Estimating Pareto's Coefficient with a Top-Wealth Database

Seminar Thesis

by Fabian Nemeczek (5064660), Sebastian Kahl (4314783),
Marcelo Avila (4679876) and Tobias Albers (5065815)

Berlin, 15.03.2019

Measuring Inequality

Winter Term 2018/19

Supervised by

Prof. Dr. Carsten Schröder

Assistant Supervisor

Maximilian Wenzel

Contents

List of Figures

List of Tables

List of Acronyms

α	Pareto's coefficient
$\widehat{\alpha}_j$	estimated Pareto's coefficient of implicate j
$\overline{\widehat{\alpha}}$	mean of estimated Pareto's coefficient over all implicates
σ^2_W	within variance
σ^2_B	between variance
σ^2_T	total variance
c	intercept of OLS estimation
cdf	cumulative distribution function
ccdf	complementary cumulative distribution function
CI	confidence interval
df	degrees of freedom
F	cumulative distribution
hh	households
imp.	implicate or imputation
i	household units
j	net wealth implicates
m	number of implicates per variable
MICE	multiple imputation by chained equations
nw	net wealth
pdf	probability density function
p95, p99	95th or 99th percentile
P	complementary cumulative distribution
PT	abbreviation for Pretest
SP	abbreviation for Socio-Economic Panel (SOEP)
T	test statistic of Welch's t-test
v_i	re-weight for household i in the Pretest dataset
w_i	weight for household i in the SOEP dataset
y	net wealth variable
$\underline{y}$	lower bound or threshold value of the Pareto distribution

1 Introduction

Since Piketty's (2014) "Capital in the Twenty-First Century", the discussion about a fair distribution of income and wealth is up to date in politics and science. A high degree of economic inequality may lead to social, political and economical disturbances. There is a risk for the social cohesion because poverty and hopelessness are the base for radical political players. A high concentration of wealth at the top of the wealth distribution may increase the political power of a small group of rich people which pushes their own interests. The economic consequences of high poverty are a reduction of purchase power and credit limitation, which may slow down economic growth. In general, high inequality in income and wealth violates the sense of justice of a meritocratic society. Especially when high fortune is inherited across generations (Piketty, 2014).

At the beginning of 2019, the briefing paper "Public good or private wealth?", published by the non-government organization Oxfam (2019), dominated the public discussion about wealth and distributive justice. According to their study, worldwide wealth is more concentrated than ever before. In 2018, 26 people owned the same as the 3.8 billion people who make up the poorest half of humanity worldwide. Even for Germany the evaluation is disastrous. The top percent of the wealth distribution owned the same as the lower 87 percent. Hence, across developed countries, Germany has a considerably high degree of inequality in wealth.

Publications like these have substantial weaknesses: The data basis is inadequate and the method of prediction is dubious. For example, Oxfam only utilize data from the *Forbes World's Billionaires List* for high-worth individuals and from the Swiss bank *Credit Suisse* for the general wealth distribution. However, both sources of data deal with rather different definitions of wealth. In general, valid and comparable data about high-worth individuals or households and their assets are nearly non-existent, including Germany (Losse, 2019).

The lack of data of high-worth individuals and households in statistics has many reasons. One reason for the non-existence of public data of very rich individuals in Germany is that it is not obligatory to register wealth. Furthermore, the German wealth tax is also suspended since 1997. Also, it is unlikely to collect data on rich individuals in limited random samples because their numbers are rather small and their willingness to participate in surveys is assumed to decrease with increasing wealth (Westermeier and Grabka, 2015).

In an effort to address the issue raised above, the German Institute for Economic Research (DIW Berlin) constructed a new survey design for identifying high-worth individuals. The study is based on the assumption that wealthy people hold a large proportion of their assets in company shares. Schröder et al. (2018) assembled a database suitable to overcome the lack of information at the top-tail of the wealth distribution. On the basis of this database a first successful Pretest was completed in 2017, resulting in 124 successfully interviewed high-worth individuals, which form one part of our data foundation for the following analysis.

In economics, it has been discussed for decades whether the top-tails of wealth follow the Pareto distribution (Brzezinski, 2014). Simply put, the principle of the Pareto distribution implies that, starting from a certain threshold value or lower bound, the logarithmized wealth follows a linear relationship to the logarithmized complementary cumulative density function. The slope of this linear relationship is the Pareto's coefficient α, which can also be used as an instrument to measure inequality. The smaller the coefficient, the more unequal wealth is distributed among the population owning more wealth than the threshold value. Our paper explores the question of whether the top-tail of the wealth distribution in Germany follows a Pareto distribution and, if so, which Pareto's coefficients we can observe.

Related Literature

To approach the lack of data at the top-tail of the wealth distribution, many researchers use information received from so-called "rich lists" published by magazines to add them to data from population studies. Mostly, the applicable data on rich individuals and households is derived from public information, evaluations of financial markets and business media or interviews (Vermeulen, 2018). Utilizing these data sources, many researchers compute the Pareto coefficient in order to impute the missing top-tails. Westermeier and Grabka (2015) proceeded similarly when they analyzed high-worth individuals and households in Germany. They simulated the wealth shares of rich households by using about 55 rich individuals from the Forbes World's Billionaires List and the upper tail of the wealth distribution of the SOEP. Based on data from the SOEP in 2012 and by setting lower bounds between 900,000 and 1,350,000 euros for the wealth on household level, they estimate a Pareto distribution with values of the Pareto's coefficients between 1.35 and 1.38. Consecutively, they estimate that the top percent of the German wealth distribution holds 31 to 34 percent of the overall wealth in Germany, whereas the top 0.1 percent holds 14 to 16 percent. By excluding the Forbes data, they estimate that the top percent possesses 18 percent, while the top 0.1 percent holds 5 percent of total wealth in Germany.

Bach et al. (2015) set the focus on the comparison of different sources of information on rich households and the imputation of the top wealth distribution for different Western European countries. Furthermore, they deploy Gini coefficients to compare the inequality of the wealth distribution with and without imputing the top-tails. The concentration of wealth is higher when taking also the national rich lists into account compared to using only the data provided by the Forbes magazine. Using data from 2011, the authors utilize around 200 entries from the German Manager Magazine and combine them with the Household Finance and Consumption Survey (HFCS) data provided by the European Central Bank. By setting a lower bound of 500,000 euros for the wealth on household level, they estimate a Pareto distribution with a value of the Pareto's coefficient of 1.37. Due to the imputation of the top-tail of the wealth distribution, the share of the total wealth of Germany held by the most rich percent of the households in Germany increases from 24 to 33 percents. Likewise, the share of the top 0.1 percent increases from 4 to 17 percents. Accordingly, the Gini coefficient of the overall distribution also rises from 0.75 to 0.78.

Dalitz (2018) presents different methods for the estimation of the Pareto's coefficient. The author recommends the maximum likelihood estimation as the most efficient method for estimating a Pareto Distribution. Dalitz (2018) utilizes the same data sources as Bach et al. (2015), but uses data from the year 2012 instead. The author takes into account the multiply imputed values of the net wealth variable from the HFCS. By using a maximum likelihood estimation and by considering the different implicates and the different lower bounds between 558,000 and 619,800 euros for the wealth on household level, the author estimates a range for the Pareto's coefficient between 1.44 and 1.49. By imputing the top-tail on the basis of the estimation, he calculates that the top percent of the wealth distribution holds between 28 and 31 or between 32 and 34 percent of the overall wealth in Germany, depending on which model parameter was used.

Although the reviewed studies use different strategies to estimate the Pareto's coefficients on the basis of different data foundations, the estimated Pareto's coefficients are rather similar. Also, the calculations of the shares of wealth held by the most rich households are alike. Nevertheless, all authors stress the high degree of uncertainty that arises from estimations based on rich lists. Bach et al. (2015) assume that estimations of the top-tail of the wealth distribution based on external data, like rich lists, underestimate the real wealth and are inconsistent with corresponding values from survey data. The process of data collection of rich lists is assumed to be rather invalid and might result in wealth variables which are inconsistent with the design of common survey studies.

In this paper, we engage in the exploration of the top tail of the wealth distribution in Germany. For this, we use data resulting from the Pretest applied by Schröder et al. (2018). Following Westermeier and Grabka (2015), we utilize net wealth data on household level from the German Socio-Economic Panel (SOEP) in 2012. On this data foundation we estimate Pareto's coefficients for different lower bounds. Analogously to Dalitz (2018), we proceed like this with each implicate separately and deal with the multiply imputed variables according to Rubin (1987). Similar to Bach et al. (2015), we then impute the top-tails of the German wealth distribution. A considerable advantage of our analysis is that the data of the Pretest was collected by the same questionnaires as in the SOEP. In contrast to rich lists, the wealth information from the Pretest is therefore consistent and comparable with the underlying household sample. Due to this we aim to receive more robust results when estimating the Pareto distribution.

The paper is organized as follows: In chapter 2 we describe the data and variables that our analysis is based on. Subsequent to a short review about the SOEP and about the lack of data at the very top of the wealth distribution in Germany follows a summary of how the top-wealth database was assembled in the Pretest by Schröder et al. (2018). Moreover, we illustrate the probability sampling design used to identify high-worth individuals in the Pretest and compare the SOEP and the Pretest on the basis of descriptive statistics. We begin chapter 3 by describing the theoretical background of the Pareto distribution. Following this, we present the methodology we use to estimate and evaluate the Pareto's coefficients. Also, we describe how we impute the net wealth of the top-tail of the wealth distribution in the SOEP on the basis of the preceding estimation and evaluation strategy. In chapter 4 we present and discuss our main results. These include the estimated Pareto's coefficients, the applied imputation of the wealth distribution and the sensitivity analysis. In chapter 5 we conclude the paper by summarizing our findings and discussing the relevance and shortcomings of those. Lastly, we present a brief outlook as well as recommendations for further research.

2 Data & Variables

In this chapter we present the data and variables used in our analysis. In section 2.1 we describe the SOEP data from the wealth module 2012. In section 2.2 we present the method of assembling a top-wealth database which was created to identify high-worth individuals. In section 2.3 we illustrate the sample probability design applied by Schröder et al. (2018) and describe the process from the sample design to the results of the Pretest. Finally, in section 2.4, we compare the SOEP and the Pretest on the basis of descriptive statistics.

2.1 Socio-Economic Panel (SOEP)

In order to prove whether the top-tail of the wealth distribution of Germany follows a Pareto distribution and to estimate the Pareto's coefficient, we revert to the wealth data from two data sources: The SOEP and the results from the Pretest. The SOEP is a continuous household panel which collects socio-economic data from a representative sample of persons living in private households in Germany. It is conducted by the DIW Berlin. The SOEP collects wealth information on individual and household levels every five years. Positive (home market value, other property, financial assets, building-loan contract, private insurances, business assets and tangible assets) and negative asset values (debts home market value, debts other property and consumer credits) are surveyed. We use the "HWEALTH" dataset for our analysis, which includes wealth information on household level. We use the actual wealth sample from 2012.

