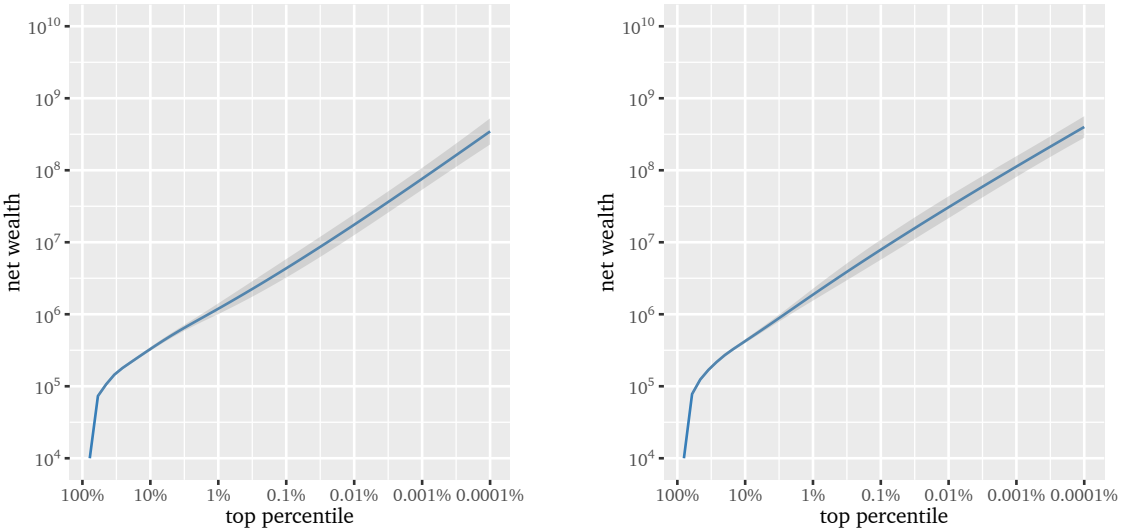
A.4 Italy



(a) non adjusted (b) adjusted

The five imputed data sets are drawn on top of each other. 95% confidence interval in grey. Bullets indicate the knots of the spline.

Figure A.7: Tail function (Italy)

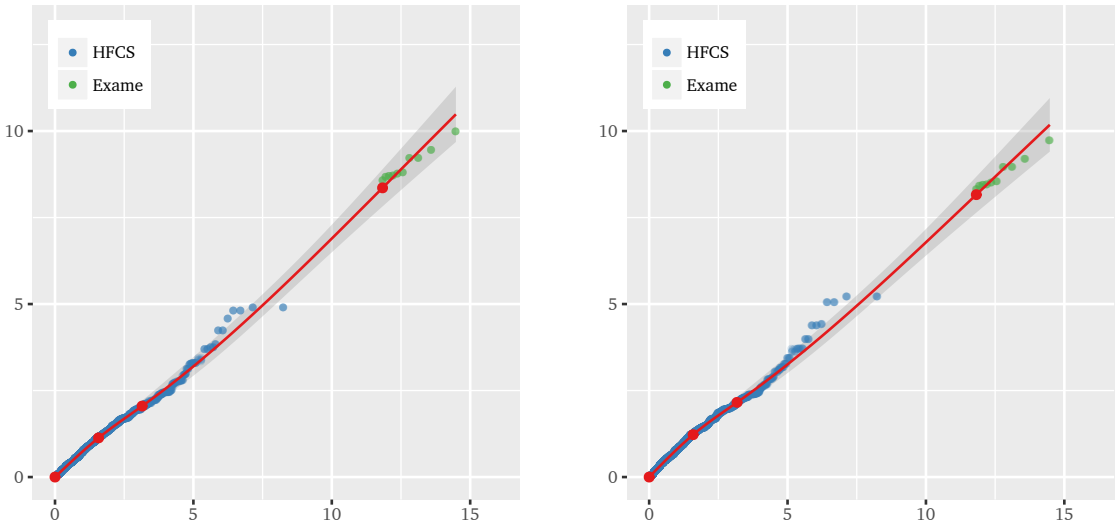


(a) non adjusted (b) adjusted

Constant 2010 euros. 95% confidence interval in grey.

Figure A.8: Quantile function (Italy)

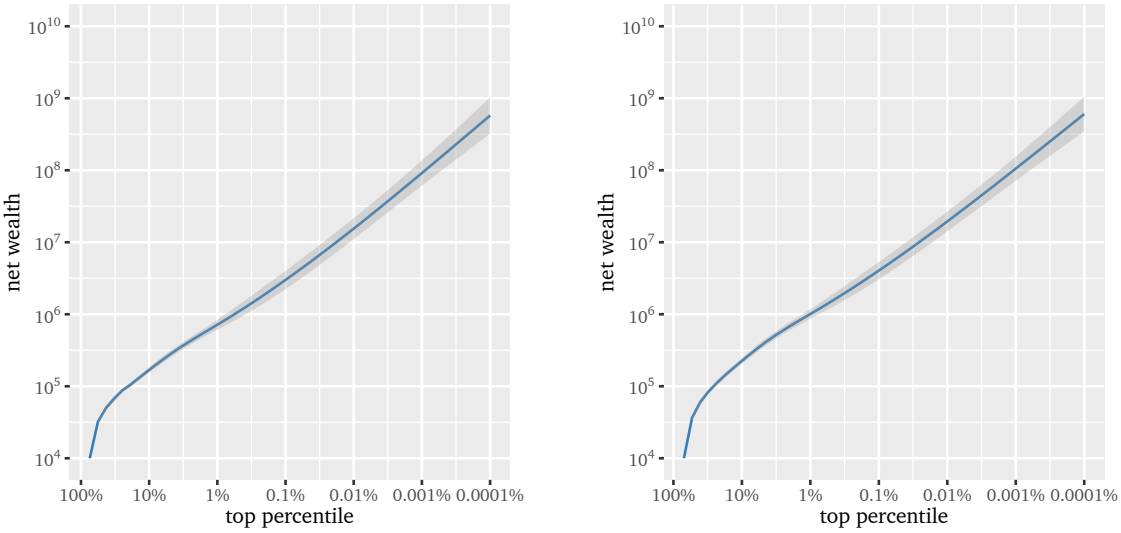
A.5 Portugal



(a) non adjusted (b) adjusted

The five imputed data sets are drawn on top of each other. 95% confidence interval in grey. Bullets indicate the knots of the spline.

