Freie Universität Berlin

Estimating Top Wealth Shares Using Survey Data-An Empiricist's Guide

Christian Westermeier

School of Business & Economics
Discussion Paper

Economics

2016/21

Estimating Top Wealth Shares Using Survey Data–An Empiricist's Guide

Christian Westermeier

Contact at DIW Berlin – German Institute for Economic Research
Mohrenstraße 58, 10117 Berlin
Phone +49 30 89789-223
Fax +49 30 89789-115
Mail cwestermeier@diw.de

Abstract

Survey data tends to be biased toward the middle class. Often it fails to adequately cover the highly relevant group of multi-millionaires and billionaires, which in turn results in biased estimates for aggregate wealth and top wealth shares. In order to overcome the under coverage and obtain more reliable measurements of wealth inequality, researchers are simulating the top tail of wealth distributions using Pareto distributions both with and without information on high-net-worth-individuals from rich lists. In a series of Monte Carlo experiments, this study analyzes what assumptions need to be fulfilled in order for such an exercise to yield reliable results. If survey weights are uninformed about the relationship between non-response and wealth, as is to be expected empirically, the former case will underestimate top wealth shares and the latter may overestimate it, while both methods yield estimates of aggregate wealth that are still inherently biased downwards. In an application using German survey wealth data, it is shown that re-weighting the provided frequency weights based on exogenous information possibly affects the estimates more severely than choosing the right parameters of the Pareto distribution. However, empirically the three separate assumptions on the non-response yield wildly different estimates.The validity of exogenous dataâand the rich list dataâremains a matter of trust on the part of the empiricist.

Keywords

Differential non-response, non-observation bias, Pareto distribution, survey data, top wealth shares

Note

The author gratefully acknowledges funding from the Hans Böckler Foundation (project number: S-2012-610-4).
In the appendix, data from the Eurosystem Household Finance and Consumption Survey is used. The results published and the related observations and analysis may not correspond to results or analysis of the data producers.

1. Introduction

In the wake of a renewed discussion on inequality, distributive justice, and social cohesion, the distributions of income and wealth are, again, in the focus of science, media, and policy makers (Piketty 2014). While research on income and its distribution in Europe and across the world is widely available, research on wealth is comparatively scarce. One reason might be that until recently the availability of comparable data for Europe was severely limited, which changed after the Eurosystem's Household Finance and Consumption Survey (HFCS) became accessible to researchers (European Central Bank 2013a, 2013b). Although the latest report by the OECD (2015) thoroughly analyzes income inequality, there is only one chapter dedicated to research on wealth inequality. It addresses two major problems: finding comparable data sources, and the fact that information on the long-term developments is even harder to come by.

One issue is that researchers need to rely on survey data, if tax return data is not available.[1] However, survey data typically has the problem of a middle-class bias, it lacks a sufficient number of observations for the margins of the distribution. Due to the pronouncedly skewed distribution of net worth, the upper tail of the wealth distribution is of utmost relevance when analyzing wealth inequality. Some wealth surveys try to overcome this problem by oversampling rich households.[2] However, even with an oversampling of affluent households, there is the tendency that the truly rich households–in particular multi-millionaires and billionaires–are still not adequately represented in such surveys (Westermeier and Grabka 2015). In order to overcome the under coverage of high-net-worth-individuals and -households in wealth survey data, researchers started simulating the top tail of wealth distributions using Pareto distributions both with and without information on high-net-worth-individuals from rich lists.

The aim of this study is to shed light on some aspects of enhancing lacking survey data using Pareto simulated top tails that are previously neglected. Using Monte Carlo experiments, we show that wealth data, which is plagued by differential non-response,

[1] Even if wealth tax data is available, the information does not typically cover the whole population, as only taxable wealth components are recorded.

[2] "Relatively wealthy households account for a disproportionate share of the total wealth, and existing evidence suggests that the likelihood that they will not complete interviews when included in a sample is disproportionately high. Thus, there are potentially both bias and variance implications stemming from the treatment of wealthy households. Standard designs used when measuring income or expenditure might not be adequate for measuring wealth." (OECD 2015, page 147).

as opposed to a non-observation bias, might not be treated with a simple maximum likelihood estimation of the top tail, as estimates are still inherently biased downwards. Including rich list data and switching to regression estimation impacts top wealth shares, but the total net worth is still biased downwards. In the last step of the simulation, I show what potential effects are to be expected, if publishers of rich lists data systematically overestimate the top fortunes. Overall, all empirically encountered estimations of the aggregate wealth and top wealth shares using corrected data yield inherently biased results, once survey weights are uninformed and no additional data is available for calibration. As shown in an application using German survey data, if survey weights are re-calibrated to carry information on the distribution of households from exogenous sources, the estimates change tremendously. U.S. and Spanish survey data, which include sampling via wealth strata, are the best guesses as to how response behavior and wealth may be related.

In one stream of the existing literature, the bulk of studies explore the consistency of 'rich list' data from magazines such as Forbes with the power law distribution (Klass et al. 2007; Brzezinski 2014). Generally, these studies fall under the label econophysics (see Chatterjee et al. 2005) and are concerned with questions of exact statistics and alternative models describing the distribution of wealth (Clauset et al. 2009). Some of the most recent works concentrate on the question of whether rich-list data, such as the yearly-published list of billionaires by the American Forbes magazine, can be better described using other distributional assumptions (Brzezinski 2014, Capehart 2014). While the questions studied among the researchers in the econophysics camp are valid questions to study, the more relevant questions for public finance and policy makers are, (1) do power law distributions approximate the reality well enough; and (2) can we draw conclusions for the estimates of wealth distribution and top wealth shares? The statistical properties of rich-list data alone are of limited use, once a researcher needs to impute for missing observations at the upper tail of the distribution between rich-list data and survey data.

The second stream of literature is decidedly more empirically oriented and studies whether Pareto distributions are a useful complement to survey data. For instance, Vermeulen (2014) shows in a Monte Carlo experiment that the inclusion of rich lists' entries (such as the Forbes magazine) increases the precision of estimators for both the Pareto index and, as a result, the key figures of the entire wealth distribution, if compared to survey estimates without top-net-worth-holders from rich lists. Using data from the HFCS, he presents results for adjusted wealth distributions based on arbitrarily chosen minimum values for the Pareto distribution. Bach et al. (2014) carry out a similar exercise using survey wealth data and rich list data from Germany. Eckerstorfer et al. (2015) rely only on survey data and present a method for the identification of a Pareto distribution's minimum value using statistical hypothesis testing and Austrian data from the HFCS. They assume that, due to the skewness of wealth distributions, there is a non-observation bias at the top in survey data, as very rich households are randomly missing from the sample. Vermeulen (2014), on the other hand, based his simulation on the assumption that the under coverage at the top does not solely happen by chance, based on several reports on response rates in the US Survey of Consumer Finances (SCF), instead concluding that the response rates decrease due to differential non-response (as reported by Kennickell and Woodburn 1997; Kennickell and McManus 1993). The term encapsulates the observation that the non-response rate is increasing, the higher the net worth value of a household is, i.e. the richer a household, the lower the probability that it is included in a survey sample. Survey data alone then yields severely downward biased results, even more so without a dedicated oversample for very rich households, as included in the SCF (Kennickell 2007, Kennickell 2009) and some countries that are part of the HFCS sample (European Central Bank 2013a). The 2011 Spanish subsample of the HFCS includes such an oversample for very wealthy households; it is based on individual wealth tax file information from the 2007 wealth tax.[3] All people with taxable wealth over 108,000 euros in Spain were subject to this tax. The wealth strata were chosen based on the percentile distribution of households filing a wealth tax return. The resulting cooperation rates show that non-participation is much less likely for the middle class (wealth below 500,000 euro) than for the upper class (wealth greater than 6 million euro). The former strata had cooperation rates exceeding 50 percent, the latter below 26.5 percent (Bover et al. 2014, page 27). Findings from the 2002, 2005, and 2008 Encuesta Financiera de las Familias conducted in Spain confirm the progressively increasing non-response rates (Bover 2004, 2008, 2011). As for the mechanisms that might cause differential non-response, they remain largely unexplored, but it is straightforward to think of a series of unknown household characteristics that might correlate with both net worth and response probability (e.g. availability of the household's head and time use, the value of opportunity costs), none of which are observed by survey providers or used in the post-stratification of the survey weights, thus yielding weights that do not reflect the distribution of households in the sampling population.

In section 2, a series of Monte Carlo experiments is conducted, adjusting several of the

[3] The wealth tax in Spain was discontinued afterwards, but re-established in 2011.

assumptions and testing various methods used by researchers to correct for the under coverage of high-net-worth-individuals and -households in survey studies. In section 3 the findings are applied to to German survey wealth data, showing how re-calibrating the survey weights might affect top wealth shares, and compares the results. Section 4 concludes.

2. A simulation study

The correction for the missing rich in survey data, using a Pareto distribution as an approximation for the upper tail, involves several steps. First, the parameters of the Pareto distribution must be identified. In section 2.1, non-observation bias is illustrated using a similar simulation set-up to Eckerstorfer et al. (2015). It is shown that both sample estimates and Pareto-corrected estimates become more precise as more observations are sampled from the tail of a wealth distribution (specification 1). Next, the assumption of non-observation bias as the motivating factor is changed to differential non-response and it is shown that non-informative survey weights, ceteris paribus, result in downward biased estimates, even though a Pareto-correction is applied (specification 2).

2.1. Non-observation bias versus differential non-response

It is assumed that the top net worth population of a fictive country consists of 600,000 households with a net worth greater than 1 million; they are distributed following a Pareto distribution

$$f_p(w) = \begin{cases} 0 & w < w_m \\ \frac{\alpha w_m^{\alpha}}{w^{\alpha+1}} & w \geq w_m \end{cases}, \tag{1}$$

w_m is the threshold parameter of a pareto distribution, also called minimum value. All data exceeding this threshold follow a Pareto distribution. The parameter α is known as the Pareto index or the scaling parameter, which determines the shape of the distribution–the lower α the higher the inequality of the wealth distribution in the upper tail exceeding the threshold parameter w_m.

In the first Monte Carlo experiment, α is set to 1.3 and w_m equals 1 million (cf. Eckerstorfer et al. 2015). To estimate the Pareto index from the data, the maximum likelihood estimator is used, as it is the preferable estimator compared to the regression estimator (Clauset et al. 2009). The units of the sample are denoted by $i = 1, \ldots, n$, hence, w_i equals the net worth of household i. Then, the maximum likelihood estimator for Pareto index α is given by

$$a_{ml} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{w_i}{w_m} \right]^{-1}. \tag{2}$$

Samples with varying sample sizes are drawn from this population: the number of Pareto distributed households n varies between 100 and 2,000 observations in steps of 100.[4] As each step involves 1,000 samples, in total 20,000 samples are drawn and the respective Pareto index is calculated using equation (2).

In contrast to Eckerstorfer et al. (2015), the x-axis does not show the sample size as a percent of the population, it is shown in absolute numbers. The precision of the estimation depends on the absolute sample size rather than the relative sample size, i.e. the overall size of the population is only relevant for the extrapolation of the aggregate wealth. As the sampled total net worth in the top panel of figure 1 depicts, the median aggregate wealth is likely biased downwards before correction. However, estimates of the Pareto index are unbiased (against the median) and the precision expectedly gets higher as the sample size increases. On the lower right panel in figure 1 the total net worth is recalculated, based on the extrapolation of the Pareto estimates. As it is known how many units exceed w_m from the overall population, it is straightforward to calculate the resulting total net worth via the expected mean:

$$E(W) = \begin{cases} \infty & \alpha \leq 1 \\ \frac{\alpha w_m}{\alpha - 1} & \alpha > 1 \end{cases}. \tag{3}$$

[4] Based on my own calculations, the 2012 sample of the German Socio-economic Panel Study (SOEP) has 270 households that have a net worth exceeding 1 million euro, in the German subsample of the HFCS data there are 246 households with a net worth greater than 1 million euro (means over 5 implicates, see table 2). In the Austrian subsample only 113 households exceed 1 million, while in Belgium, Finland, and Italy the number varies between 200 and 300 households. In Spain and France the number is well above 1000 households, which in turn seems to greatly affect variance between separate implicates of the multiply imputed data (see Appendix C).

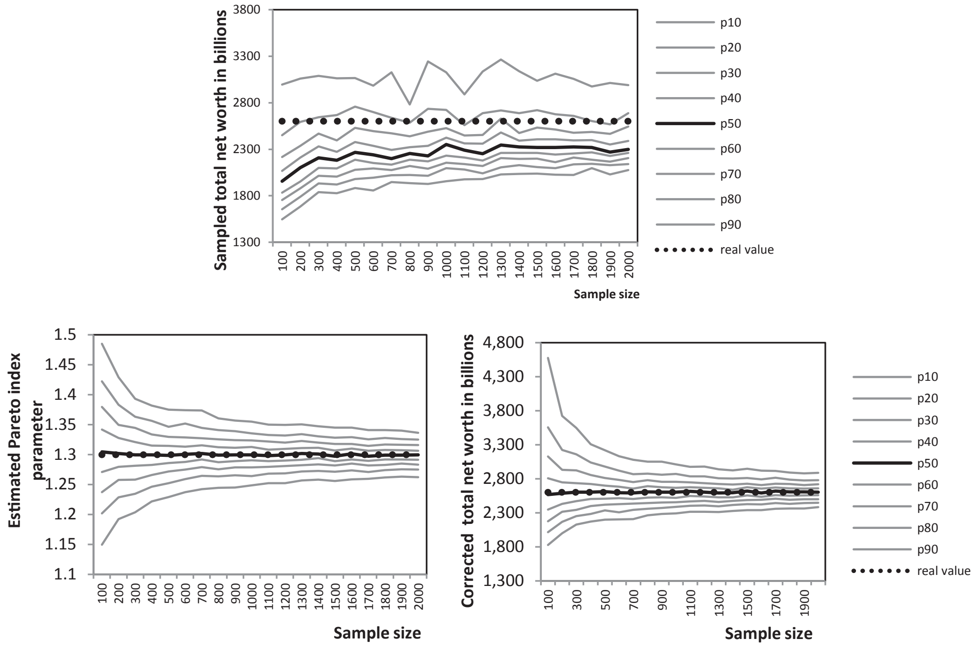

Figure 1: Specification 1. Deciles of the estimated a_{ml} and total net worth for sample sizes between 100 and 2,000 of a Pareto distributed population with an actual $\alpha = 1.3$, $w_m = 1{,}000{,}000$ and population size $N = 600{,}000$.