The focus of our analysis is based on the household net wealth, meaning the accumulated asset values minus debts and other liabilities per household. So it is not surprising to observe negative wealth values when the actual debts of a household are higher than the held assets. Missing or implausible values are already imputed and edited, respectively. Therefore, per unit exist five implicates that describe the net wealth of each household. Furthermore, we apply the cross-sectional weights from the SOEP to extrapolate the observations to the whole household population in Germany.

2.2 Assembling a Top-Wealth Database

As mentioned above, high-worth individuals are underrepresented in population surveys. To receive a better understanding about the very top of the wealth distribution of Germany, Schröder et al. (2018) developed a new survey design to close the data gap by accurately identifying and interviewing high-worth individuals. Thus, an appropriate database that contains information about the wealth of high-worth individuals in Germany had to be created. For this, the database Orbis (Bureau van Dijk) which has world-wide coverage of information from around 300m public and private companies was utilized by Schröder et al. (2018). The database Orbis grants access to a wide range of information. This includes reports on corporate ownership structures and data about individuals associated with companies, which are both of particular importance for the study. The authors transformed the accessible information into a new database composed of individual information about company shares, as well as names, private and work addresses and contact information.

By accumulating the company shares of shareholders the new database contains the approximated accumulated wealth as well as the names and contact information of 1.5m individuals living in Germany. Based on the assumption that wealthy individuals own company shares, it is concluded that the top one percent of the German wealth distribution is represented to a great extent in this database. Since the adult population in Germany consists roughly of 66m individuals, only the wealthiest 660,000 individuals were considered in the Pretest. Those compose the database of the top one percent of the wealth distribution in Germany (Schröder et al., 2018, pp. 9-12; 21).

2.3 Sample Probability Design of the Pretest

In order to achieve the goal of identifying particularly rich individuals in the analysis, the researchers applied a probability sampling design called *stratified sampling*. This method empowers the researcher to highlight certain subgroups when particular emphasis is put on the analysis of specific subgroups or when these groups are nearly inaccessible.

To apply a stratified sampling method, initially, the examined population must be divided into subgroups, called strata, on the basis of specific characteristics. Subsequently, random and independent samples have to be drawn from each stratum. The samples can be drawn proportionally to the number of observations of the respective stratum to the total population (*proportionate stratified sampling*). Alternatively, the samples may also be drawn on the basis of an appropriate weighting scheme (*disproportionate stratified sampling*). The next step is to merge the drawn samples from all strata into a random sample, before lastly, in the case of disproportionate stratified sampling, new weights must be allocated to the random sub samples to compensate for the disproportionate drawing probabilities (Cochran, 1977).

As discussed above, Schröder et al. (2018) based their analysis on the assumption that the willingness to participate in survey studies decreases with increasing wealth. Additionally, the goal of the new sampling design is to target particularly the most rich individuals to increase the knowledge about the very top of the wealth distribution in Germany. To oversample those, a disproportionate stratified sampling method was applied. The database of the top one percent of the wealth distribution in Germany was therefore sorted in terms of their approximated wealth and divided into three equally sized strata of 220,000 individuals.

As common in face-to-face population surveys, before drawing the samples and conducting the interviews, a two-step sample design was applied. This was done by selecting sample points which reflect heterogeneous regional structures based on response behavior and socio-economic composition, while minimizing the cost of interviewing. From the approximately 14,000 remaining observations, random and independent samples were drawn. Stratum 3 in which the richest individuals of the sample are suspected are assigned the highest drawing probability of 4/7. Following this concept, an observation is drawn from stratum 2 or stratum 1 with a probability of 2/7 or 1/7, respectively.

Ultimately, 2000 observations were drawn and after adjusting the gross sample for unavailable addresses and individuals which could not be contacted, 124 interviews were realised from the successfully contacted individuals, as shown in Table 2.1 (Schröder et al., 2018, pp. 21-24).

<table>
<tr><td colspan="4">Sample: 2000 Individuals
(Stratum 1: 296 Ind.; Stratum 2: 588 Ind.; Stratum 3: 1119 Ind.)</td></tr>
<tr><td colspan="3">Adjusted Gross Sample:
1652 ind.</td><td>Address Unavailable:
348 ind.</td></tr>
<tr><td colspan="2">Contacted:
532 ind.</td><td>Not Contacted:
1120 ind.</td><td></td></tr>
<tr><td>Final Response:
124 ind.</td><td>Non-Response:
408 ind.</td><td></td><td></td></tr>
</table>

Table 2.1: Case Numbers - from the Drawn Sample to Realised Interviews

Due to the sampling design, an appropriate re-weighting scheme has to be applied. For this, each stratum has to be re-weighted by assigning weights determined by the relative shares of each stratum's amount of observations and dividing them by the total number of strata.

2.4 Comparison of SOEP and Pretest

The data of the Pretest was collected in 2017. The sample includes a net wealth variable with five implicates per unit and a ordinal variable indicating which stratum the respective household belongs to. The Pretest includes wealth information on household level. The advantage is that the data was polled with the same questioning instruments as the SOEP. Therefore, the net wealth variables from the SOEP and the Pretest are comparable and consistent for the following analysis.[1] Exceptions are the values of vehicle and educational loans, that are only included in the Pretest - but these are quantitatively negligible.

It should be noted that both datasets use different methods of multiple imputation of missing data. The multiple imputation by chained equations (MICE) procedure is used in the Pretest and the SOEP applies a combination of MICE and row-and-column imputation technique. For further information about imputed data in the SOEP wealth module see Grabka and Westermeier (2015). Not having corresponding cross-sectional weights to our disposal, we can solely rely on the re-weighting of the Pretest as explained above and extrapolate up to 413.000 households - the top one percent of the population in Germany 2017 (Statista, 2019).

	count	mean	p50	p75	p90	p99	min	max
SOEP nw imp. 1	14,982	0.17	0.05	0.20	0.41	1.57	-3.81	50.78
Pretest nw imp. 1	124	11.95	2.10	5.98	22.42	202.71	-4.43	207.02

Source: SOEP 2012 (v33.1), Pretest 2017.

Note: Net Wealth (nw) implicate 1, in million euros, unweighted, for simplicity rounded.

Table 2.2: Descriptive Statistics of SOEP and Pretest

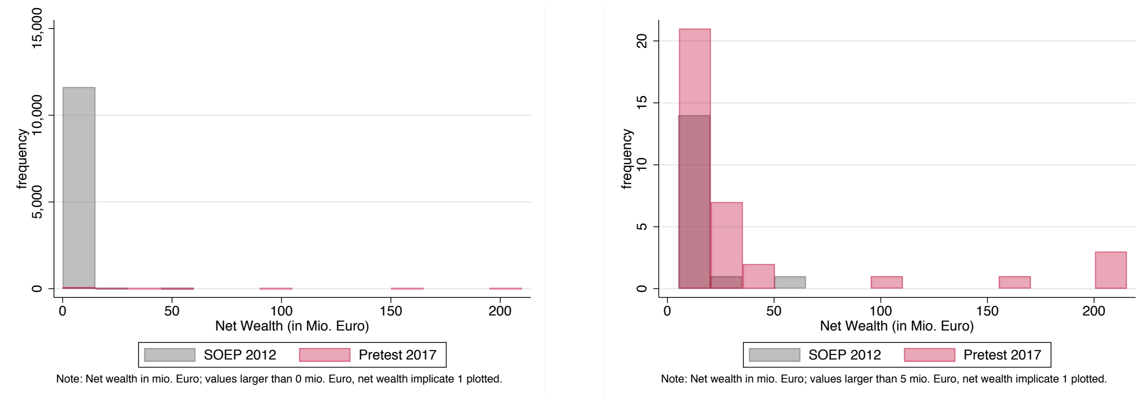
For the estimation of Pareto's coefficient it is necessary that the net wealth observed from Pretest households covers the top-tail of the wealth distribution in Germany. Thus, it is desirable to observe households which are wealthier than the richest in the SOEP. Table 2.2 shows the

[1]Note: It would be unreliable just to merge both datasets and start the analysis with the combined data as a "augmented SOEP". On the one hand, there is a period of time of five years between both surveys. On the other hand, the applied SOEP weights already take into account the whole population of Germany. A joint analysis by including Pretest data would overestimate the top wealth.

descriptive statistics of both datasets. For simplicity, we just compare the first implicates of each dataset. The SOEP contains information of about 15,000 households, the Pretest consists of 124 observations. The mean of the SOEP is about 170,000 euros, whereas in the Pretest the mean is about 12m euros. The median in the SOEP lies at 50,000 euros, while the median of the Pretest is around 2.1m euros. The richest observed household in the SOEP has around 50m euros and the richest household in the Pretest has more than 200m euros.

From the histograms depicted in figure 2.1 we see the density of the net wealth variables for both datasets. Because our analysis is focused on high-worth households and for graphical simplification, we only illustrate positive net wealth values. Figure 2.1(a) demonstrates that the wealth distribution of the SOEP data (grey) has a high concentration under values of 10m euros. This means that the vast majority of the SOEP respondents own less than 10m euros, whereas the net wealth from the Pretest (red) is distributed more uniformly and even across higher values.

Figure 2.1(b) displays only observations with a net wealth higher than 5m euros to disclose more details about the top-tail of the distribution. Only few units from the SOEP are observed and concentrated between 5m and 30m euros, while various households from the Pretest are located in the area of three-digit millions of euros of net wealth. Thus, the observations of the Pretest lie in the upper-tail of the wealth distribution in Germany which is insufficiently covered in the SOEP.



(a) Net Wealth Implicate 1 ($nw > 0$ euros) (b) Net Wealth Implicate 1 ($nw > 5m$ euros)

Figure 2.1: Histograms of SOEP and Pretest

3 Methodology

In this chapter we present the methodology used in our analysis. In section 3.1 we describe the theoretical properties of the Pareto Distribution. In section 3.2 we present the estimation procedure to obtain the Pareto coefficients. In section 3.3 we focus on the evaluation strategy and describe, first in section 3.3.1, how we deal with multiply imputed variables and then, in section 3.3.2, how we test the hypothesis of equality between the estimated Pareto's coefficients. Finally, in section 3.4, we describe the estimation procedure we apply to impute the top-tail of the distribution.

3.1 Theoretical Background on the Pareto Distribution

In his pursuit to approximate political economy to the natural sciences, the Italian economist Vilfredo Pareto rigorously studied, among other subjects, the wealth distribution across several countries and long periods of time. More precisely, Pareto finds that the proportion of households owning at least an certain amount of wealth decreases linearly with a relative increase in wealth, as described in the second volume of Cours d'Économie Politique (1897, pp. 304–309). Further, Pareto demonstrates that this *linearity in proportion* is strikingly stable across countries and across time. This concept is the foundation of the Pareto Distribution which we describe in this section.