Figure A.9: Tail function (Portugal)

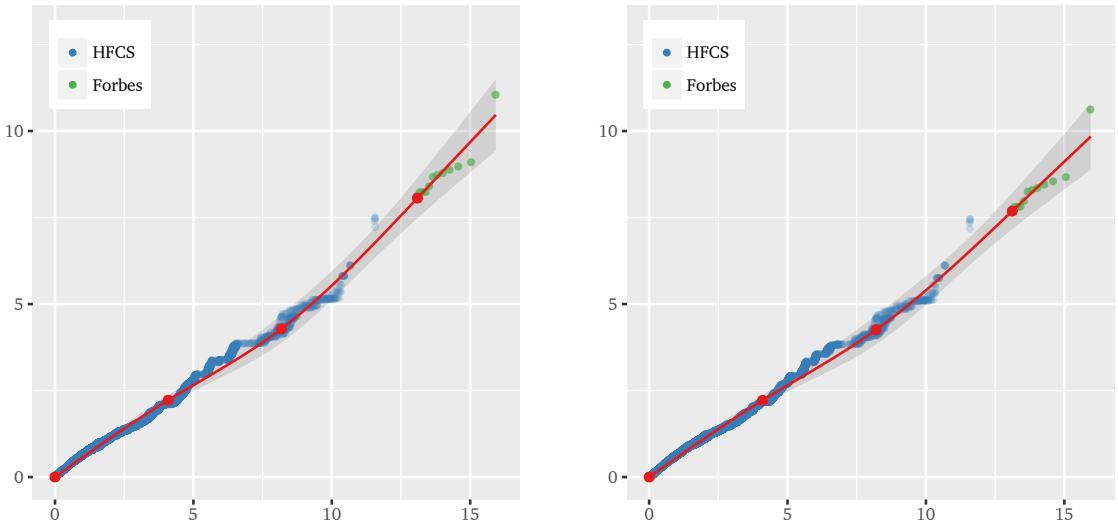


(a) non adjusted (b) adjusted

Constant 2010 euros. 95% confidence interval in grey.

Figure A.10: Quantile function (Portugal)

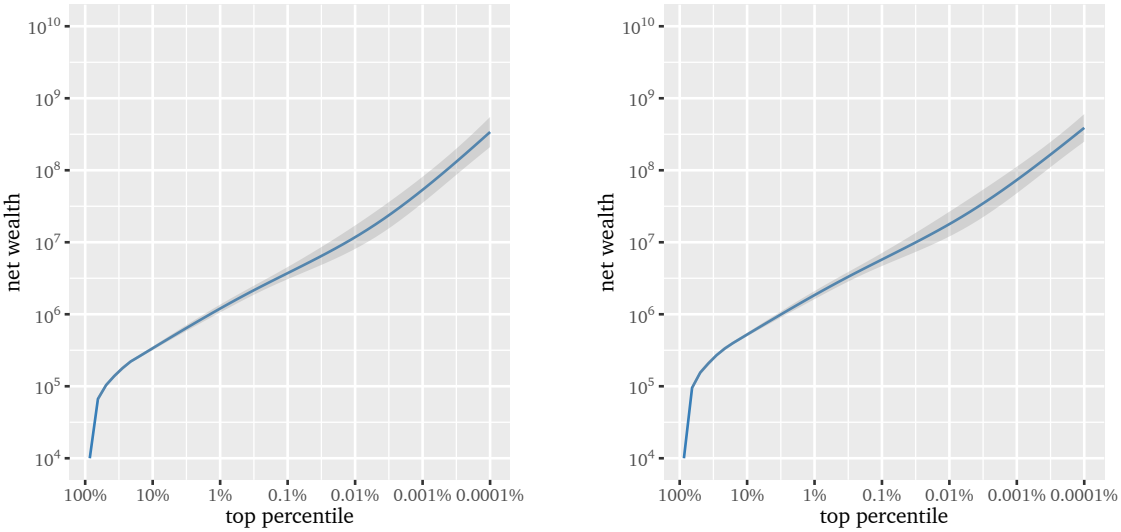
A.6 Spain



(a) non adjusted (b) adjusted

The five imputed data sets are drawn on top of each other. 95% confidence interval in grey. Bullets indicate the knots of the spline.

Figure A.11: Tail function (Spain)

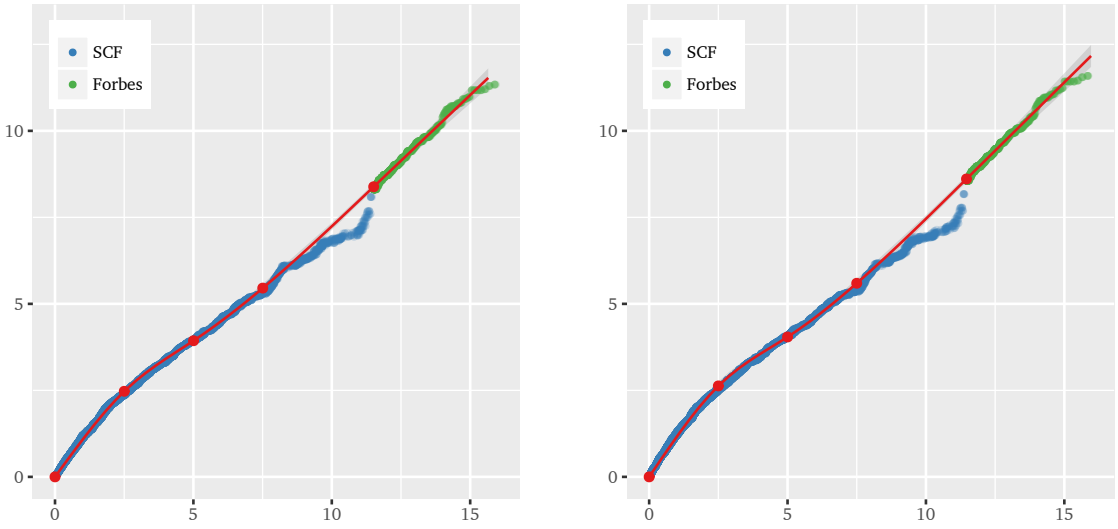


(a) non adjusted (b) adjusted

Constant 2010 euros. 95% confidence interval in grey.