The corrected estimates for the total net worth are unbiased, the precision increases with the sample size. In this case, any downward bias would be the result of a non-observation bias. Sample selection randomly excludes very rich households and the resulting estimated totals are too low.

However, surveys in Spain and the US show that the non-response rate is increasing with the level of wealth (Kennickell and Woodburn 1997; Bover et al. 2014, page 27). Hence, differential non-response might be a more viable explanation for the lack of statistical power at the top of wealth distributions and biased top wealth shares in survey samples. Eckerstorfer et al. concentrate their study on a statistically sound method to determine the correct minimum value w_m and apply the method to HFCS data from Austria, assuming only a non-observation bias. They note that their method could also be applied if the data suffer from differential non-response. As shown next, differential non-response yields biased estimates of Pareto index α and, thus, biased totals, even if both the threshold parameter w_m and the population size exceeding w_m are known.

In specification 2, the same Monte Carlo experiment as in specification 1 is repeated, assuming that the probability to refuse to participate in the survey increases with the level of wealth. For this simulation, any assumption of the mechanism would suffice, as long as it progressively increases the probability of non-response with the level of net worth. Vermeulen (2014) calculates non-response probabilities for the 1992 sample of the U.S. Survey of Consumer Finances, as reported in Kennickell and Woodburn (1997): the mechanism is (approximately) described by $\Pr(\text{non-response}) = 0.1 + 0.04 \ln(w)$. This means a person with a net worth value of 1 million would refuse with a probability 65.3% and a person with a value of 10 million with a probability of 74.47%.[5] However, it is assumed that the survey provider increases the gross sample size by a factor of 3 in order to ensure that the net sample size stays roughly the same as in the first simulation. Implementing the same mechanism in our data and drawing random gross samples of, again, increasingly large gross sample sizes between 300 and 6,000 units yields net samples of roughly the size as in specification 1.

In figure 2 we estimate Pareto index α, again using the maximum likelihood estimator a_{ml}. However, the true mechanism of non-response is unknown to the researcher; hence, the empiricist assumes a random sample. This assumption is empirically warranted if no external data for calibration is available. Additionally, it is mathematically identical to an estimation of the Pareto index α from survey data with complex sampling without using weights as in Eckerstorfer et al. (2015).[6]

The total net worth calculated from the sample is severely biased downward (top panel in figure 2). Additionally, estimating Pareto index α using the survey data with an unknown non-response mechanism results in an overestimation of the Pareto index (bottom left panel in figure 2). In the Monte Carlo experiment the estimates are about 0.14 units too high. Hence, the first result is that if the survey data are plagued by increasing non-response rates, then a_{ml} cannot be consistently estimated. The effect of an overestimation of α leads to a corrected total net worth that is consistently too low (figure 2, bottom right

[5] Low response rates are not unique to surveys in the US. The response rate for the first wave of the German HFCS subsample was as low as 18.7%; overall the rate was below 50% in about half of the countries included in the HFCS (European Central Bank 2013a, page 41).

[6] In practice, the mechanism of non-response is never completely unknown to survey providers and unit non-response is addressed when calculating the survey weights, which combine sampling probabilities with additional paradata, that are possibly correlated with the wealth level of households. Thus, including survey weights in any estimation of Pareto index α is advisable, especially if an oversample of rich households is included in the sample. This was the case in Germany, Austria and several other countries included in the HFCS (European Central Bank 2013a). Estimation without survey weights will surely yield biased results stemming from the complex survey sampling.

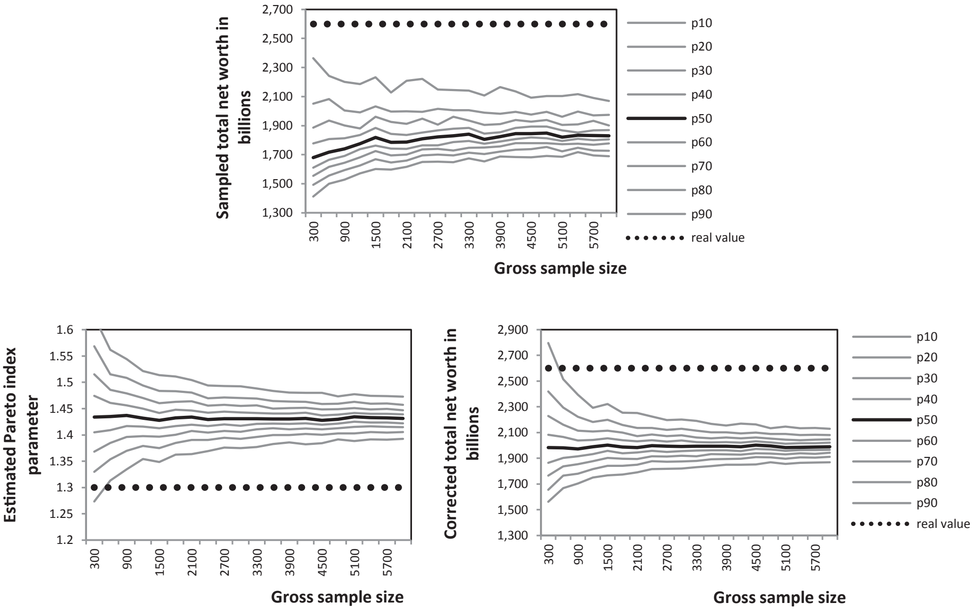

Figure 2: Specification 2. The impact of differential non-response on the estimation of α. Deciles of the estimated a_{ml} and total net worth for gross sample sizes between 300 and 6,000 of a Pareto distributed population with an actual $\alpha = 1.3$, $w_m = 1{,}000{,}000$ and population size $N = 600{,}000$.

panel), in this case, roughly 550 Billion or 20% less than the real value. Note that it is still assumed that both the value of the threshold parameter w_m and the number of households exceeding it are known, both of which are unknown in practice. In Appendix A it is shown analytically that the results of specification 2 are caused by differential non-response, as the sample probability density function (pdf) differs from the population's pdf.

As shown in specification 2, once the survey weights are uninformed about the underlying non-response mechanism, a simple ML estimation of the Pareto index followed by a re-assessment of the top wealth tail is yielding results that are still biased. Taking this assumption further, in specification 3 it is assessed how the ML estimation of the Pareto index as a function of the threshold parameter, which is unknown in practice, behaves, assuming the sample suffers from differential non-response. In specification 4 the effect of adding rich list data is shown, in which case an empiricist needs to switch to a weighted regression estimator. In addition, specification 5 tests the impact of biased rich list data on the estimates assuming both informed and uninformed survey weights. The theoretical

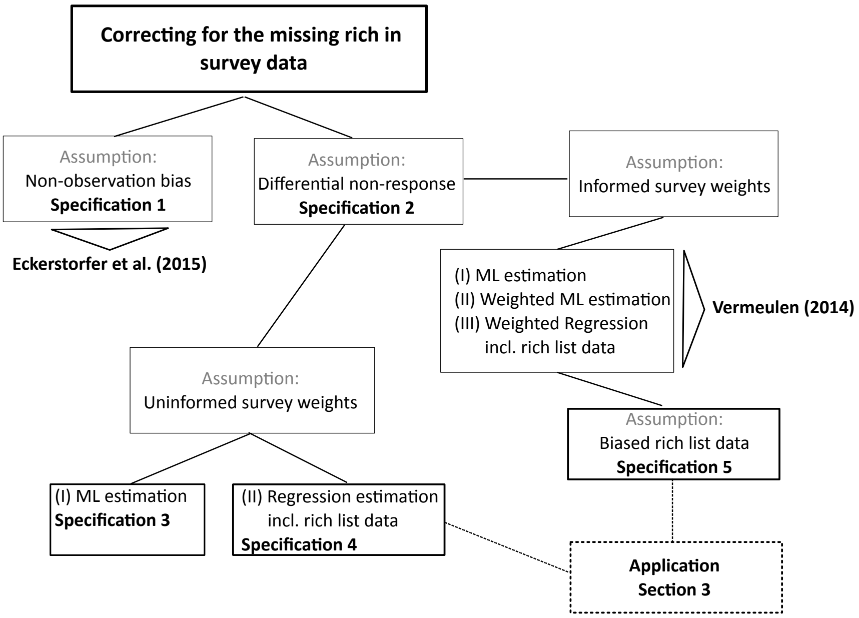

Figure 3: Overview. Correcting for the missing rich in survey data: assumptions, specifications and literature.

results of the simulations will then be compared to German wealth data in an application in section 3; it is also discussed how a re-calibration of survey weights based on assumptions concerning the relationship between response and wealth levels might be conducted.

2.2. Maximum likelihood estimation of Pareto index α as function of the threshold parameter w_m

In a survey environment the threshold parameter w_m is crucial for the correct estimation of Pareto index α but unknown. If we set w_m too low, we include data in the estimation of α that do not follow a Pareto distribution and, thus, will end up with biased results. Eckerstorfer et al. (2015) note, the inclusion of observations below the true minimum value of w_m yields downward biased estimates of α, the exclusion of data above w_m yields upward biased estimates of α. The source cited in Clauset et al. (2009, page 10) simulates the estimation behavior of α as a function of w_m, but the data below the threshold parameter follow an arbitrarily chosen exponential distribution in the referenced simulation. In fact, it is easy to prove another case using a different representation below the threshold parameter. Empirically, the expected value of α does not follow a clear pattern, as the shape of the plot depend on the empirical distribution of the data below the threshold parameter w_m (see Appendix B and C for simulated and empirical results, respectively).

In specifications 3, 4, and 5, a population of 30 million households is assumed and net worth is distributed following a lognormal distribution below a certain threshold w_m and following a Pareto distribution above w_m. This means the wealth distribution is given by

$$f_p(w) = \begin{cases} 0 & w \leq w_m \\ \frac{1}{w\sigma\sqrt{2\pi}} e^{-\frac{(\ln w - \mu)^2}{2\sigma^2}} & 0 < w < w_m \\ \frac{\alpha w_m^\alpha}{w^{\alpha+1}} & w \geq w_m \end{cases} . \tag{4}$$

In effect, this population is characterized by a strictly positive net worth, which does not diminish the results of the simulation, as one may solely look at the upper parts of the distribution exceeding w_m, the Pareto distributed tail. The parameters and sample sizes are chosen to roughly resemble the West German population from Socio-Economic Panel study–with increased wealth concentration at the top–, so that the effect of non-response can be illustrated: With parameters $\mu = 11$ and $\sigma = 1.8$, and the Pareto distribution characterized by w_m=1,000,000 and $\alpha = 1.4$, the result is a population with an aggregate wealth of about 9.9 trillion, which translates to a mean net worth of roughly 330,000, the top percentile holds a share of 37.7% and the top 0.1% a share of 19.5% of the total net worth. It is assumed that an individual (or household) responds to the survey with a probability of P(Response) $= 0.62 + 0.02 * \ln(w_i) - 0.0024 * \ln(w_i)^2$. This relationship between wealth and survey response is directly derived from the 2012 Spanish EFF strata, as documented in Bover et al. (2014, page 27). Figure 4 depicts the prototypical densities of the (log) wealth distribution before and after taking the response probabilities into account, as well as the response probabilities as a function of (log) wealth.

What is the resulting shape if the Pareto index α is estimated as a function of the threshold parameter, which is unknown in practice? In this simulation the ML estimator is used, while the survey weights are uninformed about the non-response mechanism. As the gross sample size in this simulation is 30,000 households, the non-response results in a net sample size of roughly half the size. Figure 5 shows that the resulting estimates of the Pareto index using maximum likelihood estimation are, in this case, against the median,

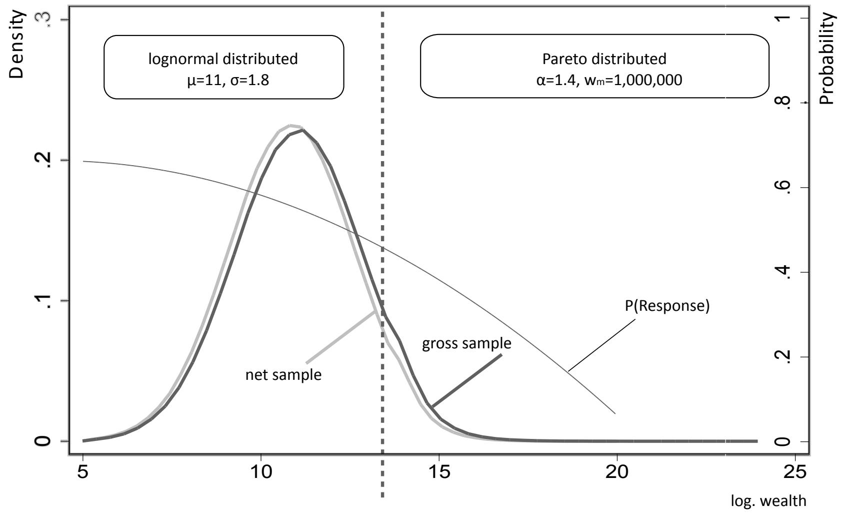

Figure 4: Simulation set-up. Assumptions in the Monte Carlo experiments in specifications 3, 4 and 5.

about 0.15 units too high at the threshold value $w_m = 1$ million. In addition, the higher the assumed threshold value is set, the less precise the estimation of Pareto index α is. An upward biased Pareto index directly translates to downward biased estimates of the inequality of the Pareto distributed wealth. Had a researcher only used the raw sample, he would estimate a total net worth of 6.8 trillion (see figure 5, right panel). The ML estimation of the Pareto index and a simulation of the tail based on the results barely improve on lacking survey data, if non-informative survey weights are included.