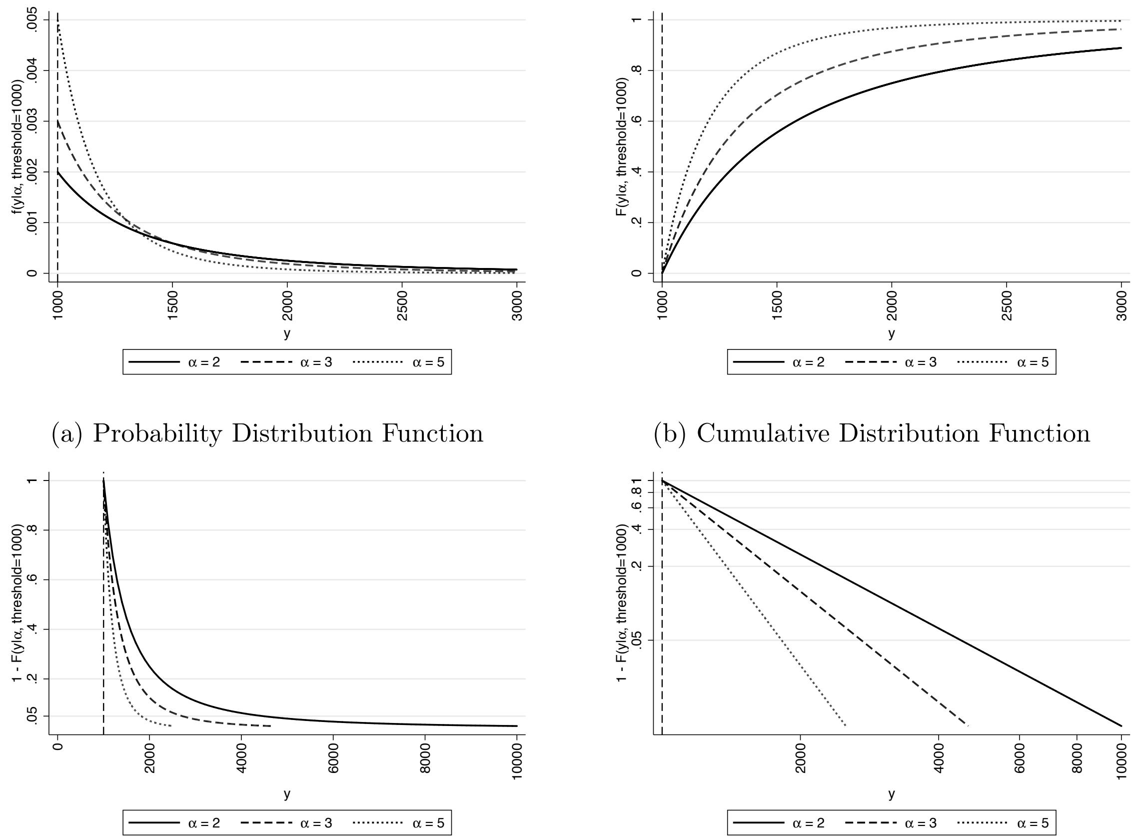
A random variable Y following a Pareto (type I) distribution has its probability density function (pdf) defined as

$$f(y) = \frac{\alpha \underline{y}^{\alpha}}{y^{\alpha+1}}, \quad \text{where } 0 < \underline{y} \leq y \text{ and } \alpha > 0, \tag{3.1}$$

and the corresponding cumulative distribution function (cdf) can be described as

$$F(y) = Pr(Y \leq y) = 1 - \left(\frac{\underline{y}}{y}\right)^{\alpha} \tag{3.2}$$

where y is the wealth measure of interest, $\underline{y}$ the threshold value, which defines the lower bound of the distribution. Finally, α, called Pareto's coefficient, determines the basic shape of the distribution. The lower the value of α, the fatter the tail of the distribution, that is, the more unequal is wealth distributed. Note that the above functions are only defined for values that are positive and bigger than the lower bound. In Figures 3.1(a) and (b), the pdf and cdf are depicted for three different values of α and a threshold of 1,000 for illustration purposes.



(a) Probability Distribution Function

(b) Cumulative Distribution Function

(c) Complementary cdf (linear)

(d) Complementary cdf (log-log, Zipf plot)

Figure 3.1: Functional Forms of the Pareto Distribution ($\underline{y}$=1000, α=$\{2,3,5\}$)

In our analysis we are more interested in the complementary cumulative distribution function (ccdf), defined as $P \equiv 1 - F(y)$, which can be interpreted as the proportion of households owning wealth valued at more than y. The ccdf is defined as

$$P = Pr(Y > y) = \left(\frac{\underline{y}}{y}\right)^{\alpha} \tag{3.3}$$

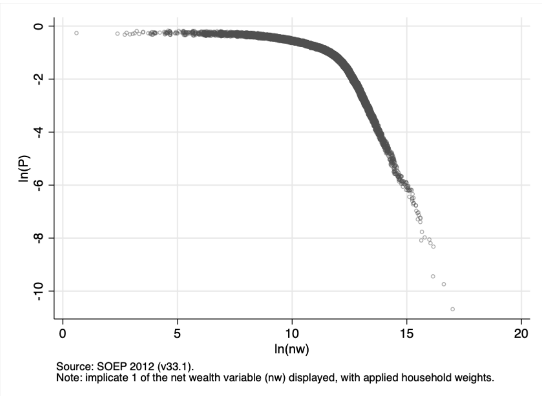
Further, taking the logarithm of equation 3.3 results in

$$\ln(P) = c - \alpha \ln(y), \text{ with } c = \alpha \ln(\underline{y}) \tag{3.4}$$

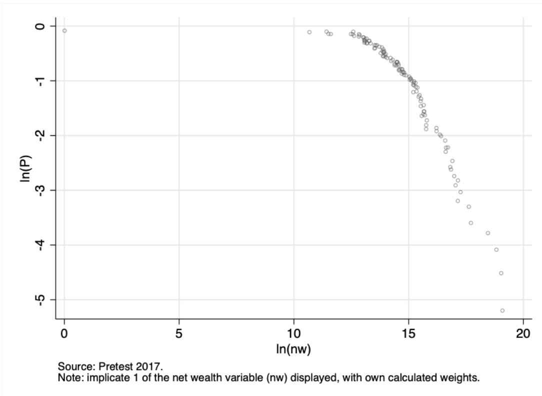
Here, the linear relationship between the ccdf and y, both in logarithm, as well as the role of α become evident. This property of the Pareto distribution is highlighted when plotted as a *Zipf* *plot*, that is, a double logarithmic scaled plot, as illustrated in Figure 3.1(d). Equation 3.4 is the base of our estimation strategy, which is described in the following section.

3.2 Estimating Pareto's Coefficients

In this section, we describe the procedure taken to estimate Pareto's coefficient α, following a similar approach to the regression method of estimating a power law as described by Vermeulen (2018, pp. 16–18).



(a) Net Wealth Implicate 1 of the SOEP



(b) Net Wealth Implicate 1 of the Pretest

Figure 3.2: Zipf Plots of SOEP and Pretest

Note that our analysis is based on multiply imputed datasets. The same procedure described below is, therefore, repeated across all implicates. How we deal with the multiply imputed dataset is described in section 3.3.1. Similarly, this also applies to the fact that we conduct the estimations on both the Pretest and the SOEP datasets separately. Further, we also check the stability of α by conducting the same estimations based on two threshold values. In total, the same procedure is repeated 20 (5 x 2 x 2) times. For simplicity, we subsequently maintain focus on only one of such a procedure.

First, we generate an empirical cdf by constructing a rank ($n(y_i)$) of the households ordered by wealth levels divided by number of households (N). To obtain an empirical ccdf, we subtract the generated empirical cdf from 1, such that

$$\left(1 - \frac{n(y_i)}{N}\right) \cong P_i = \left(\frac{\underline{y}}{y_i}\right)^{\alpha} \text{ where } y_i \geq \underline{y} \qquad (3.5)$$

Both N as well as the ranking $n(y_i)$ are weighted by the frequency weights in order to obtain a representative estimation of the whole population (see Vermeulen (2018, pp. 17–18)).

Then, taking the logarithm of Eq. 3.5, we obtain our regression equation, which we estimate by ordinary least squares (OLS) and can be depicted as

$$\ln P_i = c - \alpha \cdot \ln y_i + \varepsilon_i \text{ where } y_i \geq \underline{y} \; \forall \; i \tag{3.6}$$

with $c = \alpha \ln(\underline{y})$. Note that the resulting slope of the regression is expected to be negative, but since $\alpha > 0$, we always refer to this coefficient in absolute terms. Further, observe that the Pareto distribution is only defined for those observations with higher net wealth than the threshold value. Based on Eq. 3.6, a graphical analysis as depicted in Fig. 3.2 is the first step to assess the basic shape of the distribution and to examine the range of threshold values after which the decrease in $\ln P_i$ should be approximately linear.

The choice of threshold is of great importance for estimation procedure, since the estimated coefficients may present strong variability, depending on the chosen value. This variability, however, indicates that the Pareto distribution might not be the best approximation to the empirical data distribution. In order to maintain consistency between the estimation procedures, we seek to preserve the same threshold levels on the estimation based on the SOEP and on the Pretest data. For that, the data should in both cases cover the threshold values with a significant number of observations. Further, we should set the thresholds on a range where the linearity in proportions is already apparent, and this can be assessed visually by looking at Fig. 4.4 where we estimate and plot the Pareto coefficient for different net wealth values. Following Jenkins (2017), we found that the thresholds set at the respective 95th (ranging from 563.570 to 573.000 euros) and 99th (1.31m to 1.37m) percentiles of the SOEP data account for these requirements. Note that the thresholds vary slightly across implicates and all observations with net wealth under the threshold are disregarded for the regressions.

3.3 Evaluation Strategy

3.3.1 Dealing with Imputed Net Wealth Variables

As already discussed, the net wealth variable of both datasets, SOEP and Pretest, is multiply imputed. However, the respective multiple imputation methods which are applied in the SOEP and the Pretest do not correspond with each other. Thus, an evaluation according to Rubin (1987) allows us to appropriately interpret the Pareto's coefficients in the presence of multiply imputed data. In a series of steps we follow Rubin (1987) and calculate the *total variance* to control for the increased uncertainty of the true parameters due to the multiple imputation

(Rubin, 1987, pp. 15-22). The corrected variance formula allows us to perform tests and estimate confidence intervals for the true parameter of the Pareto's coefficient α.

Let i denote the rows of each dataset, i.e. the households, and let j denote the columns, i.e the net wealth implicate of the household. As explained above, for each dataset m Pareto's coefficients are computed per threshold value and thus, for each column j we compute one estimate $\widehat{\alpha}_j$. Since it is desirable to receive only one Pareto's coefficient $\widehat{\alpha}$ per dataset and treshold value, the arithmetic mean of the m estimated coefficients $\widehat{\alpha}_j$ is used as an estimate of the parameter at each threshold value for the SOEP and Pretest data, respectively:

$$\overline{\widehat{\alpha}} = \frac{1}{m}\sum_{j=1}^{m} \widehat{\alpha}_j, \quad m = \{1, ..., 5\}. \tag{3.7}$$

For the remaining part of this section, we refer to the arithmetic mean as conducted in Eq. 3.7) as the *estimate* of the Pareto's coefficient $\overline{\widehat{\alpha}}$. In the next series of steps we compute the total variance according to Rubin (1987) for each estimate of Pareto's coefficient $\overline{\widehat{\alpha}}$. For the adjusted variance formula which reflects the increased uncertainty due to the multiply imputed net wealth values, the average within-variance and the between-variance of the estimate of α have to be calculated. In accordance with the notation of the previous sections, let $\ln y_{i,j}$ denote column j's value of the logarithmic approximated cumulative net wealth variable of household i. Also, let $\ln P_{i,j}$ denote the logarithmic value of the cumulative population share in column j that possesses not more wealth than household i in column j.