Figure A.12: Quantile function (Spain)

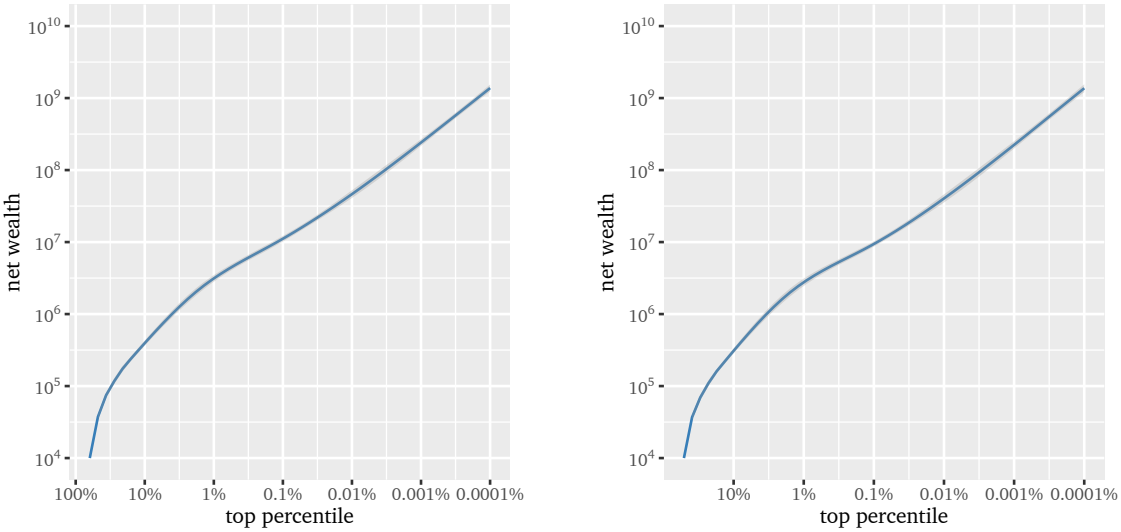
A.7 United States



(a) non adjusted (b) adjusted

The five imputed data sets are drawn on top of each other. 95% confidence interval in grey. Bullets indicate the knots of the spline.

Figure A.13: Tail function (United States)



(a) non adjusted (b) adjusted

Constant 2010 euros. 95% confidence interval in grey.

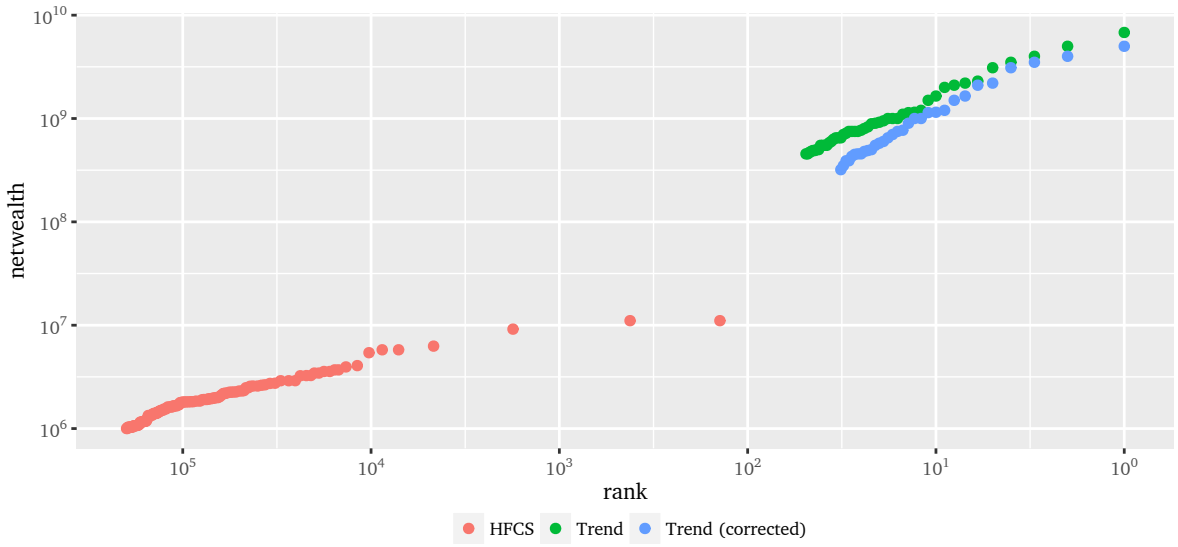
Figure A.14: Quantile function (United States)