By the raw sample, the top-0.1% of the population holds 10.1% of the net worth; the target value would be 19.5%. The re-assessment slightly improves on the raw sample as figure 6 shows: after applying a Pareto correction the top wealth shares are somewhat higher, as the wealth is redistributed from households in the 90th to 99th percentiles, whose wealth is overestimated before correction, to the top percentile. However, while the estimates are slightly improved, they are still far from acceptable. Generally, if the survey weights are uninformed, the number of Pareto distributed households is too low, and their distribution too equal, both before and after correction. For comparison's sake, the same simulation with informed survey weights using a weighted ML estimator is repeated in

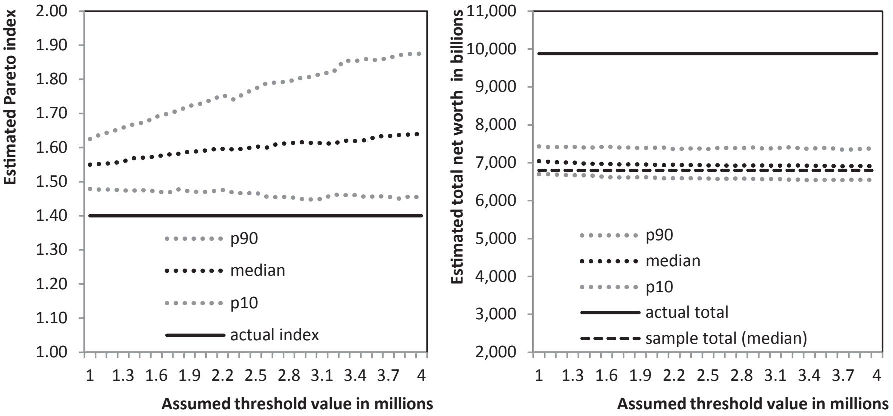

Figure 5: Specification 3. Impact of differential non-response on the maximum likelihood estimates for the Pareto index α and total net worth, plotted as a function of the value assumed for w_m. 1,000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

Appendix D.

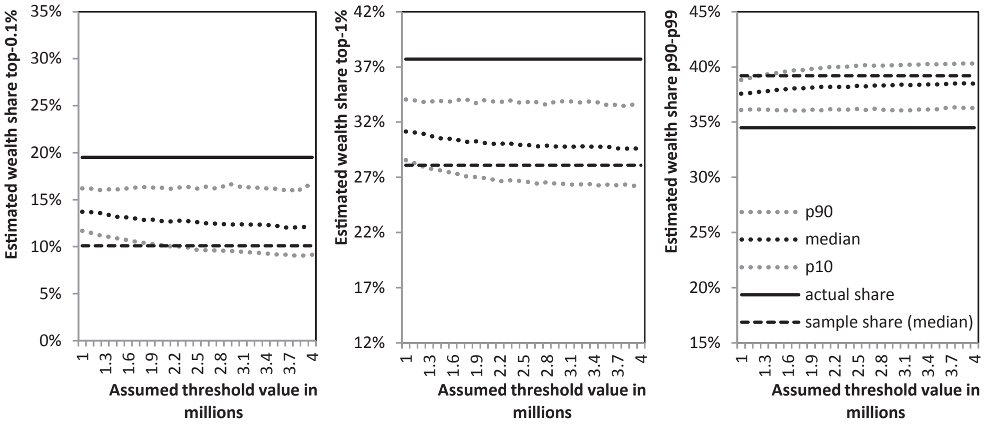

Figure 6: Specification 3. Impact of differential non-response on the top wealth shares before and after Pareto correction, plotted as a function of the value assumed for w_m. 1,000 samples each, drawn from test distributions, Eq. 4, see also Fig. 4.

2.3. The regression method including rich list data

In a next step, it is assumed that information on the net worth of the top 50 net worth holders is available to the researcher from an external source. This is where the so-called rich lists come into play. It is evaluated, whether external data may be used in order to obtain unbiased estimates, if a survey sample suffers from differential non-response and the very rich are missing as a result. As in specification 3, we draw a net sample of about 15,000 households. Then, we add the 50 wealthiest households taken from the base population for the estimation (each carrying a survey weight of one). In this case, using a sample combined of two sources, the maximum likelihood estimator would yield biased results, hence leaving us with the regression method as the only option. As above, N_i denotes the frequency weight of household i; $N_{w>w_i}$ is the sum of the frequency weights exceeding the net worth of household i; thus, it corresponds to the rank of a household, if survey weights are included. The regression estimator then is the estimated parameter from a regression of the log of the net worth on the log of the rank of all households holding a net worth of w_m and higher:

$$\ln(N_{w>w_i}) = c - a_{reg}\ln(w_i). \tag{5}$$

It is assumed that the weights are uninformed, i.e. the frequency weight for any household is the inverse of the sampling probability. The same specification is, again, repeated with informed survey weights in Appendix D.

In figure 7 the results on the left hand side depict the median estimates and selected percentiles of the estimation of Pareto index α. If the non-response mechanism is unknown, α is underestimated at the threshold $w_m = 1000000$ by 0.04 units, and steadily decreases thereafter as less and less households are included in the estimation. Apparently, if the survey sample is affected by differential non-response, including rich list data will result in an underestimation of the Pareto index, the inequality in the top tail is too high. The target estimate of an unbiased total net worth would be 9.9 trillion. An obvious question arises: if the Pareto index α is too low, why are estimates of the aggregate wealth too low? The adjustment for the missing rich here is plagued by two separate biases with countering effects: on the one hand, the Pareto index is too high, resulting in an overestimation of wealth concentration for the Pareto distributed part of the wealth distribution. On the

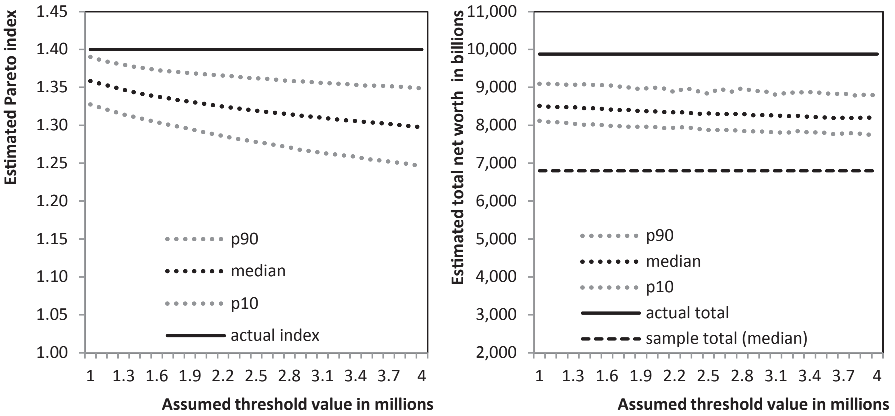

Figure 7: Specification 4. Including rich list data in the regression estimation. Regression estimates for the Pareto index α and corrected total net worth, plotted as a function of the value assumed for w_m. Uninformed survey weights. 1,000 samples drawn from the test distribution, Eq. 4, see also Fig. 4.

other hand, the number of observations above the Pareto threshold w_m is too low. In effect, too few observations are distributed too unequally in the re-assessed tail of the wealth distribution, and the latter effect dominates the computation of the aggregate wealth. Only if the survey weights are informed about the exact mechanism of non-response, is an empiricist able to determine how many households are in the tail of a wealth distribution.

As figure 8 shows, while the aggregate wealth of the population is underestimated by 1.4 trillion, the shares of the wealthiest 1 % and 0.1 % of households are likely to be improved. To be more precise, it is likely that the corrected estimates overshoot the mark: as the re-assessment of the top tail is fed with both wealth data that is biased to the middle-class and rich list data at the very top, a (relative) redistribution from the lower deciles to the top takes place (see sample shares and median of the wealth held by 90th to 99th percentiles), resulting in top wealth shares that are too high and wealth levels held by the upper middle-class that are too low. As the 90th to 99th percentiles are equally important for the aggregate wealth, but their net worth is falsely assessed, the aggregate wealth is still biased downward; the top wealth shares are biased upward.

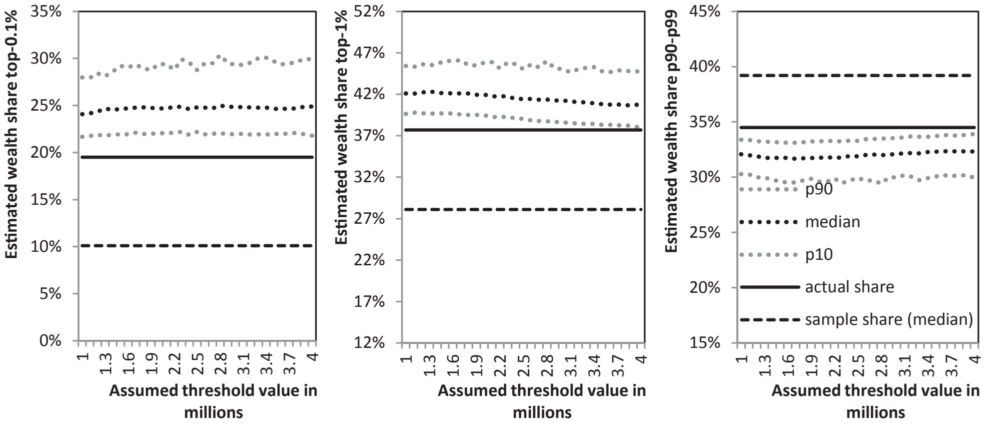

Figure 8: Specification 4. Including rich list data in the regression estimation. Top wealth shares plotted as a function of the value assumed for w_m. Uninformed survey weights. 1,000 samples drawn from the test distribution, Eq. 4, see also Fig. 4.

2.4. The impact of biased rich list data

The dubious nature of data taken from rich lists published in magazines largely remains unresolved. Assuming that mistakes in the journalistic black box are merely random would have a negligible effect on the estimated Pareto indices of the top tail. However, if the lists' entries are too high or too low, they have a significant impact on the estimations. Admittedly, since neither the sources of data nor the method of obtaining the information are made public, the details of such lists ultimately cannot be verified. There are results hinting at an overvaluation of assets in the Forbes magazine. When US federal tax authority researchers compared the tax data of deceased persons and the Forbes list, they discovered that the list overestimated net worth by approximately 50 percent, primarily due to assessment difficulties, fiscal distinctions, and poor assessment of liabilities (Raub et al. 2010).

In the last specification of this simulation study, the possible impact of an overestimation of billionaires' wealth in rich lists on the relevant estimates is assessed. In specification 5a it is assumed that the survey weights are informed about the non-response mechanisms, in specification 5b the survey weights are non-informative. The Pareto index is estimated using the weighted regression estimator including the top 50 rich list entries, however, they are multiplied with a random normal variable with a mean of 1.4 and a standard error of 0.15, resulting in an overestimation of billionaire's wealth by 40 % on average.

	5a - survey weights **informed**				5b - survey weights **uninformed**			
Assumed	Pareto	Total	Wealth share	Wealth share	Pareto	Total	Wealth share	Wealth share
threshold	index	net worth	top-1%	p90-p99	index	net worth	top-1%	p90-p99
w_m **in millions**		in billions	in %	in %		in billions	in %	in %
1.0	1.341	10,910	43.1	31.8	1.332	9,192	46.0	30.0
1.2	1.339	10,900	43.3	31.6	1.322	9,198	46.7	29.5
1.4	1.337	10,910	43.2	31.6	1.315	9,046	46.3	29.4
1.6	1.334	10,910	43.5	31.4	1.310	9,029	46.0	29.4
1.8	1.332	10,910	43.5	31.4	1.305	9,009	46.2	29.4
2.0	1.330	10,870	43.2	31.6	1.302	8,984	46.2	29.4
2.2	1.328	10,730	43.2	31.7	1.299	8,979	45.9	29.3
2.4	1.327	10,880	43.4	31.4	1.296	8,948	45.9	29.5
2.6	1.328	10,830	43.1	31.5	1.293	8,865	45.6	29.7
2.8	1.328	10,770	43.0	31.7	1.289	8,821	45.4	29.7
3.0	1.329	10,780	43.1	31.6	1.286	8,955	45.6	29.6
3.2	1.327	10,720	43.0	31.6	1.283	8,849	45.1	29.9
3.4	1.327	10,760	42.8	31.7	1.279	8,808	44.9	30.1
3.6	1.326	10,710	42.7	31.6	1.275	8,774	44.9	30.0
3.8	1.325	10,750	42.8	31.6	1.274	8,833	44.9	30.0
4.0	1.324	10,730	42.9	31.6	1.273	8,753	44.7	30.1
Actual values	1.400	9,875	37.7	34.5	1.400	9,875	37.7	34.5
Before correction (at $w_m = 1$ million)	1.547	6,800	28.1	39.2	1.547	6,800	28.1	39.2

Table 1: Specifications 5a and 5b. Median estimates weighted regression method using biased rich list data. Estimated Pareto index, total net worth and shares as a function of the value assumed for w_m. Test distributions given by Eq. 4, see also Fig. 4.