To increase the readability of the variance formula of each coefficient j in Eq. 3.10 and Eq. 3.13, we set up auxiliary functions. An appropriate mean of the logarithmic net wealth variable must take the frequency weights which reflect the estimated frequency of comparable households in the underlying population into consideration. Applying the frequency weights of the SOEP (in index: SP), denoted as w_i, the weighted mean of each net wealth implicate is then given by

$$\overline{\ln y}_{SP,j} = \frac{\sum_{i=1}^{n} w_i * \ln y_{i,j}}{\sum_{i=1}^{n} w_i}, \tag{3.8}$$

while the sum of the squared error terms of the estimation of each coefficient $\widehat{\alpha}_j$ is composed as follows:

$$\sigma^2_{SP}(\varepsilon_{i,j}) = \sum_{i=1}^{n} (\ln P_{i,j} - (c_j + \alpha_j * w_i * \ln y_{i,j})). \tag{3.9}$$

The resulting variance of each parameter $\widehat{\alpha}_j$ is then given by:

$$\sigma^2\left(\widehat{\alpha}_{SP,j}\right) = \frac{\sigma^2_{SP}\left(\varepsilon_{i,j}\right)}{\sum_{i=1}^{n}\left(w_i * \ln y_{i,j} - \overline{\ln y}_{SP,j}\right)^2}. \tag{3.10}$$

This specification of the variance of alpha is consistent with the simple linear regression model set up in section 3.2.

Since a corresponding weighting scheme for the Pretest data is not available yet, Eq. 3.8 - 3.10 have to be adjusted for the Pretest (in index: PT). As mentioned above, the Pretest observations were drawn according to a disproportionate stratified sampling method. Therefore, the Pretest data requires to be re-weighted to compensate for the different high probabilities that a drawn household is from one of the respective strata. Thus, we applied a re-weighting scheme in which each household's weight is composed as the inverse of the relative share of interviewed households i that are obtained in the stratum divided by the total number of strata. Therefore, let v_i denote the re-weight for each household depending on which stratum in the Pretest dataset it belongs to. Accordingly, the alternate formulas for the mean of the net wealth variable and the sum of squared error terms of each estimation are computed as

$$\overline{\ln y}_{PT,j} = \frac{\sum_{i=1}^{n}(v_i * \ln y_{i,j})}{\sum_{i=1}^{n} v_i} \tag{3.11}$$

and

$$\sigma^2_{PT}(\varepsilon_{i,j}) = \sum_{i=1}^{n}(\ln P_{i,j} - (c_j + \alpha_j * v_i * \ln y_{i,j})) \tag{3.12}$$

respectively.

Subsequently, the estimator for the variance of the Pareto's coefficient for column j in the Pretest dataset, composed of 3.11 and 3.12, results as following:

$$\sigma^2\left(\widehat{\alpha}_{PT,j}\right) = \frac{\sigma^2\left(\varepsilon_{i,j}\right)}{\sum_{i=1}^{n}\left((v_s * \ln y_{i,j}) - \overline{\ln y}_{PT,j}\right)^2} \tag{3.13}$$

Making use of the respective variance formulas as specified in Eq. 3.10 and Eq. 3.13, for both, SOEP and Pretest, the within-variance of an coefficient $\overline{\widehat{\alpha}}_j$ is calculated as

$$\sigma^2_W(\widehat{\alpha}_j) = \frac{\sigma^2(\widehat{\alpha}_j)}{n}, \tag{3.14}$$

whereupon

$$\sigma_W^2(\overline{\widehat{\alpha}}) = \frac{1}{m}\sum_{j=1}^{m=5}\sigma_W^2(\widehat{\alpha}_j) \tag{3.15}$$

represents the average within-variance, i.e. the arithmetic mean of the m within-variances of the Pareto's coefficients $\widehat{\alpha}_j$ of each dataset and at each threshold value.

The corresponding between-variance, i.e the sum of the squared deviations of each m Pareto's coefficients $\widehat{\alpha}_j$ from their respective arithmetic mean is then computed as follows:

$$\sigma_B^2(\overline{\widehat{\alpha}}) = \sum_{j=1}^{m=5}(\widehat{\alpha}_j - \overline{\widehat{\alpha}})^2. \tag{3.16}$$

Eventually, from the average within-variance 3.15 and the between-variance 3.16 the total variance of the estimate of alpha $\overline{\widehat{\alpha}}_j$ is given as

$$\sigma_T^2 = \underbrace{\frac{1}{m}\sum_{j=1}^{m=5}\frac{\sigma^2(\widehat{\alpha}_j)}{n} + (1-m^{-1}) * \sum_{j=1}^{m=5}(\widehat{\alpha}_j - \overline{\widehat{\alpha}})^2}_{\sigma_W^2(\overline{\widehat{\alpha}})+(1-m^{-1})*\sigma_B^2(\overline{\widehat{\alpha}})}, \tag{3.17}$$

which is consistent with the total variance as conducted by Rubin (1987, p.21). Using the total variance formula of Pareto's coefficients, confidence intervals can be computed, which in the remaining part of this paper will be referred to as CI. The CI's lower bound 3.18 and upper bound 3.19 are computed as

$$\overline{\widehat{\alpha}} - [\mathcal{N}_{0.975}]_{df(1)} * \sigma_T \tag{3.18}$$

$$\overline{\widehat{\alpha}} + [\mathcal{N}_{0.975}]_{df(1)} * \sigma_T \tag{3.19}$$

3.3.2 Testing the Equality of Pareto's Coefficients

We start from the hypothesis that the wealth distribution is approximately Pareto distributed among the top-tail of the wealth distribution. If this is the case, we expect that, due to the linearity in the Pareto distribution, when observing wealth in logarithm form, the estimated coefficients should remain stable across the estimations based on different thresholds and datasets. Thus, to test the equality of Pareto's coefficients estimated according to Rubin (1987) statistically, we require a test that examines whether two estimates have the same mean while considering different variances and sample sizes between the two estimations. Therefore, *Welsh's t-test* appears to be sufficient to test the equality of the estimated Pareto's coefficients. We applied the test procedure according to Welch (1947) as described below.

Consistently with section 3.2 and subsection 3.3.1, for both the SOEP and the Pretest data we assume that the estimates with the total variance computed in Equations 3.7 and 3.17, respectively, are appropriate estimators for the simple linear regression model as set up in Eq. 3.6 in section 3.2. Adjusting to the presence of multiply imputed data according to Rubin (1987) allows the presumption that the estimates of alpha follow the normal distribution, in case the standard error in the simple linear regression model is a normally distributed random variable with mean 0 and unknown variance σ^2.

Therefore, the null hypothesis $H_0 : \overline{\widehat{\alpha}}_{SP} = \overline{\widehat{\alpha}}_{PT}$ stating that two estimates of Pareto's coefficients for the SOEP and the Pretest dataset are equal can be tested against the alternative hypothesis $H_1 : \overline{\widehat{\alpha}}_{SP} \neq \overline{\widehat{\alpha}}_{PT}$ using the test statistic

$$T = \frac{\overline{\widehat{\alpha}}_{SP} - \overline{\widehat{\alpha}}_{PT}}{\sigma_\Delta} \sim \tau_{df}, \tag{3.20}$$

where

$$\sigma_\Delta = \sqrt{\frac{\sigma_T^2\left(\overline{\widehat{\alpha}}_{SP}\right)}{n_{SP}} + \frac{\sigma_T^2\left(\overline{\widehat{\alpha}}_{PT}\right)}{n_{PT}}} \tag{3.21}$$

is an unbiased estimator of the variance of each of the two samples. The degrees of freedom associated with this variance estimate is approximated using the *Welch–Satterthwaite Equation*:

$$df \quad \approx \quad \frac{\left(\frac{\sigma_{SP}^2}{n_{SP}} + \frac{\sigma_{PT}^2}{n_{PT}}\right)^2}{\frac{\sigma_{SP}^4}{n_{SP}^2 * (n_{SP}-1)} + \frac{\sigma_{PT}^4}{n_{PT}^2 * (n_{PT}-1)}} \tag{3.22}$$

If the null hypothesis is true, the test statistic is expected to follow a t-distribution with *df* degrees of freedom. Hence, the null hypothesis can be rejected at a significance level of 0.05 if T exceeds the corresponding value of the cumulative distribution function of the t-distribution, given the significance level and the degrees of freedom. If T exceeds t_{df}, both estimates of Pareto's coefficient $\overline{\widehat{\alpha}}$ are statistically different at a significance level of 95%. This method allows to examine the equality of the Pareto's coefficients $\overline{\widehat{\alpha}}$ as computed in section 3.3.1 across datasets for equal threshold values. For robustness checks, we additionally use the Welch's t-test to test the equality of two estimates of alphas within the same dataset and different threshold values.

3.4 Imputing the Top Net Wealth

As part of the analysis, we intend to estimate and impute the total wealth held at the top 5%, 1% and 0.1% of the distribution in Germany, as well as their respective percentile values. In this section, we describe this imputation procedure.

Based on the estimated Pareto's coefficient $\overline{\widehat{\alpha}}$ and the lower bound $\underline{y}$, we are able to estimate the apportioning of wealth among the top-tail of the distribution using the inverse of the ccdf, as in

$$\hat{y}_i = \frac{\underline{y}}{(1 - F(y_i))^{1/\overline{\widehat{\alpha}}}} \text{ where } y_i \geq \underline{y} \text{ and } F(y_i) \in [p^{th}; 1) \tag{3.23}$$

where $p^{th} \in \{0.95; 0.99\}$, respectively.

Remember, however, that we assume that the data is Pareto distributed only at the interval over the 95th and the 99th percentiles, respectively. With this in mind, we have to modify the term in parenthesis in the denominator of Eq. 3.23 so we can correctly impute the top wealth values based on our estimation results. Focusing for a moment only on the 95th percentile threshold, the basic idea of this correction is to generate another variable, call it *synthetic distribution*, to expand the range within the 95th and the (approximately) 100th percentile, so that the 5% portion of the distribution to be estimated indeed covers the whole *synthetic* distribution from the 0th to the 100th percentile. Note that the 100th percentile, where $F(y) = 1$ is never reached, due to the construction of ccdf ($P \equiv 1 - F(y)$) and $\ln 0$ is undefined. For simplicity, we drop the term *approximately* subsequently.