Appendix B

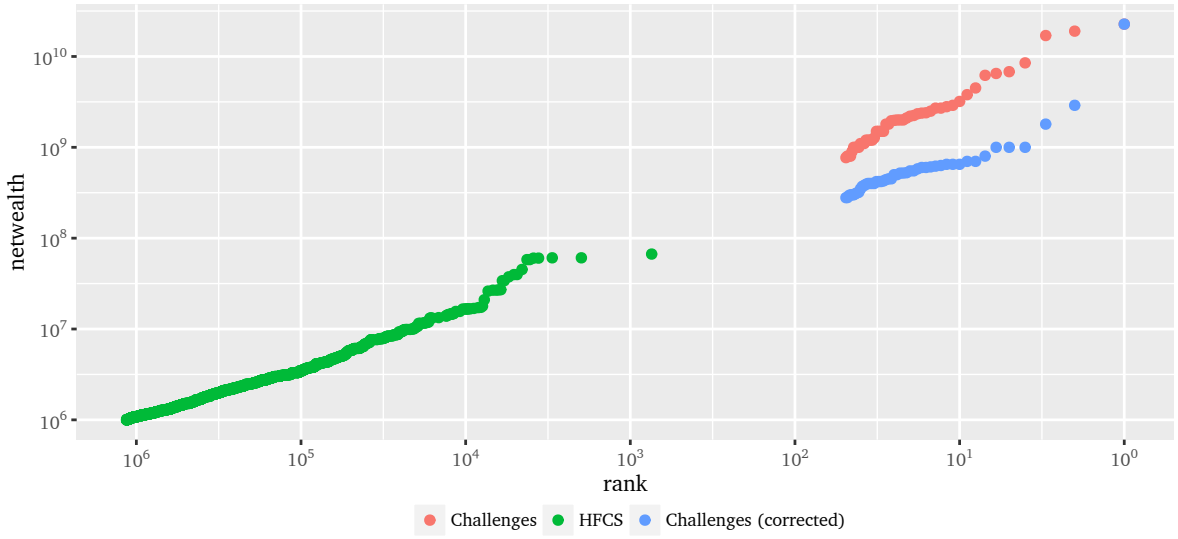
Correction of wealth rankings

This appendix shows the wealth rankings, before and after correction, on a Pareto diagram alongside the survey data adjusted to the national accounts. Only the first implicate is shown for better visibility. All amounts are in constant 2010 euros.

When no corrected ranking is shown, it means that there was not enough families in it for the correction to change anything (that concerns *Forbes* and *Exame*).

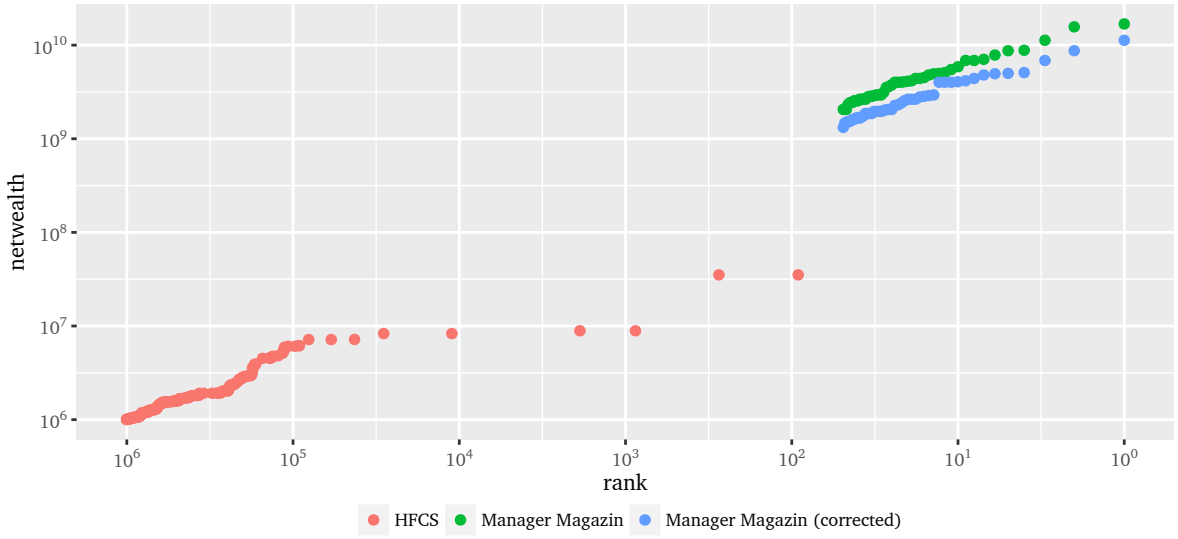


Constant 2010 euros. Adjusted survey data. Only the first implicate is shown.



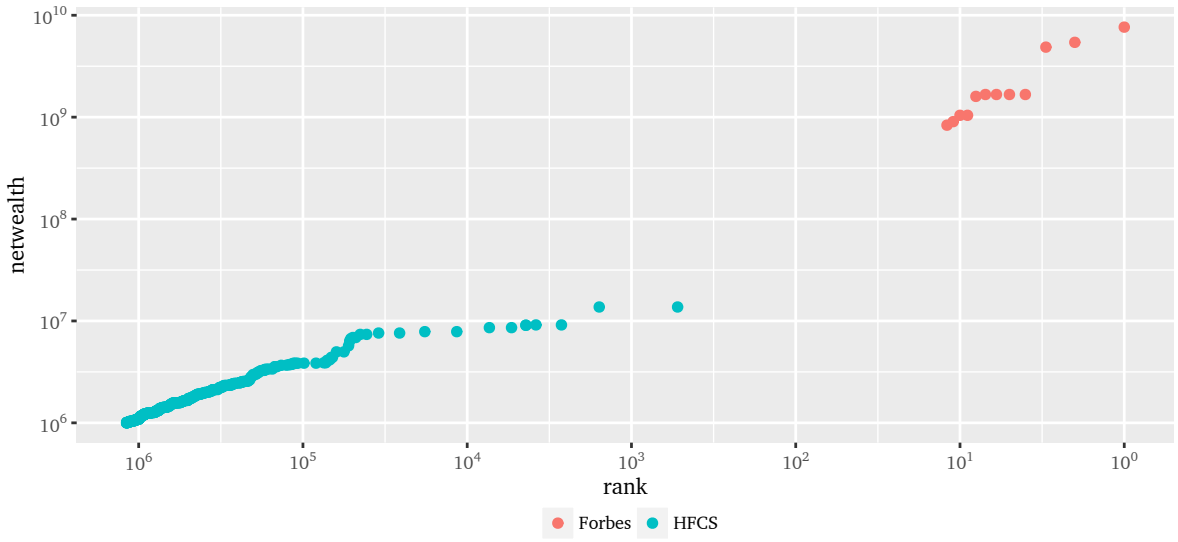
Constant 2010 euros. Adjusted survey data. Only the first implicate is shown.