The impact of biased rich list entries on the estimation of the Pareto index is severe. Table 1 shows the results, depending on the assumed value of the threshold parameter, as well as the actual values and median values calculated from the raw samples (at $w_m = 1$ million). At the threshold value, the index is about 0.07 units too low and appears to decrease slightly thereafter. If the survey weights are non-informative, the Pareto index quickly decreases, the higher the threshold value is assumed to be. The impact on the corrected total net worth is substantial. Against the median the total net worth is either about 1 trillion too high, or at least 700 billion to low, which in turn means that, if both the rich list is biased upward and the weights are uninformed, the two biases somewhat offset each other. However, in both cases this directly translates to top wealth shares that are consistently too high; the aggregate wealth is then missing from the upper middle class (p90-p99). Assuming that an empiricist's survey weights are informed about the correct non-response mechanism, the impact of biased rich list data is less severe than in specification 4. Empirically, an empiricist might want to factor in both effects in the estimations, as the rich list data are likely to be biased upward and the survey data probably suffer from differential non-response. Overall, it remains questionable whether researchers might want to put a lot of confidence in the results of such an exercise. Moreover, in a study by Brzezinski (2014), Pareto distributions were consistent with the distribution of rich list data covering U.S., China, and Russia in only about one-third of the cases. This leads to the question of whether there might be alternative distributional assumptions or estimation techniques for simulating the super-rich population that is otherwise omitted from survey data. On the other hand, Capehart (2014) finds that the results of goodness-of-fit tests might change for the positive once researchers take measurement error into account.

3. Application: German survey data

The primary goal of the German Socio-Economic Panel (SOEP), and other similar surveys, is not to measure the share of the top-1-percent wealth holders. These surveys serve multiple purposes that might not even be related to wealth (for an overview of the SOEP survey, see Wagner et al. 2007). However, the claim to be representative for the whole population and certain features of the surveys make them useful for the task at hand. Additionally, as there is no tax or register data available in Germany–similar to most other countries–survey data remains as the last trustworthy and publicly available source for a scientific analysis of the distribution of wealth and wealth inequality.

As shown in section 2, once survey data suffers from differential non-response, both ML estimation without rich list data and regression estimation including rich list data will yield biased results, as both the Pareto index and the number of households exceeding the threshold value are biased. However, both biases may be corrected if the survey weights are informed about the non-response mechanism. To be more precise, the survey weights need to be explicitly informed about the relationship between non-response and wealth. Thus, this section provides an illustration: How to inform survey weights about wealth-related non-response using exogenous information and, hence, obtain the correct distribution of household net worth. Notwithstanding this announcement, it turns out that valid exogenous information is of utmost importance, as depending on the source the exercise produces wildly different estimates.

The German Socio-Economic Panel Study (SOEP) is a longitudinal representative survey collecting socio-economic information on private households in Germany. Additionally, a module collecting wealth information was included in 2002, 2007 and 2012. In 2002, the SOEP sampled high-income individuals for the last time, it is reasonable to assume that the precision at the top of the wealth distribution was much higher in 2002, and slowly decreased afterwards due to panel attrition. Table 2 summarizes the data with regard to net worth of private households.

The framework we use to estimate the upper margin of wealth distribution is, as in the simulation study, twofold and based on estimation from survey data alone and a combination of survey data with data on the absolute peak of distribution derived from all those with the respective citizenship on the list of billionaires published annually by the US Forbes magazine. However, the Forbes lists does not provide sufficient details every year to be able to determine whether these individuals are also living in the respective country.[7] Likewise, billionaires who are living in one of the countries, but did not hold the respective citizenship, were excluded from the analysis (table 3). In this process, it is assumed that each individual on the Forbes list represents a single household.[8]

Survey wave		2002		2007		2012
Mean		149838		153998		154380
Median		37247		39220		46680
p90		361239		372899		380740
p95		538470		562386		563100
p99		1272189		1375940		1349640
Share of top-1%		21.1 %		21.6 %		18.2 %
Share of top-0.1%		7.6 %		7.1 %		5.3 %
Max. in million euro		62.7		31.7		45.5
Aggregate private wealth in billion euro		5.800		6.116		6.278
Number of households	n	N	n	N	n	N
> 500000 euros	1089	2342967	986	2522275	862	2516656
> 1000000 euros	334	620910	304	683088	270	708424
> 3000000 euros	47	88204	56	133175	42	108366

Table 2: Summary statistics: Net worth of private households in Germany, according to SOEP survey 2002, 2007 and 2012. Source: SOEPv30, own calculations, means over 5 implicates of multiply imputed data.

[7] Moreover, there may also be individuals living in Germany who are not German nationals but should be classified together with other private households.

[8] It is not possible to tell from the Forbes list whether the households of these individuals include other members or not.

	Germany 2002	Germany 2007	Germany 2012
Number of entries	34	55	55
Aggregate wealth in billion euros	159.8	185.4	188.7
Max. in billion euros	30.9	15.1	19.1

Table 3: Entries in the Forbes list of billionaires at the time of SOEP survey wealth modules. US Dollar-Euro exchange rates as of March 1 of the respective years. Source: own calculations based on Forbes magazine's yearly-published list of billionaires.

In the first step, the ML estimator (see spec. 3 in section 2.2) and the regression estimator including rich list data (see spec. 4 in section 2.3) are applied to the 2002, 2007, and 2012 SOEP wealth data. It is shown how the estimation of the Pareto index, as a function of the threshold parameters, yields shapes that are reminiscent of the results in section 2.2, leading to the assumption that the data suffers from differential non-response. The empirical results are highly volatile, often driven by very few observations leaving the estimation, as the threshold parameter w_m is set higher (figure 9, left panel). Furthermore, the left hand side, depicting the weighted maximum likelihood results, shows that any regular shape hinting at an empirical w_m is missing. However, the shapes are reminiscent of what is shown in the Monte Carlo experiment simulating survey data with differential non-response: After decreasing estimations for α, there is a local minimum between 1 and 1.6 million euros. However, to locate w_m and α in this corridor would certainly be bold. The margins of error are large, given the sample size, and the Pareto index α certainly is a good deal too high (see spec. 3). Including the survey weights in the estimation would only offset the effects of differential non-response, if the weights reflect the true response probabilities along the distribution of wealth (see spec. 3b). Once the weighted regression method is applied and the respective German rich list members from the Forbes magazine are incorporated, the curves align and do not vary a lot between the survey waves (figure 9, right panel). There are two main results for regressions including rich list data: (1) once survey weights are informed about the non response mechanism, which results in an unbiased estimation of the Pareto index α after the true w_m is reached, then the curves turn into a straight line at the true Pareto index in the simulation (Vermeulen 2014, see also spec. 4b in Appendix D). (2) If the survey weights are uninformed, the estimates for Pareto index α decreased steadily, while already being downward biased at w_m (spec. 4). The empirical results using SOEP wealth data are more reminiscent of the latter case, hinting at differential non-response in the data.

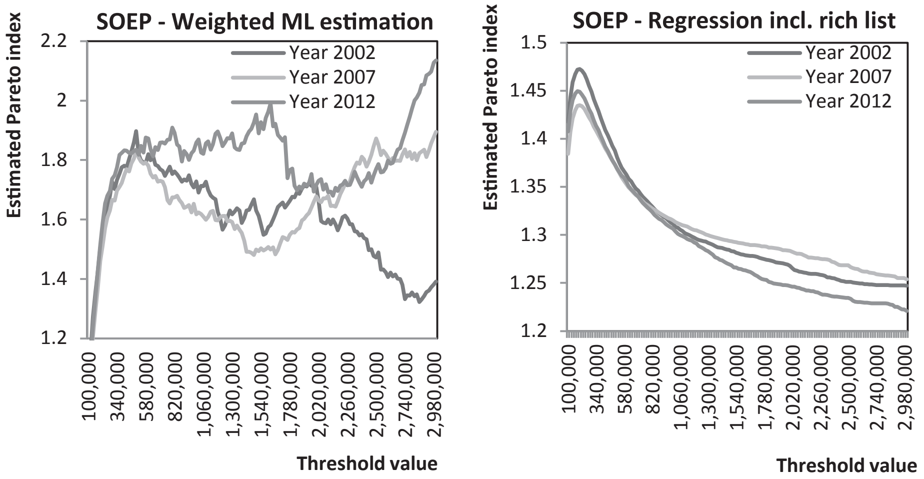

Figure 9: SOEP survey 2002, 2007, 2012. Estimation of Pareto index as a function of threshold parameter w_m. Means over 5 implicates of multiply imputed data. Source: SOEPv30, Forbes' list of billionaires, own calculations.

Next, the calculated Pareto indices, as a function of the Pareto threshold value shown in figure 9, are used to compute the aggregate private wealth in Germany. Using weighted maximum likelihood estimators and re-simulating the data barely impacts the estimates for aggregate wealth. Originally, the total net worth varied between EUR5.8 trillion in 2002 and EUR6.2 trillion in 2012. As shown in the left panel of figure 10, the estimated values are rather close to the ones observed without correction for the missing rich. This is to be expected if one solely tries to correct for non-observation bias but the data suffers from differential non-response. If differential non-response is assumed and the regression method plus rich list data from the Forbes magazine are used, one can expect that the resulting values are less, but still downward, biased (spec. 4). If it is assumed that w_m is at 1 million euro (and not varying over time), the re-assessment adds about 1 trillion euro in 2002, 1.2 trillion euro in 2007 and 1.4 trillion euro in 2012.

Re-calibrating survey wealth data based on external sources

In order for a re-assessment of top wealth with survey data to work, the most obvious first solution is to reweight the data, as it is biased to the middle class using non-informative weights (see figure 4). The steps include: (1) Based on the response as a function of wealth

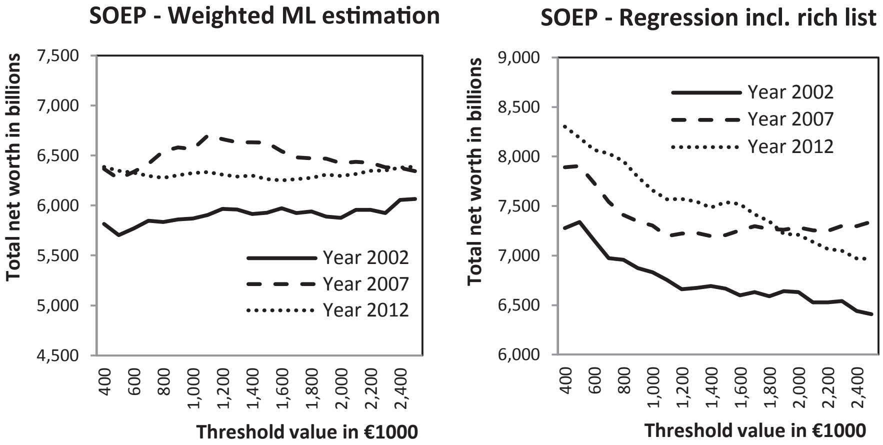

Figure 10: SOEP survey 2002, 2007, 2012. Total net worth and top wealth shares based on corrected data using regression method including rich list data, as a function of threshold parameter w_m. Means over 5 implicates of multiply imputed data. Source: SOEPv30, Forbes' list of billionaires, own calculations.

each household is assigned both the probability to respond and the inverse probability. (2) A household's inverse probability is multiplied by its uninformed frequency weight.[9] (3) The aggregate number of households is divided by the sum of (2), yielding a scalar to adjust the weights to the population size. (4) Each household's new frequency weight is then given by product of a household's specific value of (2) and the scalar (3). For this reweighting to be applicable, the functional form of non-response needs to be known, which is not the case for Germany and the SOEP data. As they are the only source of information on the subject matter, the Spanish EFF strata (Bover et al. 2014) or the U.S. SCF strata (Kennickell and McManus 1993) may be used to describe response probabilities as a function of (log) wealth. Alternatively, some financial institutions release their own reports on the (global) distribution of wealth. For instance, Credit Suisse's report delivers enough information to determine the varying response probabilities depending on the level of wealth in Germany. Hence, using the equation below, the SOEP survey weights can

[9] The uninformed frequency weight refers to the household's weight as provided through the survey distributor. For instance, in the SOEP data there is an oversample for East German households and one would want to preserve the ratio between East and West after re-calibration.

be transformed to mirror the number of households within the strata provided by Credit Suisse for 2012 for the respective reporting year (Credit Suisse 2012).

- **Spanish EFF:** $\text{P(Response)} = 0.622 + 0.020 * \ln(w_i) - 0.00242 * \ln(w_i)^2$
- **U.S. SCF:** $\text{P(Response)} = 0.425 + 0.021 * \ln(w_i) - 0.00183 * \ln(w_i)^2$
- **Credit Suisse:** $\text{P(Response)} = 0.340 + 0.023 * \ln(w_i) - 0.00256 * \ln(w_i)^2$

In all three cases including a quadratic term substantially increases the fit with the empirical strata. Note that such a re-calibration mathematically only affects the relative probabilities to respond between two households with differing wealth levels as it adjusts the frequency weights according to the functional form, even though the overall non-response rate in the SOEP might differ. The re-calibration is completely relative. It does not affect statistical power or anything else. It is best illustrated as revoking the net sample to the gross sample in figure 4, thus, eliminating the effect of differential non-response on the household level.