The generated *synthetic* distributions $\tilde{F}_{p^{th}}(y)$ are then defined as

$$\tilde{F}_{p^{th}}(y) = \left(F(y) - p^{th}\right)\left(\frac{1}{1 - p^{th}}\right) \text{ where } p^{th} \in \{0.95; 0.99\} \tag{3.24}$$

Since we are interested in the top 5%, 1% and 0.1%, it is worth mentioning that for the values of F of $0.95, 0.99$ and 0.999 we obtain $0.00, 0.80$ and 0.98 for the *synthetic* $\tilde{F}$, when setting the threshold at the 95th percentile. At the 99th percentile, we refrain from estimating the top 5%, since this value is lower than the threshold, and, therefore, undefined in the Pareto distribution. The remaining 0.99 and 0.999 percentiles (of F) is mapped to 0.0 and 0.9 (of $\tilde{F}$), respectively.

We proceed to estimate and impute the top distribution via the *synthetic* distribution, as in

$$\hat{y}_i = \frac{\underline{y}}{(1 - \tilde{F}_{p^{th}}(y_i))^{1/\overline{\widehat{\alpha}}}} \text{ where } y_i \geq \underline{y} \text{ and } \tilde{F}(y_i) \in [0; 1) \tag{3.25}$$

With the imputed values, we can easily extract the estimated percentiles, as well as calculate the total wealth of those richer than the top three percentiles of interest. Remember that this procedure is followed for the estimated $\overline{\widehat{\alpha}}$ based on the Pretest as well as on the SOEP datasets. For the imputation, however, we use only the SOEP dataset as our basis and impute values for the observations above the threshold using the estimated $\overline{\widehat{\alpha}}$ coefficients of the SOEP regression and also of the Pretest regression, each with the lower bound set at the 95th and at the 99th percentiles. Since the Pretest had the intent to fill the missing information on the top end of the SOEP dataset, we refrain from following the procedure to impute the top percentiles of the Pretest dataset for any values of $\overline{\widehat{\alpha}}$ and lower bounds. Further, we conduct the estimations across all five implicates. The results are displayed in Tables 4.3 and 4.4.

4 Results

In this chapter we present and discuss the outcomes of our analysis. In section 4.2 we describe the results of the estimated Pareto's coefficient. Further, in section 4.2, we show the results of our imputation of the top-tails of the wealth distribution in the SOEP. Additionally, in section 4.3 we verify the robustness of our results and in section 4.4 we discuss the results.

4.1 Estimated Pareto's Coefficients

The results of our estimations with a threshold value set at the 95th percentile are listed in Table 4.1 and with a threshold value set at the 99th percentile in Table 4.2.

Focusing on the 95th percentile threshold based on the SOEP, the estimated coefficients range from 1.899 to 1.949. On a 95% CI, the true value of the Pareto's coefficient lies between 1.806 and 2.013. The estimated coefficients based on the Pretest range from 0.716 to 0.766 and on a 95% CI, the true value lies between 0.646 and and 0.817. By comparing the numbers of both stated CIs, we observe that they do not overlap. This becomes apparent in Figure 4.1 where we plotted the CIs for both datasets. Also, our applied Welch's t-test, which tests $\overline{\hat{\alpha}}_{SP}$ against $\overline{\hat{\alpha}}_{PT}$ on equality, results with a p-value of approximatelly zero, indicating that we can highly reject the null hypothesis that they are equal.

At the 99th percentile threshold, the estimated coefficients based on the SOEP range from 1.978 to 2.114 and for the Pretest from 0.832 to 0.927. On a 95% CI, the true value of Pareto's coefficient based on the SOEP lies between 1.809 and 2.27 and for the Pretest between 0.707 and 1.035. Again, we can reject the null hypothesis that both means $\overline{\hat{\alpha}}_{SP}$ and $\overline{\hat{\alpha}}_{PT}$ are equal with a p-value at 0.000. If we compare the CIs of the $\overline{\hat{\alpha}}_{SP}$ based on the 95th and on the 99th percentile threshold, we see that it is slightly increasing (from 1.806 to 1.809 for the lower CI limit and from 2.013 to 2.27 for the upper CI limit). The same is noticeable for the CI of $\overline{\hat{\alpha}}_{PT}$ (from 0.646 to 0.707 and 0.817 to 1.035, respectively). Further, the CI limits for both, SOEP and Pretest, increase in absolute numbers if we increase the threshold percentile from 95th to

	$\hat{\alpha}_{SP}$	N_{SP}	$\hat{\alpha}_{PT}$	N_{PT}	*threshold*
$ln(nw_1)$	1.899	1808991	.725	303270	563800
	(.0001)		(.0003)		
$ln(nw_2)$	1.949	1806357	.722	306794	563570
	(.0001)		(.0004)		
$ln(nw_3)$	1.89	1808323	.766	304166	571000
	(.0001)		(.0004)		
$ln(nw_4)$	1.894	1809090	.729	310152	573000
	(.0002)		(.0003)		
$ln(nw_5)$	1.915	1808598	.716	304500	572000
	(.0001)		(.0003)		
$\overline{\hat{\alpha}}$	1.91		.732		
	(.0028)		(.0019)		
CI lower (.025)	1.806		.646		
CI upper (.975)	2.013		.817		

Source: SOEP (2012, v33.1) and Pretest 2017.

Note: CIs computed in Eq. 3.18 & 3.19 (see chp. 3) according to Rubin (1987), *ln(nw)* denotes the logarithmic net wealth with applied household weights for the SOEP and applied reweighting for the Pretest, *sd* in parenthesis.

Table 4.1: Estimation Results of Pareto's Coefficients (Threshold at p95)

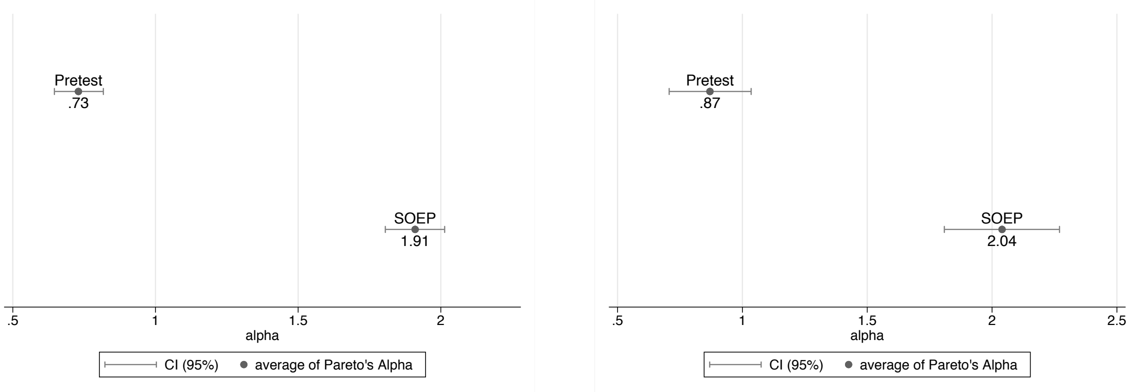
	$\hat{\alpha}_{SP}$	N_{SP}	$\hat{\alpha}_{PT}$	N_{PT}	*threshold*
$ln(nw_1)$	1.978	359338	.832	230918	1335000
	(.0005)		(.0003)		
$ln(nw_2)$	2.061	358418	.878	226496	1316000
	(.0004)		(.0004)		
$ln(nw_3)$	1.998	358441	.927	222804	1350000
	(.0005)		(.0004)		
$ln(nw_4)$	2.114	360819	.879	218216	1371624
	(.0007)		(.0004)		
$ln(nw_5)$	2.047	361211	.839	216088	1371900
	(.0005)		(.0003)		
$\overline{\hat{\alpha}}$	2.039		.871		
	(.0138)		(.007)		
CI lower (.025)	1.809		.707		
CI upper (.975)	2.27		1.035		

Source: SOEP (2012, v33.1) and Pretest 2017

Note: CIs computed in Eq. 3.18 & 3.19 (see chp. 3) according to Rubin (1987), *ln(nw)* denotes the logarithmic net wealth with applied household weights for the SOEP and applied reweighting for the Pretest, *sd* in parenthesis.

Table 4.2: Estimation Results of Pareto's Coefficients (Threshold at p99)

99th. Conducting the Welch's t-test, we can reject the null hypothesis that the means $\overline{\overline{\alpha}}_{SP}$ and $\overline{\overline{\alpha}}_{PT}$ are equal. This fact is visible in Figure 4.1 where we can clearly see that both estimated CIs of $\overline{\overline{\alpha}}_{SP}$ and $\overline{\overline{\alpha}}_{PT}$ do not overlap.



(a) Estimated CIs of $\hat{\alpha}_{SP}$ and $\hat{\alpha}_{PT}$, Threshold at p95

(b) Estimated CIs of $\hat{\alpha}_{SP}$ and $\hat{\alpha}_{PT}$, Threshold at p99

Figure 4.1: Comparison of CIs of $\hat{\alpha}_{SP}$ and of $\hat{\alpha}_{PT}$

If we compare the absolute values of the CIs from $\overline{\overline{\alpha}}_{SP}$ evaluated at the 95th percentile threshold to those at the 99th percentile threshold, we see that they overlap (at 95th $\overline{\overline{\alpha}}_{SP}$ range from 1.81 to 2.01 and at 99th $\overline{\overline{\alpha}}_{SP}$ range from 1.81 to 2.27). Analogeously, the CIs of the $\overline{\overline{\alpha}}_{PT}$ overlap, too (at 95th the CI of $\overline{\overline{\alpha}}_{PT}$ is 0.65 to 0.82 compared to 99th with 0.71 to 1.04). Conducting the Welch's t-test, however, we still have strong evidence to reject the hypothesis that the estimated coefficients are equal for the estimations conducted within the dataset and across threshold levels, both for the SOEP and the Pretest datasets.

To draw reliable conclusions about the stability of our estimated coefficients across the thresholds, we additionally conducted the Welch's t-test to test $\overline{\overline{\alpha}}_{SP}$ at the 95th against $\overline{\overline{\alpha}}_{SP}$ at the 99th percentile threshold. We can reject the null at a significance level of 1%. In fact, the means of both CIs do not lie in each other's range. Analogously for the estimated $\overline{\overline{\alpha}}_{PT}$s at the 95th and 99th percentile threshold, we can highly reject the Welch's t-test's null, too.

4.2 Imputation of the Top-Tail of the Wealth Distribution in the SOEP

We list our results of the imputation for selected top percentiles in Table 4.3. First, we imputed the top 5%, the top 1% and the top 0.1% of the wealth distribution in the SOEP by using the $\overline{\overline{\alpha}}_{PT}$ for both threshold percentiles. In a second step, we repeated the imputation, using $\overline{\overline{\alpha}}_{SP}$

instead. By doing that, we can now compare both results. The values indicate the net wealth in m euros that the top percentages at least contain. For instance, the least rich household among the SOEP's top 5% most wealthy households owns 0.57m euros.