Figure B.1: Wealth ranking: Challenges (France)



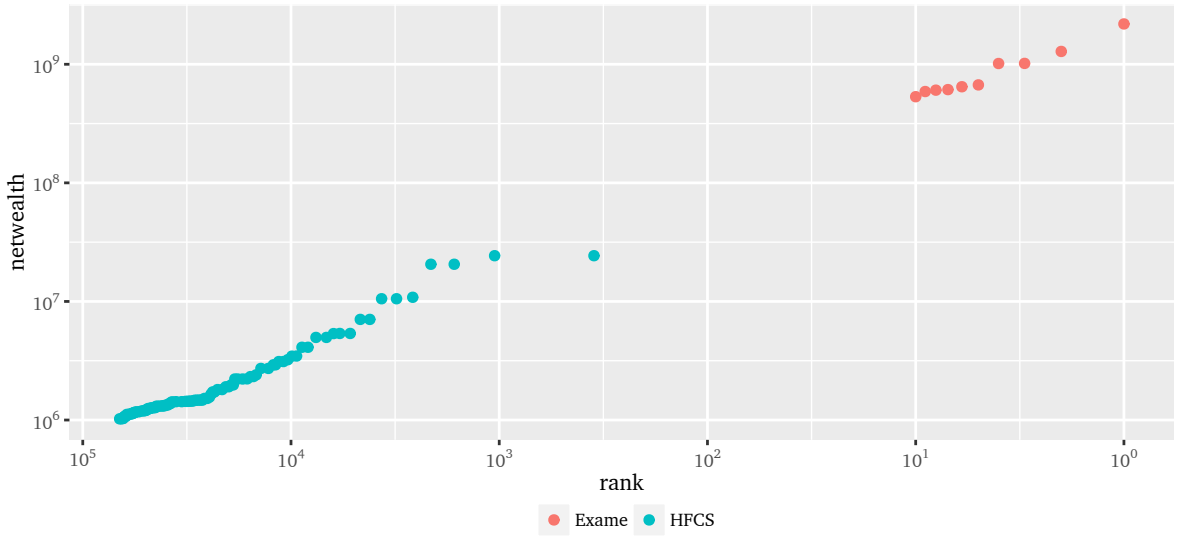
Constant 2010 euros. Adjusted survey data. Only the first implicate is shown.

Figure B.2: Wealth ranking: Manager Magazin (Germany)



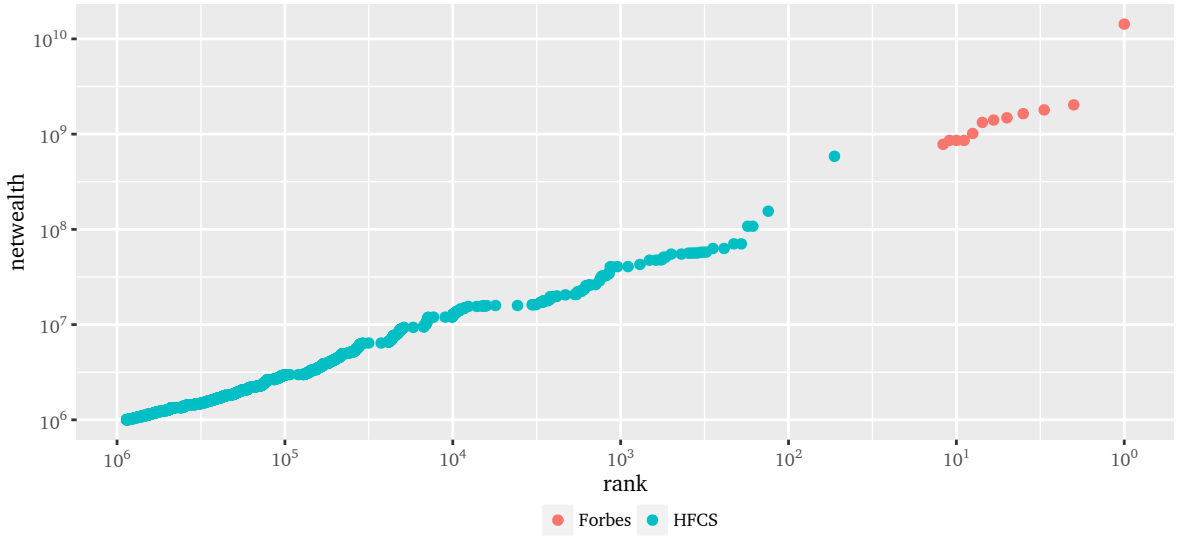
Constant 2010 euros. Adjusted survey data. Only the first implicate is shown.

Figure B.3: Wealth ranking: Forbes (Italy)



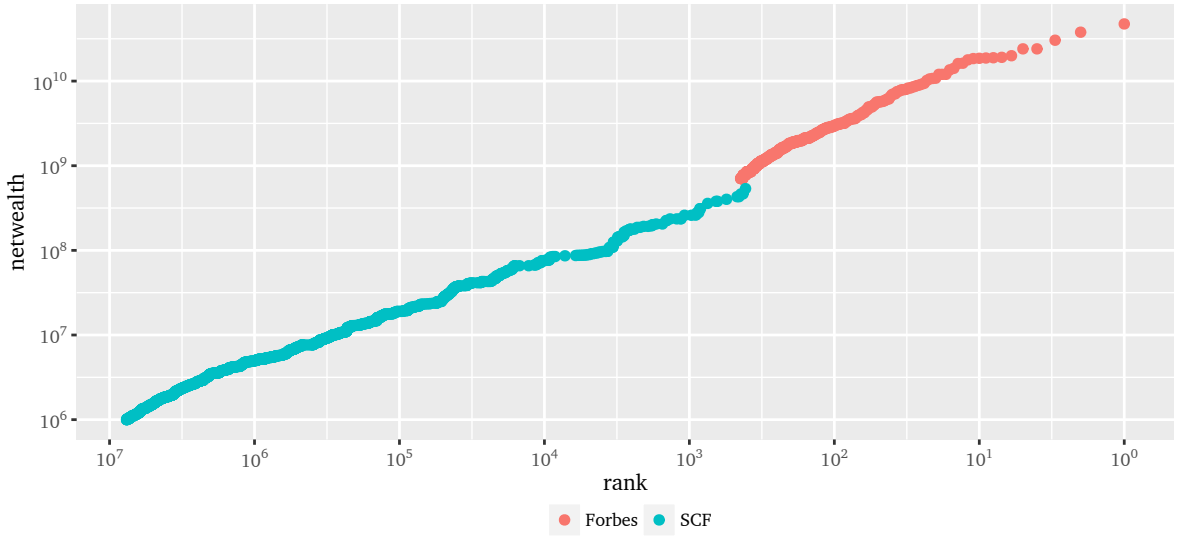
Constant 2010 euros. Adjusted survey data. Only the first implicate is shown.