The exemplary result for the 2012 SOEP data is shown in figure 11. For once, the Pareto indices are still decreasing after re-calibration using EFF and SCF strata, hinting at either the wrong functional form of the non-response generating mechanism–the middle-class bias might be more severe–or the Pareto distribution does not yield a good fit for the top tail of the German wealth distribution. For the Credit Suisse re-calibration, the Pareto indices are slightly, but steadily, increasing. Assuming the threshold value would be at about 2 million euros, inequality at the top is lowest for Credit Suisse and highest for the SCF strata. Note that it is the same for aggregate wealth (figure 11, top left panel), meaning that the Credit Suisse strata yields a substantially less unequal Pareto distribution, but also substantially higher wealth levels overall. Trusting in their data would mean that the SOEP survey underestimates upper middle class wealth (say, p95-p99), as compared to the EFF and SCF non-response assumptions, while the shares of the top-1% and top-0.1% are about the same or lower. Likewise, if one does put trust into the non-response mechanisms as reported for Spanish or U.S. surveys, the top wealth shares are considerably higher than for the raw survey data and slightly higher than for the Credit Suisse strata. Note that in every case, the Pareto indices are higher than without re-calibrating the weights and using rich list data (cf. figure 10), meaning less inequality at the top. The aggregate wealth is also higher, depending on the re-calibration between 1.5 and 5 trillion euros.

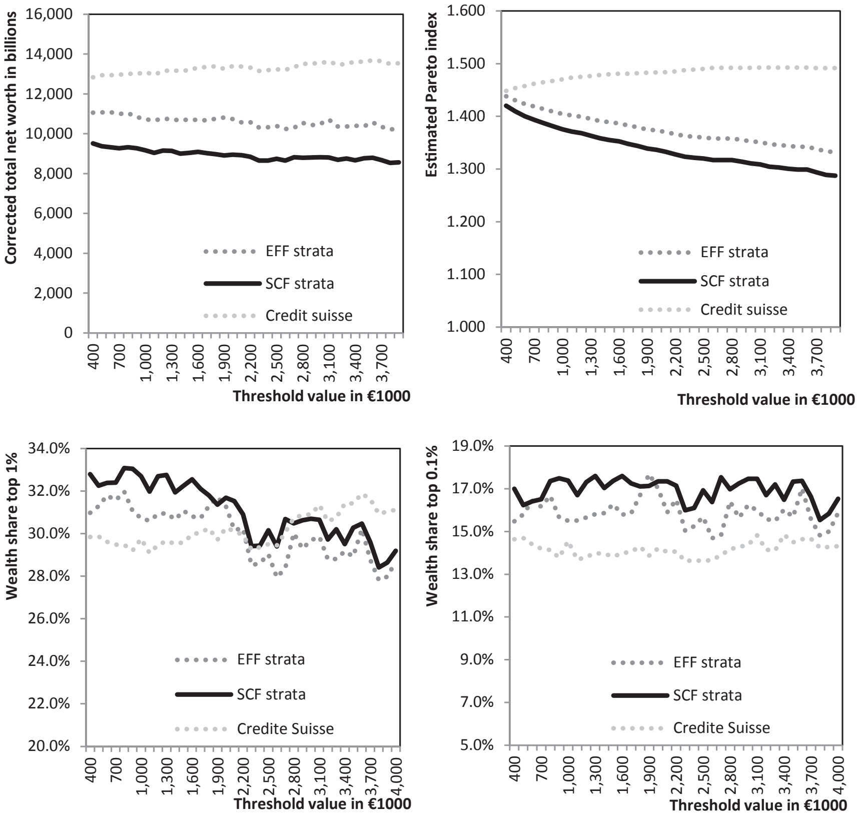

Figure 11: SOEP survey 2012. Total net worth, estimated Pareto index and top wealth shares after re-calibration of survey weights. Data sources: SOEPv30, Forbes' list of billionaires as of March 2012, Bover et al. 2014, Kennickell and McManus 1993, Credit Suisse 2012, own calculations.

As shown, re-calibrating the survey data using exogenous sources has a much more severe impact on the estimates than choosing a threshold value or setting the Pareto indices (see variation in figures 9 and 10). Thus, the most powerful source of bias in measurement of top wealth shares in survey data are non-informative survey weights. The results in this section show that the resulting corridor, depending on the non-response assumptions, is very broad. Moreover, from an empiricist's point of view, none of the results above could reasonably be ruled out from the outset. The balance sheet data, as provided by the German Federal Statistical Office (Destatis 2015), uses different definitions and delimitations of household wealth as compared to survey data, which means it is not advisable to compare it directly to survey data (Grabka and Westermeier 2015). For now, there is no benchmark available. As all three assumptions on the non-response are within the realm of possibilities, the results leave us with a glaringly wide corridor of possible values for aggregate private wealth and top wealth shares. The validity of exogenous data–and the rich list data–remains a matter of trust on the part of empiricist.

4. Summary and conclusion

It is safe to say that in countries without reliable tax return data, or otherwise obtained register data, on the distribution of wealth, policy makers remain largely uninformed about the extent of wealth concentration at the top. Empiricists started improving upon lacking survey data by assuming a Pareto distribution at the top and computing new estimates with and without rich list data.

As other simulation studies show, once the survey weights are informed about the relationship between response and wealth, the weighted maximum likelihood estimator is unbiased. Regression estimation including rich list data then improves the precision of the estimates (Vermeulen 2014). However, the simulations conducted in this study show that if the data suffer from differential non-response and the survey weights are uninformed, then the maximum likelihood estimator yields estimates of the aggregated wealth and top wealth shares that are biased downwards. Adding rich list data and switching to a regression estimation falls short of compensating for the bias, as it underestimates the total net worth, while it overestimates top wealth shares, as too few households in the tail are distributed too unequally. Moreover, there is strong indication that rich list data, such as the yearly published Forbes list, actually overestimate billionaires' wealth, which in turn yields estimates of the top wealth shares and aggregate wealth that are systematically too high. Researchers are not readily able to assess these biases.

The best remedy for lacking survey data is a re-weighting of the survey weights based on either additional assumption or valid data. As such data is not available for Germany, the Spanish EFF and the US SCF response probabilities as a function of wealth were applied in the application of this study to hypothetically show how additional data might be used to compensate for differential non-response. Additionally, households are re-weighted to match their distribution in Credit Suisse's 2012 wealth report. Before re-calibration, both the ML estimator without rich list data and the regression including rich list data yield estimates of the aggregate wealth that might still be biased downwards. After re-calibration, the aggregate wealth was higher by more than 1.5 or 5 trillion euros, depending on the assumed non-response mechanism. The 2012 top wealth shares of the top-1% and the top-0.1% increase by more than 10%. However, all estimations depend on the empirically unknown threshold parameter, the assumed relative response probabilities of the households–which might shape up differently to other countries such as Spain or the US–and the assumption that wealth is actually distributed following a Pareto distribution at the top.

If anything, the findings emphasize the need to use exogenous information in sample design, which allows for creating appropriate weights taking non-response into account. Survey providers must know the exact response probabilities to offset the effects of differential non-response as well as to calculate totals and top wealth shares reliably. Only then can developments in the long run be reasonably analyzed. Until more exhaustive data sources are accessible to researchers–or tax authorities are willing to cooperate more closely with survey providers–it might be a more viable choice to put the efforts in steadily well-run surveys that include dedicated oversamples of high-net-worth-households.

References

Bach, S., Beznoska, M., and Steiner, V. (2014). A wealth tax on the rich to bring down public debt? revenue and distributional effects of a capital levy in germany. *Fiscal Studies*, 35:67–89.

Bover, O. (2004). The spanish Survey of Household Finances (EFF): description and methods of the 2002 wave. Occasional Papers 0409, Banco de Espana.

Bover, O. (2008). The spanish Survey of Household Finances (EFF): description and methods of the 2005 wave. Occasional Papers 0803, Banco de Espana.

Bover, O. (2011). The spanish Survey of Household Finances (EFF): description and methods of the 2008 wave. Occasional Papers 1103, Banco de Espana.

Bover, O., Coronado, E., and Velilla, P. (2014). The spanish Survey of Household Finances (EFF): description and methods of the 2011 wave. Occasional Papers 1407, Banco de Espana.

Brzezinski, M. (2014). Do wealth distributions follow power laws? Evidence from ’rich lists’. *Physica A: Statistical Mechanics and its Applications*, 406:155–162.

Capehart, K. (2014). Is the wealth of the world’s billionaires not Paretian? *Physica A: Statistical Mechanics and its Applications*, 395:255–260.

Chatterjee, A., Yarlagadda, S., and Chakrabarti, B. (2005). *Econophysics of Wealth Distributions.* Springer, Milan.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Credit Suisse (2012). *Credit Suisse Global Wealth Data Book 2012.* Credit Suisse Group AG, Zurich.

Destatis (2015). *Sektorale und Gesamtwirtschaftliche Vermögensbilanzen 1999 - 2014.* Statistisches Bundesamt Wiesbaden.

Eckerstorfer, P., Halak, J., Kapeller, J., Schütz, B., Springholz, F., and Wildauer, R. (2015). Correcting for the missing rich: An application to wealth survey data. *Review of Income and Wealth.* Online first.

European Central Bank (2013a). The Eurosystem Household Finance and Consumption Survey. Methodological report for the first wave. Statistics Paper Series 1, European Central Bank, Franfurt am Main.

European Central Bank (2013b). The Eurosystem Household Finance and Consumption Survey. Results from the first wave. Statistics Paper Series 2, European Central Bank, Franfurt am Main.

Grabka, M. M. and Westermeier, C. (2015). Real net worth of households in Germany fell between 2003 and 2013. *DIW Economic Bulletin*, 5(34):441–450.

Kennickell, A. B. (2007). The role of oversampling of the wealthy in the Survey of Consumer Finances. Survey of Consumer Finances Working Paper, Federal Reserve Board.

Kennickell, A. B. (2009). Getting to the top: Reaching wealthy respondents in the SCF. Paper prepared for the 2009 Joint Statistical Meetings, Washington, DC.

Kennickell, A. B. and McManus, D. A. (1993). Sampling for household financial characteristics using frame information on past income. In *JSM Proceedings, Survey Research Methods Section*, volume 47, pages 88–97, Alexandria, VA. American Statistical Organization.

Kennickell, A. B. and Woodburn, R. L. (1997). Consistent weight design for the 1989, 1992, and 1995 SCFs, and the distribution of wealth. Survey of Consumer Finances Working Paper, Federal Reserve Board.

Klass, S. O., Biham, O., Levy, M., Malcai, O., and Solomon, S. (2007). The Forbes 400, the Pareto power-law and efficient markets. *The European Physical Journal B*, 55(2):143–147.

OECD (2015). *In It Together: Why Less Inequality Benefits All.* Organisation for Economic Co-operation and Development, Paris.

Peffermann, D., Krieger, A. M., and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8:1087–1114.

Piketty, T. (2014). *Capital in the 21st Century.* Harvard University Press, Cambridge, MA.

Raub, B., Johnson, B., and Newcomb, J. (2010). A comparison of wealth estimates for America's wealthiest descendants using tax data and data from the Forbes 400. In *National Tax Association Proceedings, 103rd Annual Conference on Taxation*, pages 128–135, Chicago, IL.

Vermeulen, P. (2014). How fat is the top tail of the wealth distribution? Working Paper Series 1692, European Central Bank.

Wagner, G. G., Frick, J. R., and Schupp, J. (2007). The German Socio-Economic Panel Study (SOEP) - Scope, evolution and enhancements. *Schmollers Jahrbuch : Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 127(1):139–169.

Westermeier, C. and Grabka, M. M. (2015). Significant statistical uncertainty over share of high net worth households. *DIW Economic Bulletin*, 5(14/15):210–219.

A. Non-observation bias versus differential non-response in survey sampling

As we saw in section 2, non-observation bias and differential non-response yield not the same results, if one tries to correct lacking survey wealth data. Survey statisticians differentiate between *sampling error* and *sampling bias*. The former case describes a situation where a survey sample's calculated value differs from a population's aggregate private wealth purely by chance, as the value from a sample is never identical to the actual value. In this case, due to the skewness of the distribution, some super rich individuals or household are excluded from the sample. Their inclusion would drive up the mean and, subsequently, the aggregate private wealth. Their absence from the sample causes a *sampling error*.

A *sampling bias*, however, is to be expected if some members of the intended population are less likely included in the sample than others: differential non-response with respect to wealth in this case. For survey statisticians, this is trivial mathematics, for empiricists, maybe less so. For the sake of simplicity, let us assume that private wealth W is independently identically Pareto distributed (see specification 2 in section 2.1 for more details):

$$f_p(w) = \begin{cases} 0 & w < w_m \\ \frac{\alpha w_m^\alpha}{w^{-\alpha-1}} & w \geq w_m \end{cases}. \tag{6}$$

A sampling error is less of a problem, because any sampled value is still drawn randomly from the underlying distribution, thus the expected value equals the population's mean (is unbiased)

$$E_p(W) = \frac{w_m \alpha}{\alpha - 1}. \tag{7}$$

If the survey sample suffers by differential non-response, the probability to be included in the sample depends on w. In the example in specification 2 it was $\Pr(i \in s|w_i) = 0.9 - 0.04 \ln(w_i)$; more generally, one could describe such a response function as $\Pr(i \in s|w_i) = a - b \ln(w_i)$. In this case, the probabilities to be sampled depends on the value

w itself and, thus, differs at different points of the wealth distribution. With a fixed at 0.9, and w_m supposed to be known, the empirically observed sample distribution exhibits much less inequality depending on b. As it is difficult to show analytically, one typically resorts to Monte Carlo experiments. However, for specification 2 it is possible to show that the sample distribution differs from $f_p(w)$.