We observe that, compared to using $\overline{\overline{\alpha}}_{SP}$, all imputed values are higher for both threshold percentiles when $\overline{\overline{\alpha}}_{PT}$ is applied: For example, utilizing $\overline{\overline{\alpha}}_{SP}$ to impute to the top tails from the 95th percentile, we estimate that the top 1% contains at least around 1.3m euros, whereas using the $\overline{\overline{\alpha}}_{SP}$ we derive an imputation of at least 5.5m euros. The difference between both imputations is much larger if we impute a higher top percentile such as the top 0.1%. With $\overline{\overline{\alpha}}_{SP}$ we then impute a value of around 4.5m euros, while for $\overline{\overline{\alpha}}_{PT}$ estimate a value of around 160m euros. The results of the imputed values are not that high if we use the Pareto's coefficients estimated from the 99th percentile threshold. For the top 0.1% we impute minimum values of around 4.3m euros with $\overline{\overline{\alpha}}_{SP}$ and of around 24.5m euros with $\overline{\overline{\alpha}}_{PT}$. Further, the higher we set the threshold, the less divergent are the imputed values using $\hat{\alpha}_{SP}$ compared to those using $\overline{\overline{\alpha}}_{PT}$. Note that the lack of variability at the estimated percentiles corresponding to each respective threshold, that is, top 5% at 95th and top 1% at the 99th percentile, is due to the fact that the values of the estimated Pareto's coefficient $\hat{\alpha}$ have negligible influence on the resulting estimation at values close to the threshold, that is, where the *synthetic* distribution is close to zero (see Eq. 3.25).

In Table 4.4 we list the results of our calculation of the total wealth held among the top three percentiles of interest. For example, with $\overline{\overline{\alpha}}_{SP}$ estimated at the 95th threshold, we calculate a total wealth of about 1.02trn euros. Using $\overline{\overline{\alpha}}_{PT}$, we calculate a total value of 150.8trn euros.

If we calculate the total wealth of the top 1% with the $\overline{\overline{\alpha}}_{SP}$ estimated at the 99th percentile threshold, we estimate the the total wealth to be around 988.7bn euros, whereas applying $\overline{\overline{\alpha}}_{PT}$ we receive a value of about 15.115.2trn euros. We notice that all values calculated with $\overline{\overline{\alpha}}_{PT}$ are higher than those with $\overline{\overline{\alpha}}_{SP}$. Similar to the results in Table 4.3, all calculated values are higher if we use $\overline{\overline{\alpha}}_{PT}$.

In Figure 4.2 we plot the Pareto distribution of the SOEP wealth distribution in grey and add the imputed top-tails imputed by using $\overline{\overline{\alpha}}_{PT}$ above the 95th and 99th percentile threshold. The imputed values from the mean of our estimated $\overline{\overline{\alpha}}_{PT}$ are displayed in red. The 95% CIs are illustrated in yellow. We notice that we impute higher values than the original data. Also, the course of the imputed values is flatter. If we compare graph (a) and (b), compared to using the 99th percentile as a threshold value, we notice that when using the 95th percentile we estimate higher values.

	$\hat{\alpha}_{SP}$			$\hat{\alpha}_{PT}$		
	CI_{low}	$mean$	CI_{high}	CI_{low}	$mean$	CI_{high}
Threshold at 95th percentile						
Top 5%	.569	.569	.569	.569	.569	.569
Top 1%	1.266	1.327	1.388	4.083	5.488	6.893
Top 0.1%	4.01	4.516	5.022	69.979	160.694	251.41
Threshold at 99th percentile						
Top 1%	1.349	1.349	1.349	1.349	1.349	1.349
Top 0.1%	3.754	4.314	4.874	12.736	24.458	36.181

Source: SOEP (2012, v33.1) and Pretest 2017.

Note: All values in mio. euros, 95%-confidence intervals calculated according to Rubin (1987), estimation based on our imputation. We do not estimate top 5% based on the 99th percentile.

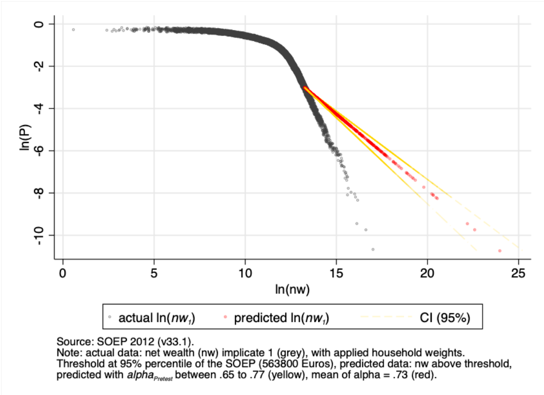
Table 4.3: Estimation of Selected Top Wealth Percentiles in the SOEP

	$\hat{\alpha}_{SP}$			$\hat{\alpha}_{PT}$		
	CI_{low}	$mean$	CI_{high}	CI_{low}	$mean$	CI_{high}
Threshold at 95th percentile						
Top 5%	2055.642	2185.564	2315.485	33881.166	153131.6	272382.03
Top 1%	918.172	1024.592	1131.011	31876.257	150788.97	269701.68
Top 0.1%	286.46	342.525	398.591	27246.521	142362.9	257479.29
Threshold at 99th percentile						
Top 1%	880.585	988.656	1096.727	3612.883	15115.246	26617.609
Top 0.1%	243.014	314.572	386.131	2500.333	13567.3	24634.267

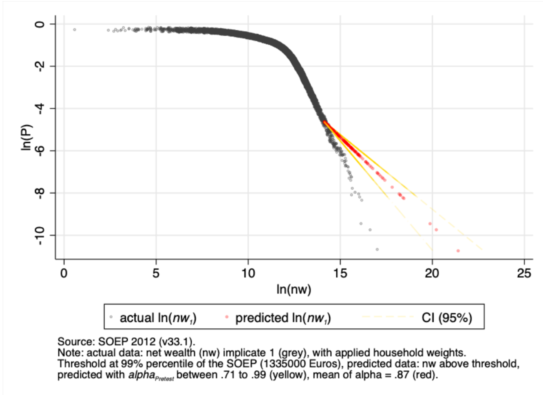
Source: SOEP (2012, v33.1) and Pretest 2017.

Note: All values in billion Euros, 95%-confidence intervals calculated according to Rubin (1987), estimation based on our imputation. We do not estimate top 5% based on the 99th percentile.

Table 4.4: Estimation of Total Wealth of Selected Top Wealth Percentiles in the SOEP

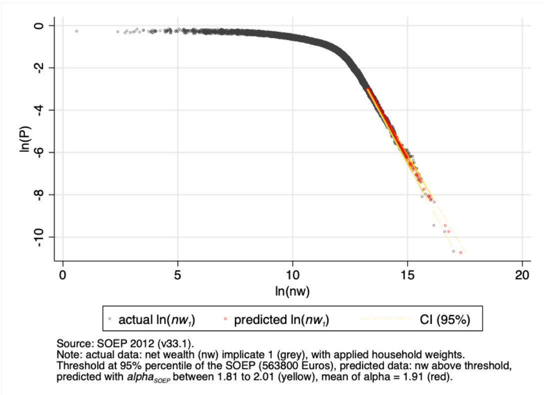


(a) Implicate 1, p95 Threshold

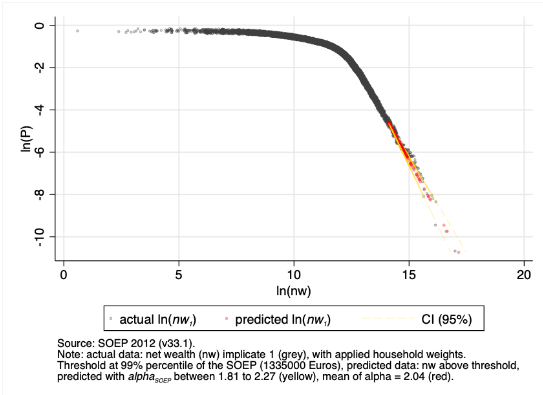


(b) Implicate 1, p99 Threshold

Figure 4.2: Actual vs. Imputed Top Net Wealths of the SOEP ($\hat{\alpha}_{PT}$, p95 and p99)



(a) Implicate 1, p95 Threshold



(b) Implicate 1, p99 Threshold

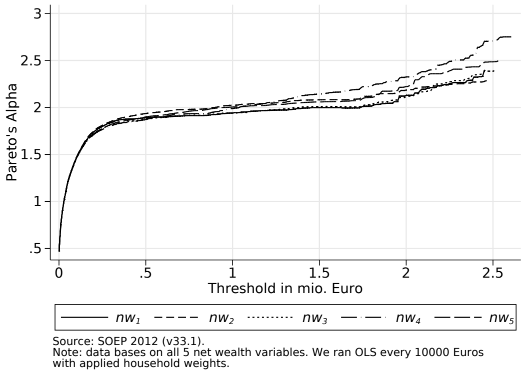
Figure 4.3: Actual vs. Imputed Top Net Wealths of the SOEP ($\hat{\alpha}_{SP}$, p95 and p99)

For comparative purposes, in Figure 4.3 we also plot the imputed values utilizing $\overline{\overline{\alpha}}_{SP}$. For both threshold values, the trend of the imputed top-tails coincides with the actual data.

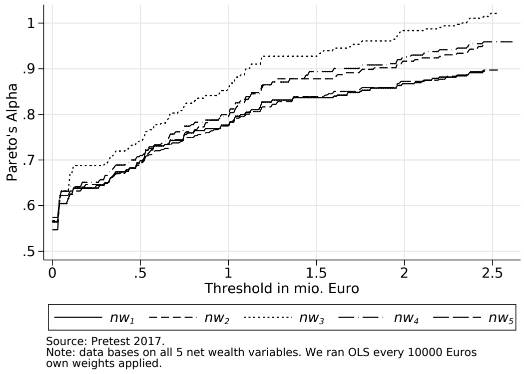
4.3 Sensitivity Analysis of Threshold Levels

By using threshold values between 0 and 2.5m euros in steps of 100.000 euros, we estimated the corresponding Pareto's for both datasets. The resulting "α-threshold-combinations" are depicted in Figure 4.4. We observe that for threshold values below 0.5m euros, the Pareto's coefficients increase with increasing threshold values. From 0.5m euros to about 1.5m euros, the Pareto's Coefficients seem to stabilize between 1.9 to 2.1. For threshold values of 1.5m and higher, the corresponding values of α increase with rising threshold values.

The plotted Pareto's coefficients of the implicates of the Pretest are increasing over the whole range (from 0.0m to 2.5m euros). We do not recognize any stagnation. Due to the limited amount of observations we observe a flat pathway of the coefficients around threshold values of around 1.5m and 1.8m euros.



(a) SOEP Implicates



(b) Pretest Implicates

Figure 4.4: α-threshold-combinations of SOEP and Pretest

4.4 Discussion

In regards to our analysis, we evaluate two main aspects: First, we discuss about the stability of the Pareto's coefficient and the choice of the threshold values. Second, we examine the integration of the Pretest to the SOEP.