Figure B.4: Wealth ranking: Exame (Portugal)



Constant 2010 euros. Adjusted survey data. Only the first implicate is shown.

Figure B.5: Wealth ranking: Forbes (Spain)



Constant 2010 euros. Adjusted survey data. Only the first implicate is shown.

Figure B.6: Wealth ranking: Forbes (United States)

Appendix C

Generalized Least Squares estimator: proofs

Explicit formula for the estimator

Recall the following notations:

$$\boldsymbol{y} = \begin{bmatrix} \log(X_{(1)}/\omega) \\ \log(X_{(2)}/\omega) \\ \vdots \\ \log(X_{(n)}/\omega) \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} H_n - H_{n-1} \\ H_n - H_{n-2} \\ \vdots \\ H_n - H_0 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} H_n^{(2)} - H_{n-1}^{(2)} & H_n^{(2)} - H_{n-1}^{(2)} & \cdots & H_n^{(2)} - H_{n-1}^{(2)} \\ H_n^{(2)} - H_{n-1}^{(2)} & H_n^{(2)} - H_{n-2}^{(2)} & \cdots & H_n^{(2)} - H_{n-2}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ H_n^{(2)} - H_{n-1}^{(2)} & H_n^{(2)} - H_{n-2}^{(2)} & \cdots & H_n^{(2)} - H_0^{(2)} \end{bmatrix}$$

The GLS estimator is defined as:

$$\hat{\beta}^{\text{GLS}} = \underset{\beta}{\operatorname{argmin}} (\boldsymbol{y} - \beta \boldsymbol{x})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \beta \boldsymbol{x})$$

That problem has an explicit solution:

$$\hat{\beta}^{\text{GLS}} = (\boldsymbol{x}' \boldsymbol{\Sigma}^{-1} \boldsymbol{x})^{-1} \boldsymbol{x}' \boldsymbol{\Sigma}^{-1} \boldsymbol{y} \tag{C.1}$$

Because $\boldsymbol{\Sigma}$ is symmetric positive definite, we can apply Cholesky decomposition and write it as:

$$\boldsymbol{\Sigma} = \boldsymbol{P}' \boldsymbol{P}$$

where $\boldsymbol{P}$ is a lower triangular matrix with positive diagonal entries. Therefore, we can rewrite (C.1) as:

$$\hat{\beta}^{\text{GLS}} = [(\boldsymbol{P}^{-1} \boldsymbol{x})' (\boldsymbol{P}^{-1} \boldsymbol{x})]^{-1} (\boldsymbol{P}^{-1} \boldsymbol{x})' (\boldsymbol{P}^{-1} \boldsymbol{y})$$

We can check with some algebra:

$$\boldsymbol{P} = \begin{bmatrix} 1/n & 0 & 0 & \cdots & 0 & 0 \\ 1/n & 1/(n-1) & 0 & \cdots & 0 & 0 \\ 1/n & 1/(n-1) & 1/(n-2) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1/n & 1/(n-1) & 1/(n-2) & \cdots & 1/2 & 0 \\ 1/n & 1/(n-1) & 1/(n-2) & \cdots & 1/2 & 1 \end{bmatrix}$$

Furthermore:

$$\boldsymbol{P}^{-1} = \begin{bmatrix} n & 0 & 0 & \cdots & 0 & 0 \\ -(n-1) & n-1 & 0 & \cdots & 0 & 0 \\ 0 & -(n-2) & n-2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

Hence:

$$\boldsymbol{P}^{-1}\boldsymbol{x} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}'$$

Thus, if we write:

$$\boldsymbol{P}^{-1}\boldsymbol{y} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}'$$

the expression of the estimator simplifies to:

$$\hat{\beta}^{\text{GLS}} = \frac{1}{n}\sum_{k=1}^{n} a_k$$

We have $a_1 = n\log(X_{(1)}/\omega)$ and for all $k \in \{2, \ldots, n\}$:

$$a_k = (n-k+1)[\log(X_{(k)}/\omega) - \log(X_{(k-1)}/\omega)]$$

Finally:

$$\begin{aligned} \hat{\beta}^{\text{GLS}} &= \log(X_{(1)}/\omega) + \frac{1}{n}\sum_{k=2}^{n}(n-k+1)[\log(X_{(k)}/\omega) - \log(X_{(k-1)}/\omega)] \\ &= \log(X_{(1)}/\omega) + \frac{1}{n}\sum_{k=2}^{n}(n-k+1)\log(X_{(k)}/\omega) - \frac{1}{n}\sum_{k=2}^{n}(n-k+1)\log(X_{(k-1)}/\omega) \\ &= \log(X_{(1)}/\omega) + \frac{1}{n}\sum_{k=2}^{n}(n-k+1)\log(X_{(k)}/\omega) - \frac{1}{n}\sum_{k=1}^{n-1}(n-k)\log(X_{(k)}/\omega) \\ &= \frac{1}{n}\sum_{k=1}^{n}\log(X_{(k)}/\omega) \end{aligned}$$

In each term of the sum, there is no dependence in k other than in the index of order statistics. Since all terms of the sum are interchangeable, we can drop the order statistics notation and we get:

$$\hat{\beta}^{\mathrm{GLS}} = \frac{1}{n}\sum_{k=1}^{n}\log(X_k/\omega)$$

Relation to the maximum likelihood estimator

The likelihood function can be written:

$$\mathscr{L}(\beta; X_1, \ldots, X_n) = \prod_{k=1}^{n} \frac{\omega^{1/\beta}}{\beta X_k^{1/\beta+1}}$$