For some probability density functions (pdf) $f_p(y_i|\theta)$, depending on parameters θ, it is possible to derive the sample pdf of Y_i, defined as $f_s(y_i|i \in s)$, where S denotes the selected sample. It is obtained by application of the Bayes theorem (Peffermann et al. 1998):

$$f_s(y_i|\theta^*) = f_s(y_i|i \in s) = \Pr(i \in s|y_i) f_p(y_i|\theta)/\Pr(i \in s). \tag{8}$$

The parameters θ^* are a function of θ and the parameters indexing $\Pr(i \in s|y_i)$. It is important to note that, as $\Pr(i \in s|y_i) \neq \Pr(i \in s)$ for all y_i, the sample and population probability density functions are different and survey weights derived from the sampling become informative (if available). In most cases empiricists resort to Monte Carlo methods to show the impact of various non-response mechanisms on the population pdf. This paper's assumptions on the differential non-response affecting a Pareto distributed wealth tail generally yield sample pdfs that are not Pareto distributed any more, which becomes visible when examining the resulting shapes. One of the properties of a Pareto distribution is that the conditional probability distribution of a Pareto distributed random variable, given that it is greater or equals a particular value w_1 exceeding the threshold value w_m, is again a Pareto distribution with unchanged Pareto index α but minimum value w_1 instead of w_m. This property does not hold, if the sampling suffers by differential non-response, indicating that the resulting sample pdf $f_s(w_i|\alpha, w_m, \theta^*)$, with θ^* a function of parameters indexing the biased sampling procedure $\Pr(i \in s|w_i)$, is not Pareto distributed (see also counterexamples in Appendix B and D). With population distribution$f_p(w_i|\alpha, w_m) = \alpha w_m^\alpha * w_i^{-\alpha-1}$ and the probability to be sampled conditional on household wealth given as $\Pr(i \in s|w_i) = a - b\ln(w_i)$ the sample distribution can be written as

$$f_s(w|\alpha,w_m,a,b) = (a - b\ln(w))\alpha w_m^\alpha w^{-1-\alpha} / \int_{w_m}^{\infty} (a - b\ln(w))\alpha w_m^\alpha w^{-1-\alpha} \mathrm{d}w. \tag{9}$$

As in specification 2 the wealth distribution was Pareto with an index of 1.3 and a threshold value of 1000000

$$W_i \; Pareto(\alpha = 1.3, w_m = 1000000),$$

the sample distribution reduces to

$$f_s(w|a,b) = \frac{8.20245 * 10^7 * (a - b\ln(w))}{w^{2.3} * (a - 14.5847 * b)}. \tag{10}$$

For comparison's sake one might now plot both population distribution and the sample distribution from specification 2 (with $a = 0.9$ and $b = 0.04$) to visualize that the sample distribution clearly exhibits less inequality and smaller mean (figure A.1).

To prove that the sample pdf necessarily has a smaller mean than the population pdf it is possible to compute the expected means. For a Pareto distributed random variable the expected value is given by (7). Thus, in specification 2 the expected value of the

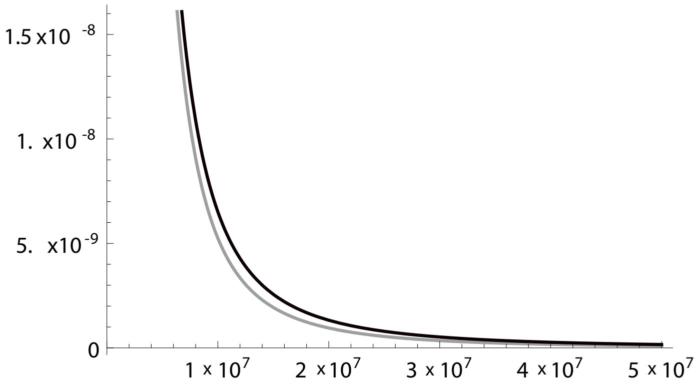

Figure A.1: Population pdf (black) and sample pdf (grey) in specification 2.

population pdf is $4\frac{1}{3}$ million. For the sample population the expected value is given by

$$\begin{aligned} E_s(W) &= \int\limits_{w_m}^{\infty} w f_s(w|a,b)\mathrm{d}w \\ &= \int\limits_{w_m}^{\infty} w \frac{8.20245 * 10^7 * (a - b\ln(w))}{w^{2.3} * (a - 14.5847 * b)}\mathrm{d}w \\ &= \frac{4\frac{1}{3} * 10^6 a - 7.43 * 10^7 b}{a - 14.5847b}. \end{aligned} \tag{11}$$

To get an idea by how much the expected value differs between sample and population pdf depending on different values of a and b, figure A.2 shows $E_s(W)/E_p(W)$ with a fixed at 0.9 (left panel) and b fixed at 0.04 (right panel).

If $b = 0.00$ the probability to be sampled is independent of the level of wealth, and, thus, there is no sampling bias (figure A.2, left panel). The greater b is, the steeper the non-response function, the more the sampled mean is biased downwards. The higher the parameter a is, the higher the overall survey response, the smaller the sampling bias (figure A.2, right panel). In specification 2 the expected downward bias between sampled expected value and the population's expected value is 32.39%. Via extrapolation this directly translates to biased estimates of the population's aggregate private wealth (see sampled total net worth in figure 6).

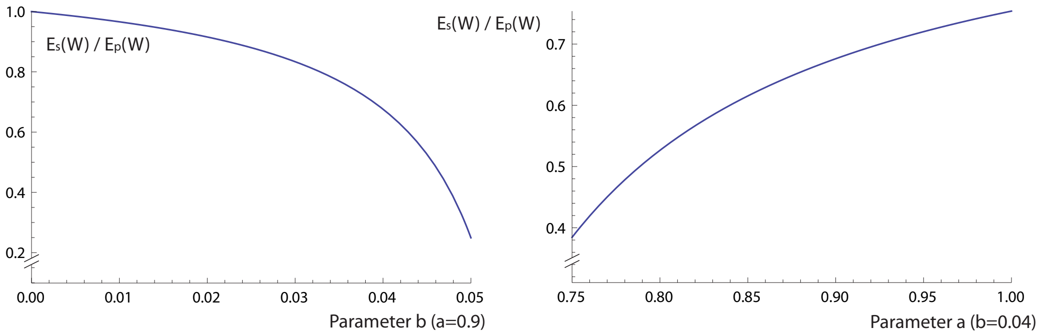

Figure A.2: Expected value of sample pdf as a function of non-response parameters a and b.

B. Simulation: Pareto index as a function of threshold parameter without non-response (ML estimation)

As in Clauset et al. (2009), I compute the estimates for α as a function of w_m, using the maximum likelihood estimator a_{ml}, and examine the resulting plot. In this Monte Carlo experiment, simple random samples of 30000 households are drawn from equation (4), without non-response. Only the median estimates for varying Pareto indices α and threshold values w_m are of interest. This simulation was carried out 1,000 times for each value of w_m.

In figure B.1, w_m is fixed at 1,000,000 and α varies between 1.2 and 1.7, in figure B.2, α is fixed at 1.4 and w_m varies between 750,000 and 1,750,000. We observe that the estimates for Pareto index α increase the higher w_m is assumed to be, until the true Pareto index value is reached, at which point none of the observations of the log-normal distributed samples are included in the estimations. Plotting ML estimates of the Pareto index as a function of w_m with survey data from the German Socio-economic Panel Survey or Euro-area HFCS data yields some results that are fairly close to the shapes in figure 6 for some countries (see section 2.2 and Appendix C).

The simulation shows that the estimated value of α as a function of w_m exhibits a robustly straight line, if the data truly follow a Pareto distribution and there is no non-response. At least a range of values could be given, which, with a very high probability, also includes the threshold value w_m. One would like to choose the value shortly after the plot becomes a straight line. In this case, setting w_m too low leads to results that overestimate the concentration of wealth in the top area (as α is too high). However, in this example it is by set-up of the simulation data that the parameters of the Pareto distribution are easily identified using a plot. In this case, the mode of data generation makes sure that there is a relatively hard transition between log-normal and Pareto distributed wealth. In empirical data the identification of the parameters is hardly as straightforward as in this Monte Carlo experiment and involves a battery of other problems, Clauset et al. (2009) offer a very detailed review of the estimation techniques and possible pitfalls. As the shape of the plot depends on the distribution below w_m in specifications 3 to 5, the resulting Pareto indices below the true w_m are not plotted.

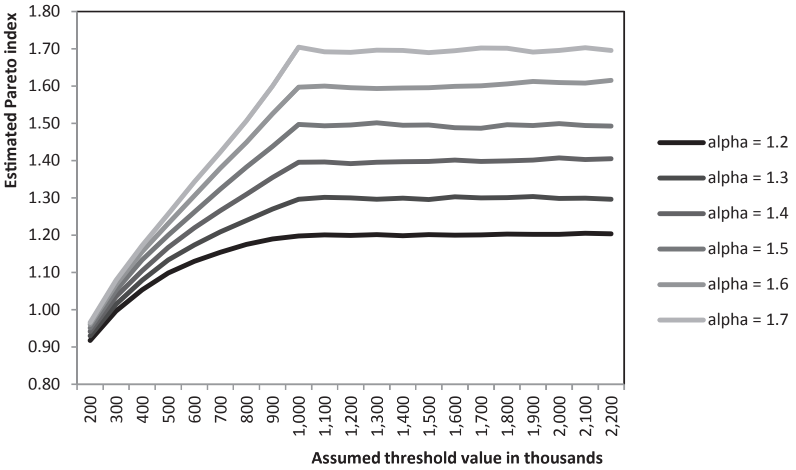

Figure B.1: ML estimation without non-response. Pareto index α plotted as a function of the value assumed for w_m, for various actual α. 1,000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

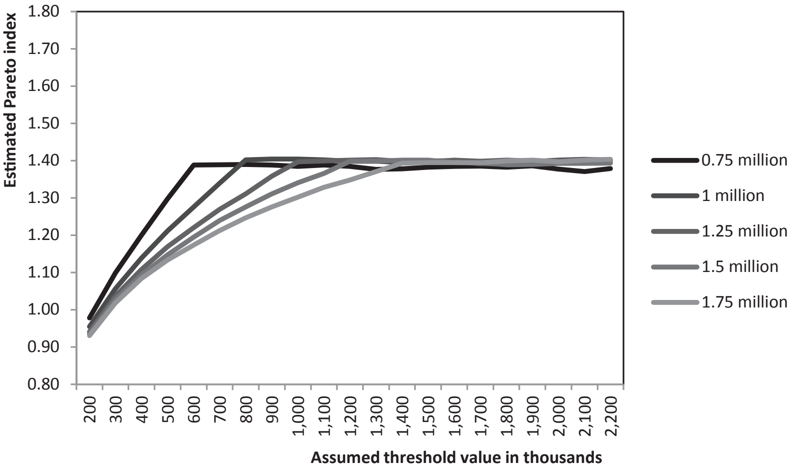

Figure B.2: ML estimation without non-response. Pareto index α plotted as a function of the value assumed for w_m, for various actual w_m. 1,000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

C. Empirical results: Pareto index as a function of threshold parameter using HFCS data (weighted ML estimation)

As seen in Appendix B, estimation of the Pareto index as a function of the threshold parameter may give a hint at the value of the threshold parameter. Here are some results of (weighted) maximum likelihood estimates using the Eurosystem Household Finance and Consumption Survey 2013 (HFCS). Empirically the resulting shapes vary: some are reminiscent of the plots resulting in the simulation, such as net wealth data from Austria, Belgium, France or Italy. Others exhibit a global maximum before seeing a decrease of the Pareto index, such as Finland and Spain. The specific shapes depend on the distribution of wealth below the threshold parameter–assuming that the data are indeed Pareto distributed. Fewer observations are included in the estimation of the Pareto index as the threshold parameter is set higher, hence, the estimates typically become more erratic.