Stability of Pareto's coefficient α

Our estimated Pareto's coefficients $\overline{\widehat{\alpha}}_{SP}$ and $\overline{\widehat{\alpha}}_{PT}$ lie between 1.8 and 2.3 and between 0.6 and 1.0, respectively. Compared to related studies as Dalitz (2018) who estimated a value of around 1.5, or Westermeier and Grabka (2015) who estimated values between 1.3 to 1.4, our estimations based on the SOEP lie above them and our estimations based on the Pretest lie below. Based on strong evidence that we receive from our conducted tests, we conclude that our estimated CIs of $\overline{\widehat{\alpha}}$ based on the SOEP and Pretest are not equal. Evaluating whether the CIs are visually overlapping, it appears unequivocally that they do not even touch. Different to studies from Jenkins (2017) who found robust results at the thresholds p95 and p99, we do not identify clear stability of our estimated Pareto's coefficients across these thresholds. We further noticed that the choice of threshold values may affect the estimated Pareto's coefficient. ALso, we find no evidence for the theoretically claimed linearity in Pareto's distributions above certain threshold values.

Although the Pretest was polled with the SOEP questioning instruments, the estimated Pareto's coefficients vary from those conducted using the SOEP data. However, sampling the German population with the same instruments is not the only requirement to end up in stable estimates. Firstly, We need to consider the different sampling designs of both datasets. Secondly, the sample size of the Pretest is considerably small ($n = 124$), which limits the validity of our results. We noticed that using the SOEP with a sample size of 14,982 households, the estimation of the Pareto's coefficient appears to stabilize around 2.0, whereas the results based on the Pretest seem to be unstable (see Figure 4.4). We believe that the sample is insufficiently small. For the problem of small-sample settings, for instance Brzezinski (2016) recommends the application of a probability integral transform statistic estimator (PITSE) which we might recommend for future analyses. However, as the the Pretest is the precursor of the main study ("*Hauptuntersuchung*"), the issue of small-sample settings might be eliminated in the future.

Integration of the Pretest to the SOEP

Other studies on the top-tails of the wealth distribution in Germany like the ones of Bach et al. (2018) or Westermeier and Grabka (2015) include rich-lists such as from the Forbes- or the Managermagazin without adjusting for e.g. socio-demographic or regional characteristics to the weighting scheme of the SOEP. For the Pretst, we do not have a weighting scheme that corresponds with the household weights of the SOEP to our disposal. Therefore, we did not integrate the Pretest to the SOEP dataset. With appropriate weights that allow the integration of the Pretest to the SOEP (wich is probably an objective of the main study), we expect more stable results and Pareto's coefficients that are close to those of the related research.

Finally, we cannot confirm the validity of Pareto's Law. We could not find clear evidence for the existence of a Pareto's coefficient that is valid for the top-tails of the wealth distribution in Germany.

5 Summary and Conclusion

Many economists suggest that the distribution of high wealth follows a Pareto distribution. This means, that starting from a certain threshold value, the logarithmized wealth displays a linear relationship to the logarithmized complementary cumulative density. The basic shape of the Pareto distribution is then determined by a Pareto's coefficient α. In this paper, we attempted to resolve the research question of whether the top-tail of the German wealth distribution follows a Pareto distribution and, if so, which Pareto's coefficients we can observe. Since information about very rich individuals and households in Germany is scarce, Schröder et al. (2018) developed a novel sampling strategy for surveying high-worth individuals. Our data foundation consists of net wealth data on household level from the Pretest and the SOEP data from the year 2012. A comparison of the descriptive statistics of both datasets displays that the Pretest contains observations at the top of the wealth distribution in Germany. Thus, the Pretest seems to be making a contribution to fixing the lack of data about high-worth individuals in the SOEP.

To estimate the Pareto coefficients, we followed the OLS regression method of estimating a power law (see Vermeulen (2018)). With our estimates, we then evaluated and imputed the net wealth owned by the top-tail of the distribution and calculated the percentile net wealth, as well as the total net wealth held by three top percentiles of interest. With the intend of checking if the Pareto distribution can adequately describe the net wealth distribution in Germany, we conducted the same procedure based on the SOEP, as well as on the Pretest datasets. For robustness, we also apply the same procedure based on two threshold values, one set at the 95th and the other at the 99th net wealth percentile of the SOEP dataset.

For estimations based on SOEP data we derive Pareto coefficients of 1.91 and 2.04 for the 95th and 99th percentile thresholds, respectively. The Pretest estimations results in coefficients of 0.73 and 0.87, for the respective 95th and 99th percentiles. Due to our conducted tests, we highly reject the hypothesis that the coefficients are equal across datasets. When focusing on each dataset separately and testing the variability across thresholds, the coefficients are more stable. Statistically, however, we can still reject the hypothesis of equality.

Based on the resulting Pareto coefficients, we estimate and impute the top-tail of the distribution with focus on the three percentiles of interest, namely the top 5%, top 1% and top 0.1%. The estimations based on the Pareto coefficients from the SOEP estimations result in more plausible figures. Focusing first on the threshold set at the 95th percentile, for the top 5% we estimate minimal values of the net wealth of 0.569m euros. For the top 1% and the the top 0.1%, we calculate minimal values of 1.32m euros (1.27m – 1.39m)[1] and 4.52m euros (4.01m – 5.02m), respectively. Based on the Pretest estimations and the corresponding small Pareto's coefficients of 0.73 to 0.87, we derive very high estimates of the top percentiles. Namely, 0.569m euros for the top 1%, 5.488m euros (4.08m – 6.89m) for the top 1% and for the top 0.1% results in 160.694m euros (69.98m – 251.41m). Focusing on the threshold set at the 99th percentile, the figures for the SOEP estimations do not differ substantially from those described above, while the figures for the Pretest are much lower, ranging from 1.3m euros for the top 1% to 24.5m euros for the top 0.1%. This discrepancy between SOEP and Pretest is a contradiction to the assumption of obtaining a linear relationship in proportions assumed by the Pareto distribution.

We observe the same pattern when estimating the total sum of net wealth held at the top three percentiles of interest. Based on the SOEP and the 95th percentile threshold, the estimated wealth owned by the top 5% amounts to 2.186trn euros (2.06trn – 2.32trn). The top 1% holds 1.02trn euros (918bn – 1.13trn) and the top 0.1% owns 342.5bn euros (286bn – 399bn). Based on the 99th percentile threshold, the SOEP estimations results in the top 1% owning 988.7bn euros (880.6bn – 1.097trn) and the top 0.1% owning 314.6bn euros (243bn – 386bn). Focusing on the Pretest regressions and the 95th percentile threshold, the estimated wealth held at the top-tail of the distribution is very excessive, with the top 5% owning 153.1trn euros (33.88trn – 272.38trn). The top 1% would own 150.789trn euros (318.76trn – 269.70trn) and the top 0.1% would hold 142.36trn euros (272.47trn – 257.48trn). Based on the 99th percentile threshold and the Pretest dataset, the results are still very inflated, but following the same pattern as those obtained with the threshold set at the 95th percentile.

Conclusion

We started from the general hypothesis that the Pareto distribution is a good approximation of the *actual* distribution of net wealth in Germany. With actual meaning that this theoretical distribution is capable of representing the real distribution of net wealth in Germany, overcoming the under-representation bias at the top-tail. If our general hypothesis is true, we expect a linear relationship in proportions between the net wealth and its empirical complementary cumulative

[1] The values of the corresponding 95% confidence intervals' lower and upper bounds are displayed in parenthesis.

distribution. Practically, from some lower bound of net wealth we expect to obtain plausible and stable Pareto coefficients across our various regressions, regardless of the choice of threshold or dataset. Our results, however, present strong evidence opposing our general hypothesis. The estimated Pareto coefficients obtained from the SOEP dataset differ highly from those of the Pretest dataset, indicating another degree of wealth concentration within these two subgroups, contradicting the expected linear relationship in proportions.

If we pay attention to the estimations within datasets only, although still statistically diverging, there are some signs of stability on the regressions conducted based solely on the SOEP dataset, as shown in Fig.4.4(a). Although the evidence is weak, this stability could suggest that the Pareto distribution can be a good approximation for a range of the distribution, from around 0.5 to 1.7m euros. The rise in the estimated coefficients above this range may well indicate that Pareto distribution is indeed not adequate or, on the other hand, that the households on the very top-tail of the distribution tend to underestimate their own wealth, wrongly implying less wealth concentration (higher Pareto's $\overline{\widehat{\alpha}}$) than reality would show. In addition, even without any underestimation, it is possible that relatively large debt held by the top-tail pushes their net wealth down in such a manner that gross wealth could possibly be better approximated by the theoretical Pareto distribution. Further, a generalised Pareto or another distribution from the power law family might better describe the wealth distribution. Although we can not strongly support these hypotheses in this analysis, they could be the focus of further research.

In conclusion, we gather strong evidence against the hypothesis that the Pareto distribution is a good approximation of the real distribution of net wealth among the top-tail of the distribution. Further, small variations in the Pareto coefficient lead to very large variations in the estimation of wealth concentration at the top-tail of the distribution. Resting on the above, caution should be taken, were one to assume that the *actual* distribution were indeed Pareto distributed when conducting estimations of the top-tail of the wealth distribution and using those as basis for policy decisions on the topic of wealth inequality. Further, the lack of reliable data on wealth held at the top-tail of the distribution expose the need of better measurement capabilities focused on this segment of the population to overcome this information gap on the *actual* wealth concentration and the resulting estimates of wealth inequality. In this regard, the upcoming main study ("*Hauptuntersuchung*") based upon the Pretest is a welcome endeavor to fill in this gap.

Bibliography

S. Bach, A. Thiemann, and A. Zucco. The Top Tail of the Wealth Distribution in Germany, France, Spain, and Greece. *DIW Berlin Discussion Paper*, 2015(1502), 2015. ISSN 1556-5068. doi: 10.2139/ssrn.2655990. URL `http://www.ssrn.com/abstract=2655990`.

S. Bach, A. Thiemann, and A. Zucco. Looking for the Missing Rich: Tracing the Top Tail of the Wealth Distribution. *DIW Berlin Discussion Paper*, 2018(1717), 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3113153. URL `https://www.ssrn.com/abstract=3113153`.

M. Brzezinski. Do wealth distributions follow power laws? Evidence from 'rich lists'. *Physica A: Statistical Mechanics and its Applications*, 406:155–162, 2014. ISSN 03784371. doi: 10.1016/j.physa.2014.03.052.

M. Brzezinski. Robust estimation of the Pareto tail index: a Monte Carlo analysis. *Empirical Economics*, 51(1):1–30, 2016. ISSN 03777332. doi: 10.1007/s00181-015-0989-9.

W. G. Cochran. *Sampling Techniques*. Wiley, New York, 1977.

C. Dalitz. Estimating Wealth Distribution: Top Tail and Inequality. 2018. URL `http://arxiv.org/abs/1807.03592`.

M. M. Grabka and C. Westermeier. Editing and Multiple Imputation of Item Non-response in the Wealth Module of the German Socio-Economic Panel. *SOEP Survey Papers*, Series C – Data Documentation(272):39, 2015.