Hence the log-likelihood function:

$$\ell(\beta; X_1, \ldots, X_n) = -n\log(\beta) + n(\log\omega)/\beta - (1/\beta + 1)\sum_{k=1}^{n}\log X_k$$

Maximizing with respect to β yields the first order condition:

$$-\frac{n}{\beta} - \frac{n(\log\omega)}{\beta^2} + \frac{1}{\beta^2}\sum_{k=1}^{n}\log X_k = 0 \Leftrightarrow n\beta - \sum_{k=1}^{n}\log(X_k/\omega) = 0$$

Therefore, we get an expression for β^{ML} which is the same as for β^{GLS}:

$$\beta^{\mathrm{ML}} = \beta^{\mathrm{GLS}} = \frac{1}{n}\sum_{k=1}^{n}\log(X_k/\omega)$$

Bibliography

Abberger, K (1998). “Cross-validation in nonparametric quantile regression”. In: *Allgemeines Statistisches Archiv* 82, pp. 149–161.

— (2001). “Penalizing function based bandwidth choice in nonparametric quantile regression”. In: *CoFE Discussion Paper* 01-10. URL: http://EconPapers.repec.org/RePEc:knz:cofedp:0110.

Aitken, A. C. (1936). “On Least Squares and Linear Combination of Observations”. In: *Proceedings of the Royal Society of Edinburgh* 55, pp. 42–48. URL: http://journals.cambridge.org/article_S0370164600014346.

Andreasch, Michael and Peter Lindner (2014). “Micro and Macro Data: A Comparison of the Household Finance and Consumption Survey with Financial Accounts in Austria”. In: *ECB Working Paper Series* 1673.

Atkinson, A. B. and Thomas Piketty (2010). *Top incomes: a global perspective*. Oxford: Oxford University Press. ISBN: 0198727747.

Avery, Robert B., Gregory E. Elliehausen, and Arthur B Kennickell (1988). “Measuring wealth with survey data: an evaluation of the 1983 Survey of Consumer Finances”. In: *Review of Income and Wealth* 34.4, pp. 339–369. ISSN: 1475-4991. DOI: 10.1111/j.1475-4991.1988.tb00575.x. URL: http://dx.doi.org/10.1111/j.1475-4991.1988.tb00575.x.

Bach, S, A Thiemann, and A Zucco (2015). “The Top Tail of the Wealth Distribution in Germany, France, Spain, and Greece”. In: *DIW Discussion Papers* 112953. URL: https://ideas.repec.org/p/zbw/vfsc15/112953.html.

Bahadur, R. R. (1966). “A Note on Quantiles in Large Samples”. In: *Ann. Math. Statist.* 37.3, pp. 577–580. URL: http://dx.doi.org/10.1214/aoms/1177699450.

Balakrishnan, N and Asit P Basu (1995). *The exponential distribution : theory, methods, and applications*. Amsterdam, United States: Gordon and Breach. ISBN: 978-2884491921.

Bover, Olympia (2010). “Wealth Inequality and Household Structure: U.S. vs. Spain”. In: *Review of Income and Wealth* 56.2, pp. 259–290. ISSN: 1475-4991. DOI: 10.1111/j.1475-4991.2010.00376.x. URL: http://dx.doi.org/10.1111/j.1475-4991.2010.00376.x.

Bricker, Jesse et al. (2012). “Changes in U.S. Family Finances from 2007 to 2010: Evidence from the Survey of Consumer Finances”. In: *Federal Reserve Bulletin* 98.2, pp. 1–80. URL: http://www.federalreserve.gov/pubs/bulletin/2012/pdf/scf12.pdf.

Clauset, A, Shalizi C Rohilla, and M E J Newman (2007). “Power-law distributions in empirical data”. In: *ArXiv e-prints*. arXiv: 0706.1062.

Csorgo, S, P Deheuvels, and D Mason (1985). “Kernel estimates of the tail index of a distribution”. In: *The Annals of Statistics*.

David, H. A. and H. N. Nagaraja (2005). *Order Statistics*. John Wiley & Sons, Inc. URL: http://onlinelibrary.wiley.com/book/10.1002/0471722162.

Davies, James B., Susanna Sandström, et al. (2009). “The Level and Distribution of Global Household Wealth”. In: Working Paper Series 15508. DOI: 10.3386/w15508. URL: http://www.nber.org/papers/w15508.

Davies, James B. and Anthony B. Shorrocks (2000). “The distribution of wealth”. In: *Handbook of Income Distribution*. Ed. by A.B. Atkinson and F. Bourguignon. 1st ed. Vol. 1. Elsevier. Chap. 11, pp. 605–675. URL: http://EconPapers.repec.org/RePEc:eee:incchp:1-11.

Dettling, L J et al. (2015). “Comparing Micro and Macro Sources for Household Accounts in the United States: Evidence from the Survey of Consumer Finances”. In: *Finance and Economics Discussion Series* 2015.86, pp. 1–69.

Dolan, Kerry A. (2012). *Methodology: How We Crunch The Numbers*. URL: http://www.forbes.com/sites/kerryadolan/2012/03/07/methodology-how-we-crunch-the-numbers/.

Eckerstorfer, P et al. (2015). “Correcting for the Missing Rich: An Application to Wealth Survey Data”. In: *Review of Income and Wealth*. ISSN: 1475-4991. URL: http://dx.doi.org/10.1111/roiw.12188.

Exame (2009). “Para mais informação, nomeadamente informação metodológica”. In: *Exame* 304.

Fack, Gabrielle and Camille Landais (2016). “The effect of tax enforcement on tax elasticities: Evidence from charitable contributions in France”. In: *Journal of Public Economics* 133, pp. 23–40. ISSN: 0047-2727. URL: http://www.sciencedirect.com/science/article/pii/S0047272715001826.