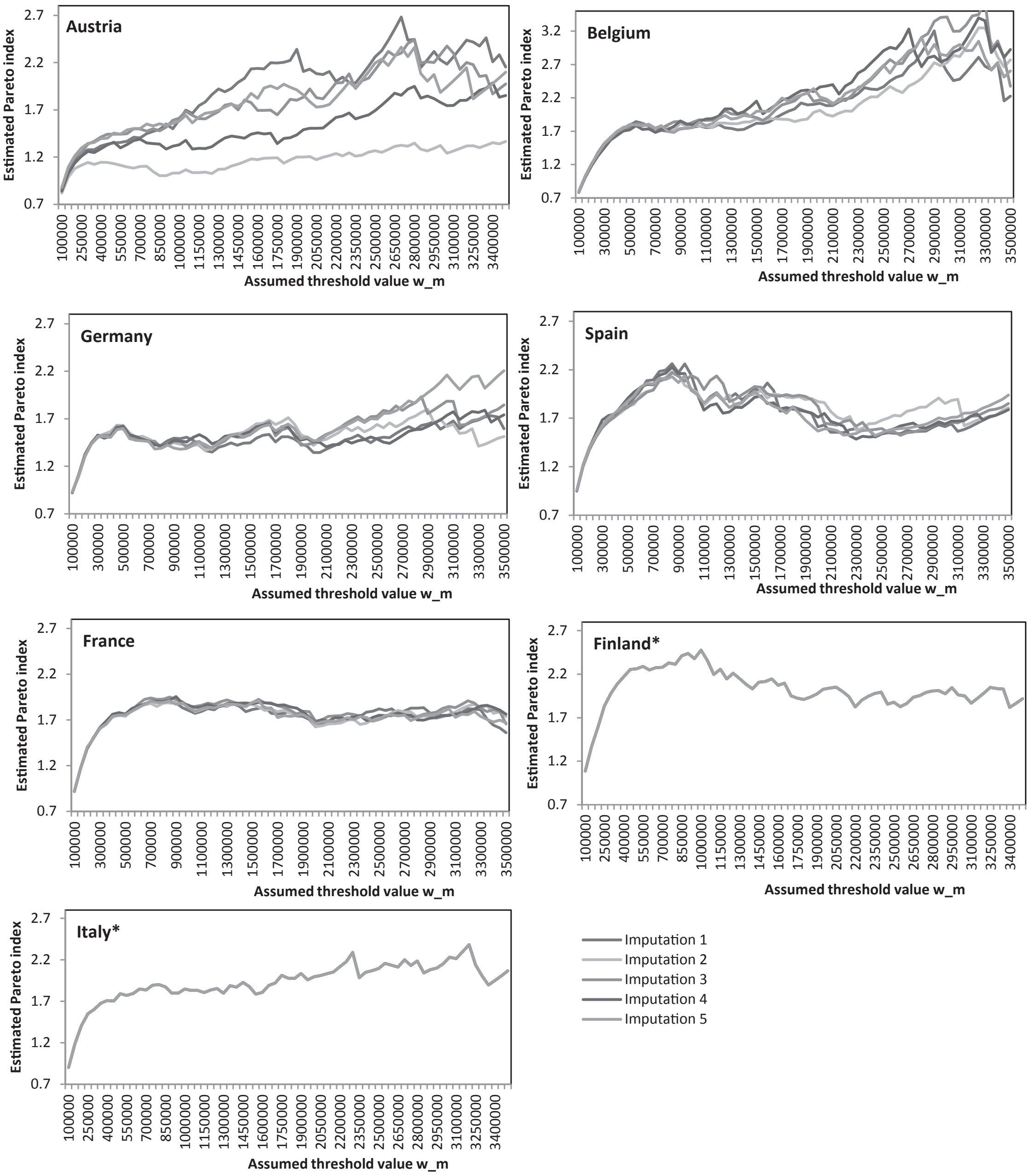

Figure C.1: ML estimation in HFCS 2013. Household net worth HFCS, Pareto index α as a function of the threshold parameter w_m. Source: HFCS 2013, own calculations. *) No multiply imputed data was provided by Finland and Italy.

D. Replication of Specifications 3 and 4 with informative weights

Since survey weights are allowed for, when calculating the Pareto parameter with a regression estimator or maximum likelihood estimator, these should, as far as possible, take into account the structure of the differential non-response. Here, for comparison's sake, it is assumed that the researcher knows about the non-response mechanism and survey weights are informed accordingly. This means the alternative weighted maximum likelihood estimator is used instead of the unweighted ML estimator. N_i is the frequency weight of a household i, $N_{w>w_m}$ is the combined frequency weights of all households exceeding the threshold value w_m:

$$a_{wml} = 1 + \frac{N_i}{N_{w>w_m}} \left[\sum_{i=1}^{n} \ln \frac{w_i}{w_m} \right]^{-1} .$$

If the survey weights are (perfectly) informed about the differential non-response the median estimates of the Pareto index are unbiased at the true value of w_m and thereafter (figure D.1, left panel). However, the margin of error is rather high as seen by the 10th and 90th percentiles. The weighted maximum likelihood estimator turns out to be unbiased, but not efficient. Overall, even if the researcher knows the exact mechanism of the differential non-response–which is usually not the case–, estimates of the Pareto index α vary strongly due to a lack of precision. This lack of precision directly translates to corrected values of aggregate private wealth (figure D.1, right panel) and top wealth shares (figure D.2), which are unbiased against the median but not very precise given the net sample size of roughly 15,000 households.

Specification 3b: Replication of Specification 3 with informed weights. Weighted maximum likelihood estimation of Pareto index α as a function of w_m. Assumption: differential non-response.

Next, it is illustrated, how informative survey weights change the results of specification 4. As in Vermeulen (2014) the incorporation of informed survey weights, a weighted regression estimator and including the top 50 entries from a rich list (assuming they are unbiased) greatly improves the precision as compared to the ML estimator in specification 3b without rich list data. Furthermore, estimation precision of the Pareto index increases (almost) independently of the chosen threshold value (figure D.3, left panel). The corrected totals using this rich list estimation are unbiased and efficient. This serves to illustrate

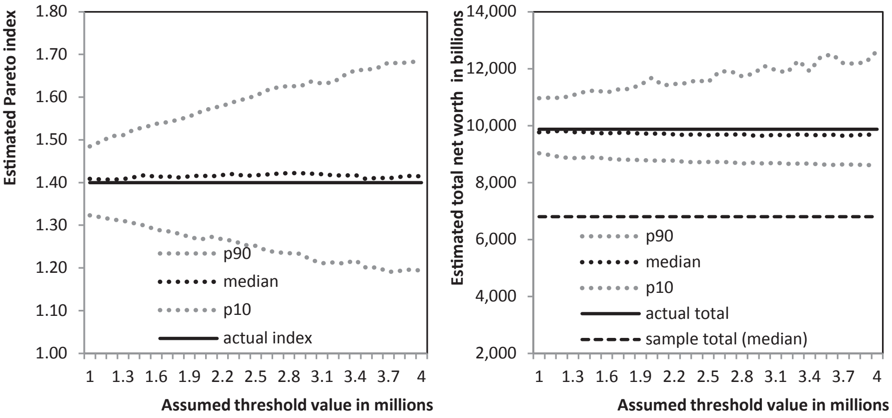

Figure D.1: Specification 3b. Informative weights, ML estimator: Impact of differential non-response on the maximum likelihood estimates for the Pareto index α and total net worth, plotted as a function of the value assumed for w_m. 1,000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

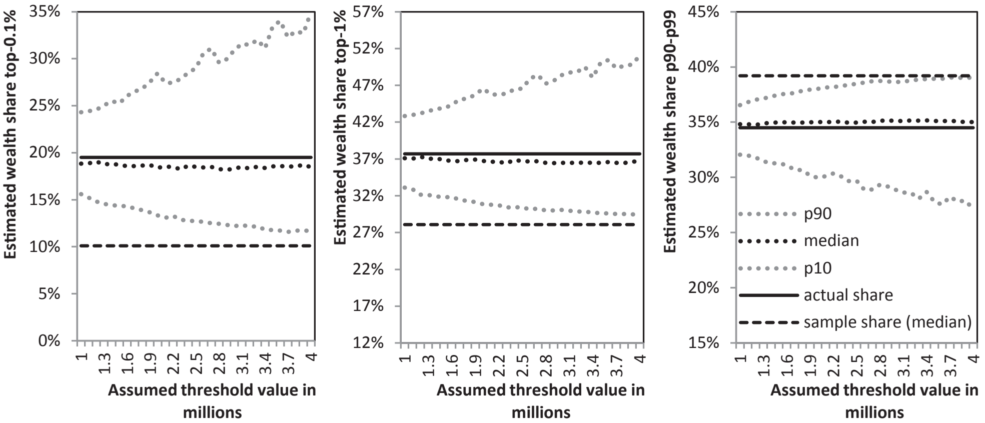

Figure D.2: Specification 3b. Informative weights, ML estimator: Impact of differential non-response on the top wealth shares before and after Pareto correction, plotted as a function of the value assumed for w_m. 1,000 samples each, drawn from test distributions, Eq. 4, see also Fig. 4.

that the real problem an empiricist faces is not the estimation of the parameters but to obtain survey data that is informed about the underlying mechanism of non-response. Only then the number of household exceeding the threshold is correct. The slight under- or overestimation of top wealth shares as depicted in figure D.4 disappears as the sample size grows larger.

Specification 4b: Replication of Specification 4 with informed weights. Weighted regression estimator (including top 50 rich list entries) of Pareto index α as a function of w_m. Assumption: differential non-response.

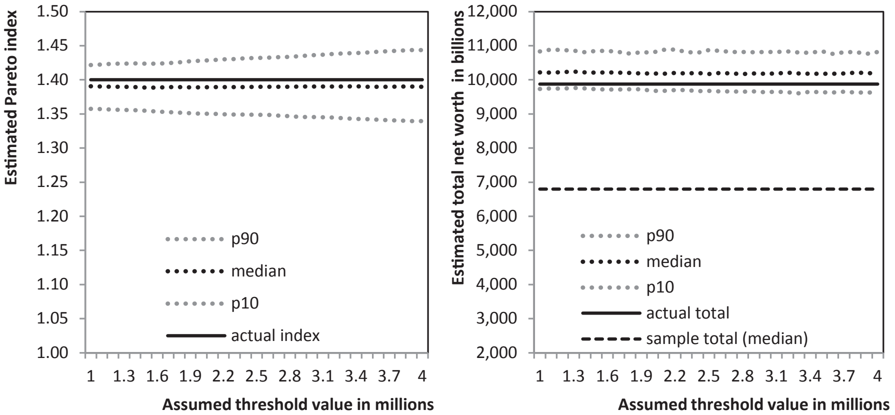

Figure D.3: Specification 4b. Informative weights, REG estimator plus rich list: Impact of differential non-response on the maximum likelihood estimates for the Pareto index α and total net worth, plotted as a function of the value assumed for w_m. 1,000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

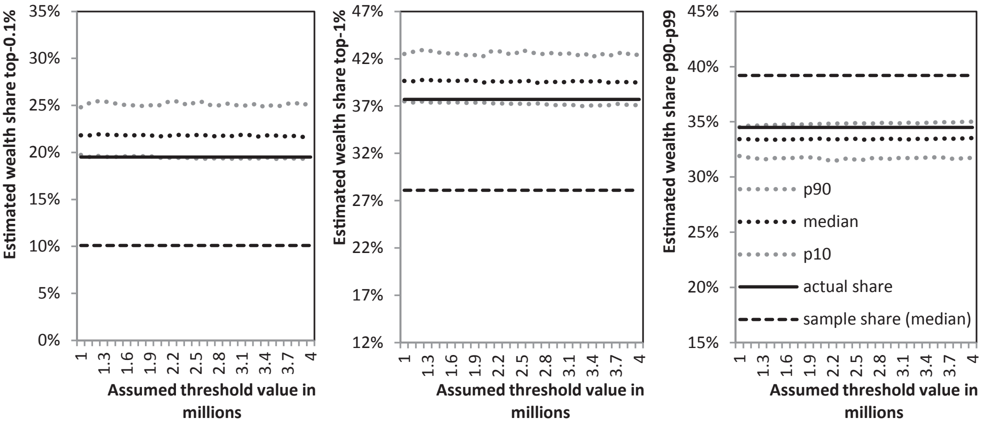

Figure D.4: Specification 4b. Informative weights, REG estimator plus rich list: Impact of differential non-response on the top wealth shares before and after Pareto correction, plotted as a function of the value assumed for w_m. 1,000 samples each, drawn from test distributions, Eq. 4, see also Fig. 4.

E. On the progressivity of non-response rates and the estimation bias

The problems when correcting for the missing rich are also closely related to the question of exactly how many households exceed the value of w_m. The number is wrong if the survey weights are uninformed. This section serves as an illustration for the observation that the overall non-response rates impact top wealth shares and aggregate wealth less than the factor, how quickly the response rates decrease depending on the households' net worth. To accomplish this, specification 4 is repeated exactly as in section 2.3, but the assumed non-response mechanism is changed. From mechanisms NR(1) to NR(4) the overall response rates are increased, as indicated by the term independent of wealth in the formulas below. However, in NR(1) the response rates decrease slower than for NR(4), as the quadratic term is smaller (see also results in Appendix A).

- **NR(1)** $\text{P(Response)} = 0.4 + 0.02 * \ln(w_i) - 0.0016 * \ln(w_i)^2$
- **NR(2)** $\text{P(Response)} = 0.5 + 0.02 * \ln(w_i) - 0.0020 * \ln(w_i)^2$
- **NR(3)** $\text{P(Response)} = 0.6 + 0.02 * \ln(w_i) - 0.0024 * \ln(w_i)^2$
- **NR(4)** $\text{P(Response)} = 0.7 + 0.02 * \ln(w_i) - 0.0028 * \ln(w_i)^2$

As in specification 4 the resulting Pareto indices and the number of households exceeding the threshold value are used to extrapolate the aggregate private wealth. Figure E.1 depicts the median relative bias of the estimates as compared to the population's target value of roughly 9.9 trillion. While the overall response rates in NR(1) are much lower than in NR(4), the underestimation of aggregate wealth is more severe in the latter case. As mentioned above, the driving mechanism is not the Pareto index itself (much less the threshold value), but the number of households an empiricist assumes to be exceeding the threshold value. This number is the lower the quicker the response rates increase.

Specifications 4c: Various non-response generating mechanisms. Weighted regression estimation of Pareto index α as a function of w_m, including rich list data. Assumption: differential non-response, underlying response mechanism unknown.

The biased aggregate wealth after correction directly translates to a bias when computing the top wealth shares (figure E.2). The top wealth shares are less overestimated under NR(1) and seem to somewhat increase with steeper non-response functions. However, since the top wealth shares are a relative measure having the aggregate wealth in the denominator, there is less variation than for the aggregate wealth for different non-response assumptions. As explained in section 2.3, in this specification are two biases, which somewhat counter

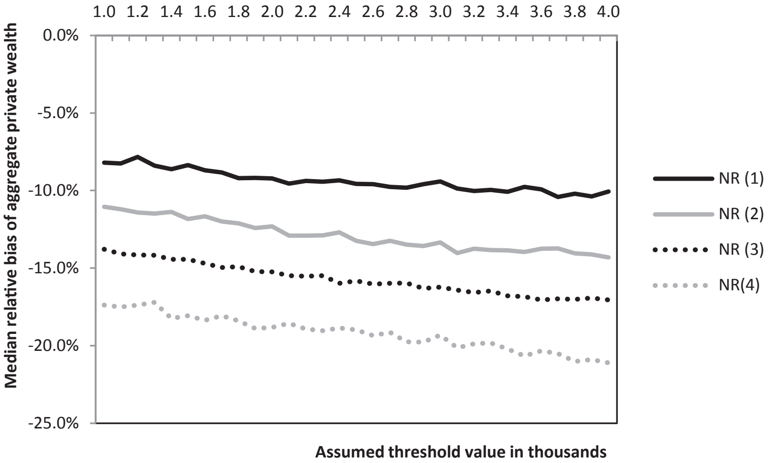

Figure E.1: Specification 4c. REG estimator plus rich list: Impact of differential non-response on the maximum likelihood estimates for the Pareto index α and total net worth, plotted as a function of the value assumed for w_m. 1,000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

each other: (1) the inequality at the top is overestimated, as the Pareto index is too low; but (2) the number of Pareto-distributed households is too low. The effect on the aggregate wealth is a downward bias; the effect on the top wealth shares is an upward bias.

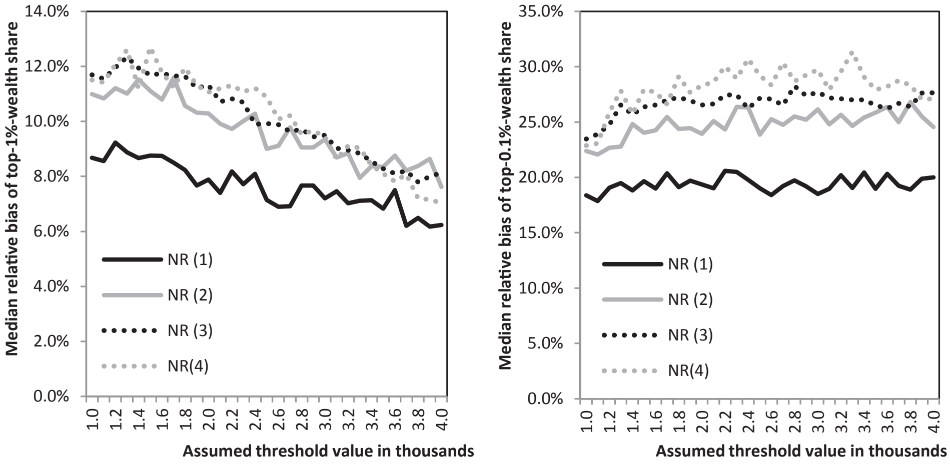

Figure E.2: Specification 4c. REG estimator plus rich list: Impact of differential non-response on the top wealth shares, plotted as a function of the value assumed for w_m. 1,000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

F. Replication of Specification 4: Are the patterns changing for varying Pareto indices or threshold parameters?

In this last section of this appendix, specification 4 is repeated once more, but this time with varying Pareto indices and threshold parameters. For an explanation of the simulation we refer to section 2.2, in figures F.1 and F.2 the parameter $w_m = 1{,}000{,}000$ is fixed and the Pareto index is varying between 1.3 and 1.6. In figures F.3 and F.4 the Parameter $\alpha = 1.4$ is fixed and the threshold value w_m varies between 0.75 and 1.75 million (see also Eq. 4 , see also Fig. 4).

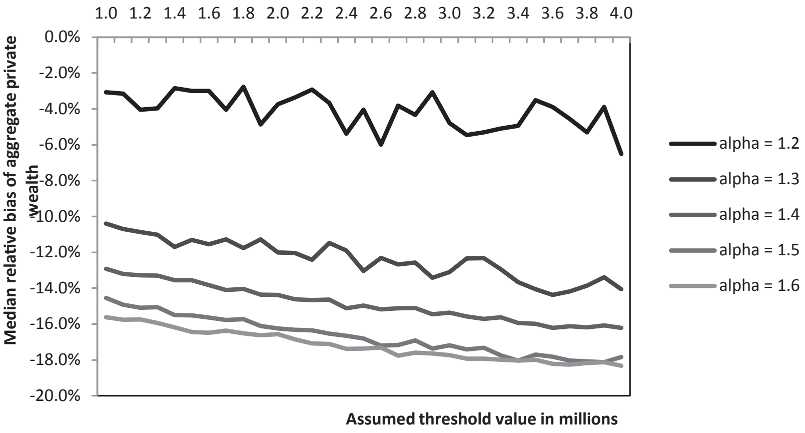

Figure F.1: Impact of Pareto index α. Impact of differential non-response on Pareto-corrected aggregate private wealth (regression including rich list data) for various values of Pareto index, plotted as a function of the value assumed for w_m. 1 000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

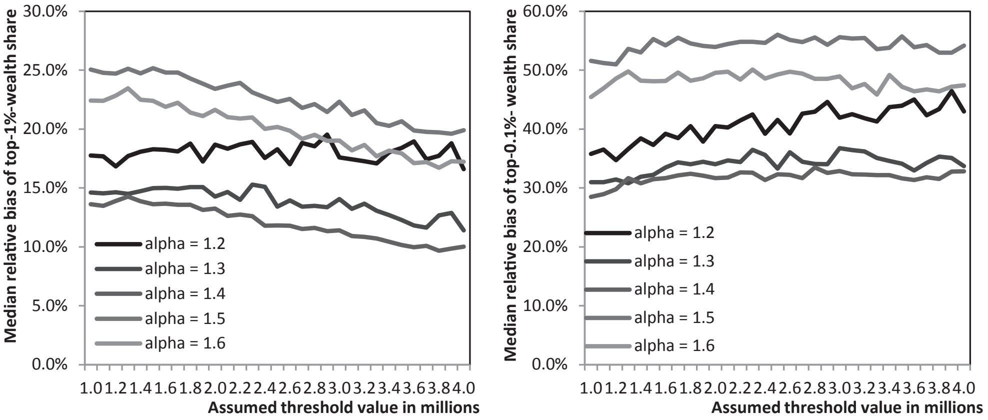

Figure F.2: Impact of Pareto index α. Impact of differential non-response on Pareto-corrected top wealth shares (regression including rich list data) for various values of Pareto index, plotted as a function of the value assumed for w_m. 1 000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

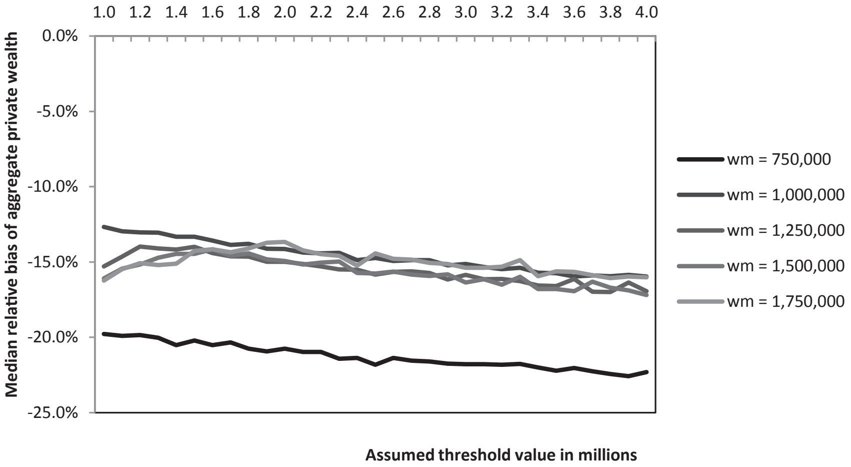

Figure F.3: Impact of (population) threshold value w_m. Impact of differential non-response on Pareto-corrected aggregate private wealth (regression including rich list data) for various values of w_m, plotted as a function of the value assumed for w_m. 1 000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

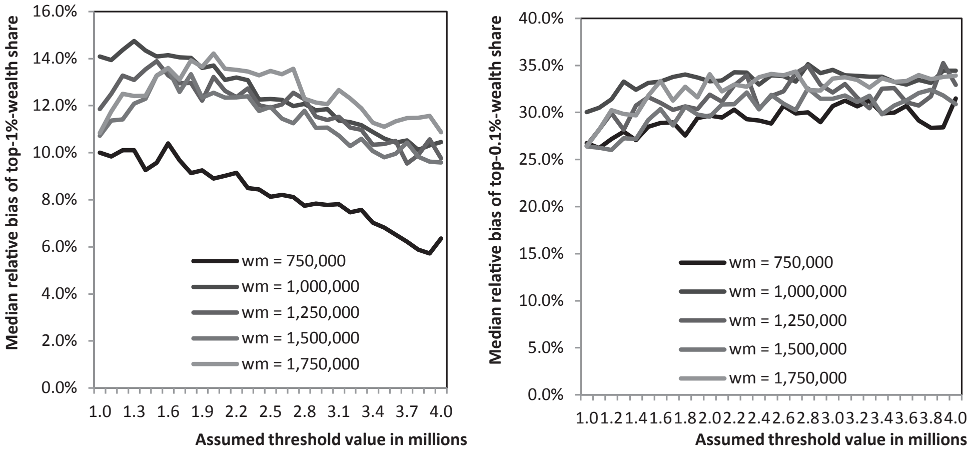

Figure F.4: Impact of (population) threshold value w_m. Impact of differential non-response on Pareto-corrected top wealth shares (regression including rich list data) for various values of w_m, plotted as a function of the value assumed for w_m. 1 000 samples each, drawn from test distributions, Eq. 4 , see also Fig. 4.

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business and Economics - Freie Universität Berlin

2016 erschienen:

2016/1 BARTELS, Charlotte und Maximilian STOCKHAUSEN
Children's opportunities in Germany – An application using multidimensional measures
Economics

2016/2 BÖNKE, Timm; Daniel KEMPTNER und Holger LÜTHEN
Effectiveness of early retirement disincentives: individual welfare, distributional and fiscal implications
Economics

2016/3 NEIDHÖFER, Guido
Intergenerational Mobility and the Rise and Fall of Inequality: Lessons from Latin America
Economics

2016/4 TIEFENSEE, Anita und Christian WESTERMEIER
Intergenerational transfers and wealth in the Euro-area: The relevance of inheritances and gifts in absolute and relative terms
Economics

2016/5 BALDERMANN, Claudia; Nicola SALVATI und Timo SCHMID
Robust small area estimation under spatial non-stationarity
Economics

2016/6 GÖRLITZ, Katja und Marcus TAMM
Information, financial aid and training participation: Evidence from a randomized field experiment
Economics

2016/7 JÄGER, Jannik und Theocharis GRIGORIADIS
Soft Budget Constraints, European Central Banking and the Financial Crisis
Economics

2016/8 SCHREIBER, Sven und Miriam BEBLO
Leisure and Housing Consumption after Retirement: New Evidence on the Life-Cycle Hypothesis
Economics

2016/9 SCHMID, Timo; Fabian BRUCKSCHEN; Nicola SALVATI und Till ZBIRANSKI
Constructing socio-demographic indicators for National Statistical Institutes using mobile phone data: estimating literacy rates in Senegal
Economics

2016/10 JESSEN, Robin; ROSTAM-AFSCHAR, Davud und Sebastian SCHMITZ
How Important is Precautionary Labor Supply?
Economics

2016/11 BIER, Solveig; Martin GERSCH, Lauri WESSEL, Robert TOLKSDORF und Nina KNOLL
Elektronische Forschungsplattformen (EFP) für Verbundprojekte: Bedarfs-, Angebots- und Erfahrungsanalyse
Wirtschaftsinformatik

2016/12 WEIDENHAMMER, Beate; Timo SCHMID, Nicola SALVATI und Nikos TZAVIDIS
A Unit-level Quantile Nested Error Regression Model for Domain Prediction with Continuous and Discrete Outcomes
Economics

2016/13 TZAVIDIS, Nikos; Li-Chun ZHANG, Angela LUNA HERNANDEZ, Timo SCHMID, Natalia ROJAS-PERILLA
From start to finish: a framework for the production of small area official statistics
Economics

2016/14 GASTEIGER, Emanuel
Do Heterogeneous Expectations Constitute a Challenge for Policy Interaction?
Economics

2016/15 HETSCHKO, Clemens; Ronnie SCHÖB und Tobias WOLF
Income Support, (Un-)Employment and Well-Being
Economics

2016/16 KÖNIG, Johannes und Carsten SCHRÖDER
Inequality-Minimization with a Given Public Budget
Economics

2016/17 ENGLER, Philipp und Juha TERVALA
Welfare Effects of TTIP in a DSGE Model
Economics

2016/18 BREYEL, Corinna und Theocharis GRIGORIADIS
Foreign Agents? Natural Resources & the Political Economy of Civil Society
Economics

2016/19 MÁRQUEZ-VELÁZQUEZ, Alejandro
Growth Impacts of the Exchange Rate and Technology
Economics

2016/20 DULLIEN, Sebastian; Barbara FRITZ, Laurissa MÜHLICH
The IMF to the Rescue: Did the Euro Area benefit from the Fund’s Experience in Crisis fighting?
Economics