Fessler, Pirmin and Martin Schürz (2013). “Cross-Country Comparability of the Eurosystem Household Finance and Consumption Survey”. In: *Monetary Policy & the Economy* 2, pp. 29–50. URL: https://ideas.repec.org/a/onb/oenbmp/y2013i2b2.html.

Francisco, Carol A. and Wayne A. Fuller (1991). “Quantile Estimation with a Complex Survey Design”. In: *Ann. Statist.* 19.1, pp. 454–469. DOI: 10.1214/aos/1176347993. URL: http://dx.doi.org/10.1214/aos/1176347993.

Henriques, Alice M. (2013). “Are homeowners in denial about their house values? Comparing owner perceptions with transaction-based indexes”. In: 2013-79. URL: https://ideas.repec.org/p/fip/fedgfe/2013-79.html.

Henriques, Alice M. and Joanne W. Hsu (2014). “Analysis of Wealth Using Micro- and Macrodata: A Comparison of the Survey of Consumer Finances and Flow of Funds Accounts”. In: pp. 245–274. URL: http://econpapers.repec.org/RePEc:nbr:nberch:12829.

HFCN (2013a). “The Eurosystem Household Finance and Consumption Survey: methodological report for the first wave”. In: *ECB Statistics Paper Series* 1. URL: https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp1en.pdf.

— (2013b). “The Eurosystem Household Finance and Consumption Survey: results from the first wave”. In: *ECB Statistics Paper Series* 2. URL: https://www.ecb.europa.eu/pub/pdf/other/ecbsp2en.pdf.

Honkkila, J and I K Kavonius (2013). “Micro and macro analysis on household income, wealth and saving in the euro area”. In: *ECB Working Paper Series* 1619.

Kennickell, Arthur B (2009). “Getting to the Top: Reaching Wealthy Respondents in the SCF”. In: *Paper prepared for the 2009 Joint Statistical Meetings, Washington, DC*. URL: https://www.federalreserve.gov/econresdata/scf/files/ASA200911.pdf.

Kolenikov, Stanislav (2010). “Resampling variance estimation for complex survey data”. In: *The Stata Journal* 10.2.

Kopczuk, Wojciech and Emmanuel Saez (2004). “Top Wealth Shares in the United States, 1916-2000: Evidence from Estate Tax Returns”. In: *National Tax Journal* 57.2, Part 2, pp. 445–487.

Milanovic, Branko (2002). “True World Income Distribution, 1988 and 1993: First Calculation Based on Household Surveys Alone”. In: *The Economic Journal* 112.476, pp. 51–92. ISSN: 1468-0297. DOI: 10.1111/1468-0297.0j673. URL: http://dx.doi.org/10.1111/1468-0297.0j673.

Opsomer, Jean, Yuedong Wang, and Yuhong Yang (2001). “Nonparametric Regressin with Correlated Errors”. In: *Statist. Sci.* 16.2, pp. 134–153. DOI: 10.1214/ss/1009213287. URL: http://dx.doi.org/10.1214/ss/1009213287.

Piketty, Thomas and Gabriel Zucman (2014). “Capital is Back: Wealth-Income Ratios in Rich Countries 1700–2010”. In: *The Quarterly Journal of Economics* 129.3, pp. 1255–1310. DOI: 10.1093/qje/qju018. URL: http://qje.oxfordjournals.org/content/129/3/1255.abstract.

Rao, J. N. K. and C. F. J. Wu (1988). “Resampling Inference With Complex Survey Data”. In: *Journal of the American Statistical Association* 83.401, pp. 231–241. URL: http://www.jstor.org/stable/2288945.

Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc. ISBN: 9780470316696. DOI: 10.1002/9780470316696. URL: http://dx.doi.org/10.1002/9780470316696.

Saez, Emmanuel and Gabriel Zucman (2016). “Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data”. In: *The Quarterly Journal of Economics* 131.2, pp. 519–578. DOI: 10.1093/qje/qjw004. URL: http://qje.oxfordjournals.org/content/131/2/519.abstract.

Schucany, W. R. (1972). “Order statistics in simulation”. In: *Journal of Statistical Computation and Simulation* 1.3, pp. 281–286. URL: http://dx.doi.org/10.1080/00949657208810022.

Taleb, Nassim Nicholas and Raphael Douady (2015). “On the super-additivity and estimation biases of quantile contributions”. In: *Physica A: Statistical Mechanics and its Applications* 429, pp. 252–260.

Treguier, Eric (2012). *Comment évalue-t-on leur patrimoine ?* URL: http://www.challenges.fr/entreprise/20120711.CHA8798/comment-evalue-t-on-leur-patrimoine.html.

Tsybakov, Alexandre B. (2009). “Introduction to Nonparametric Estimation”. In: New York, NY: Springer New York. Chap. Nonparametric estimators, pp. 1–76. ISBN: 978-0-387-79052-7. DOI: 10.1007/978-0-387-79052-7_1. URL: http://dx.doi.org/10.1007/978-0-387-79052-7_1.

United Nations (1993). *System of National Accounts, 1993*. URL: http://unstats.un.org/unsd/nationalaccount/docs/1993sna.pdf.

Vermeulen, Philip (2014). “How fat is the top tail of the wealth distribution?” In: *ECB Working Paper Series* 1692. URL: https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1692.pdf.

— (2016). “Estimating the top tail of the wealth distribution”. In: 1907.

Walck, Christian (1996). *Handbook on statistical distributions for experimentalists*. URL: http://www.fysik.su.se/~walck/suf9601.pdf.

Westermeier, C and M M Grabka (2015). “Significant statistical uncertainty over share of high net worth households”. In: *DIW Economic Bulletin*. URL: http://www.diw.de/sixcms/detail.php?id=diw_01.c.500058.de.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press Books. The MIT Press. URL: https://mitpress.mit.edu/books/econometric-analysis-cross-secion-and-panel-data.

Yu, K and M C Jones (1998). “Local linear quantile regression”. In: *Journal of the American Statistical Association* 93.441, pp. 228–237. URL: http://oro.open.ac.uk/24037/.