S. P. Jenkins. Pareto Models, Top Incomes and Recent Trends in UK Income Inequality. *Economica*, 84(334):261–289, 2017. ISSN 14680335. doi: 10.1111/ecca.12217.

W. B. Losse. Wirtschaftswoche: Wie eine neue studie die blackbox superreiche öffnen will. `https://www.wiwo.de/my/politik/deutschland/vermoegen-und-einkommen-wie-eine-neue-studie-die-blackbox-superreiche-oeffnen-will/24048394.html?ticket=ST-169956-ODVOhNWV5TqA5hcbOpFV-ap2`, 2019. Retrieved: 2019-03-13.

Oxfam. Public Good or Private Health. *Oxfam Briefing Paper*, January 2019, 2019. doi: 10.21201/2019.3651. URL https://www.oxfam.de/system/files/bp-public-good-or-private-wealth-210119-en_web.pdf.

V. Pareto. *Cours d'Économie Politique Professé a l'Université de Lausanne*, volume II. F. Rouge, Lausanne, 1897.

T. Piketty. *Capital in the Twenty-First Century.* Harvard University Press, 2014. ISBN 9780674430006. URL http://www.jstor.org/stable/j.ctt6wpqbc.

D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York, 1987.

C. Schröder, C. Bartels, M. M. Grabka, M. Kroh, and R. Siegers. A Novel Sampling Strategy for Surveying High-Worth Individuals - An Application Using the Socio-Economic Panel. *SOEPpapers on Multidisciplinary Panel Data Research*, 2018(978):40p., July 2018.

Statista. Anzahl der privathaushalte in deutschland seit 1991. https://de.statista.com/statistik/daten/studie/156950/umfrage/anzahl-der-privathaushalte-in-deutschland-seit-1991/, 2019. Retrieved: 2019-03-09.

P. Vermeulen. How Fat is the Top Tail of the Wealth Distribution? *Review of Income and Wealth*, 64(2):357–387, June 2018. ISSN 00346586. doi: 10.1111/roiw.12279.

B. Welch. The generalization of student's problem when several different population variances are involved. *Biometrika*, 34:28–35, 1947.

C. Westermeier and M. M. Grabka. Große statistische Unsicherheit beim Anteil der Top- Vermogenden in Deutschland. *DIW-Wochenbericht*, (7):11, 2015. doi: 10.1007/s10273-011-1262-2. URL https://www.diw.de/documents/publikationen/73/diw{_}01.c.496886.de/15-7-3.pdf.

Appendix

	count	mean	p50	p75	p90	p99	min	max
SOEP nw imp. 1	14,982	0.17	0.05	0.20	0.41	1.57	-3.81	50.78
SOEP nw imp. 2	14,982	0.17	0.05	0.20	0.40	1.55	-3.81	63.08
SOEP nw imp. 3	14,982	0.17	0.05	0.20	0.40	1.63	-3.81	16.21
SOEP nw imp. 4	14,982	0.17	0.05	0.20	0.41	1.68	-3.81	33.75
SOEP nw imp. 5	14,982	0.17	0.05	0.20	0.40	1.59	-3.81	37.19
Pretest nw imp. 1	124	11.95	2.10	5.98	22.42	202.71	-4.43	207.02
Pretest nw imp. 2	124	10.22	2.21	6.02	27.89	156.56	-2.58	207.02
Pretest nw imp. 3	124	8.29	1.97	4.89	18.15	156.56	-2.58	207.02
Pretest nw imp. 4	124	9.65	2.07	5.54	25.91	156.56	-12.54	207.02
Pretest nw imp. 5	124	11.27	2.05	5.54	22.97	206.55	-2.58	207.02

Source: SOEP 2012 (v33.1), Pretest 2017.

Note: Net wealth (nw) imputed, in mio. Euro, unweighted, for simplicity rounded.

Table A.1: Unweighted Descriptive Statistics of all Net Wealth Implicates

	count	mean	p50	p75	p90	p99	min	max
SOEP nw imp. 1	36,201,747	0.15	0.04	0.20	0.39	1.33	-3.81	50.78
SOEP nw imp. 2	36,201,747	0.15	0.04	0.20	0.38	1.32	-3.81	63.08
SOEP nw imp. 3	36,201,747	0.15	0.04	0.20	0.38	1.35	-3.81	16.21
SOEP nw imp. 4	36,201,747	0.16	0.05	0.20	0.39	1.37	-3.81	33.75
SOEP nw imp. 5	36,201,747	0.16	0.05	0.20	0.39	1.37	-3.81	37.19

Source: SOEP 2012 (v33.1).

Note: Net wealth (nw), imputed, weighted and displayed in mio. Euro, for simplicity rounded.

SOEP: with applied frequency weights on household-level (N=14982).

Table A.2: Weighted Descriptive Statistics of all Net Wealth Implicates of the SOEP

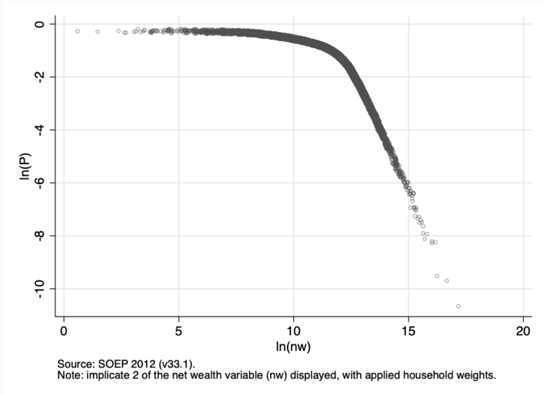
	count	mean	p50	p75	p90	p99	min	max
Pretest nw imp. 1	412,996	9.48	1.95	5.57	17.92	202.71	-4.43	207.02
Pretest nw imp. 2	412,996	8.21	1.98	5.76	21.10	156.56	-2.58	207.02
Pretest nw imp. 3	412,996	7.00	1.74	4.94	17.50	156.56	-2.58	207.02
Pretest nw imp. 4	412,996	7.76	1.76	5.01	21.45	156.56	-12.54	207.02
Pretest nw imp. 5	412,996	8.96	1.74	5.30	21.16	206.55	-2.58	207.02

Source: Pretest 2017.

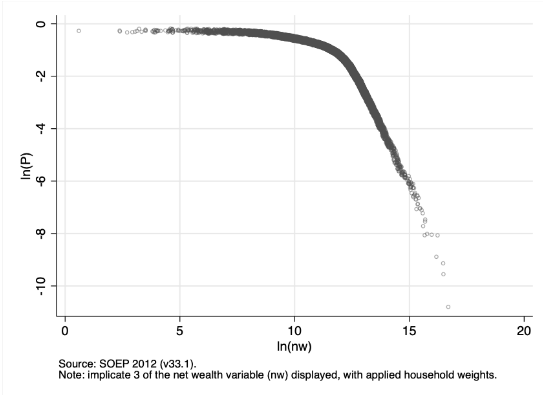
Note: Net wealth (nw), imputed, weighted and displayed in mio. Euro, for simplicity rounded.

Pretest: with own re-weighting scheme (N=124).

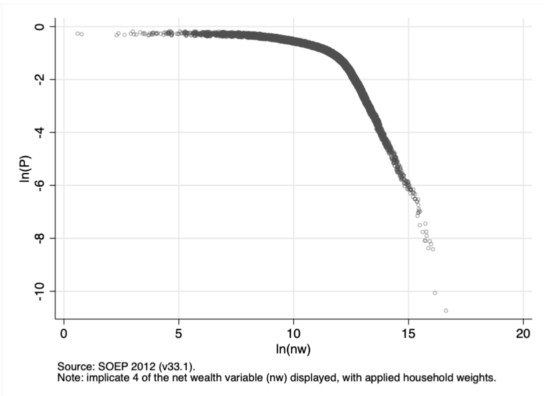
Table A.3: Weighted Descriptive Statistics of all Net Wealth Implicates of the Pretest



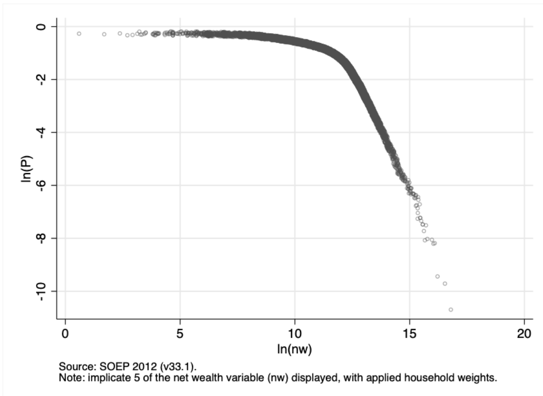
(a) Net Wealth Implicate 2



(b) Net Wealth Implicate 3

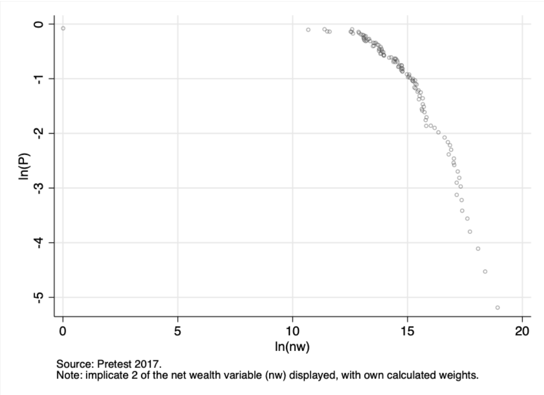


(c) Net Wealth Implicate 4

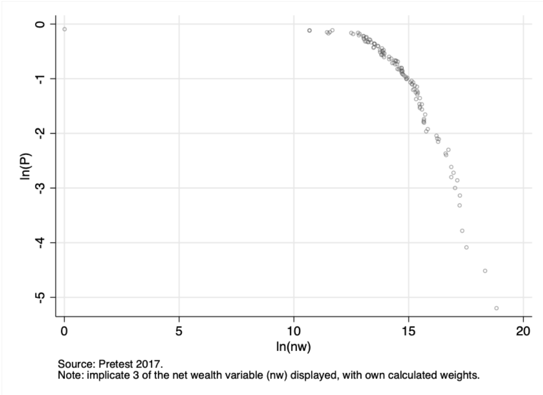


(d) Net Wealth Implicate 5

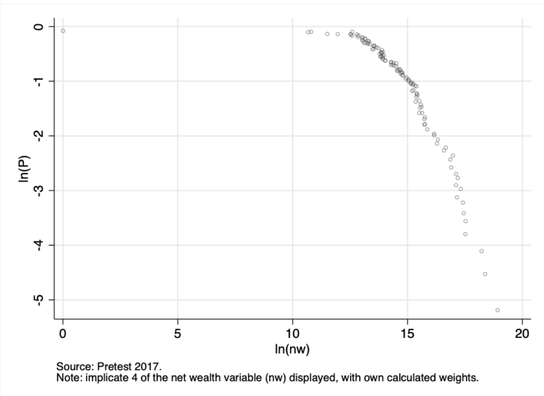
Figure A.1: Zipf Plots of Net Wealth Implicates of the SOEP



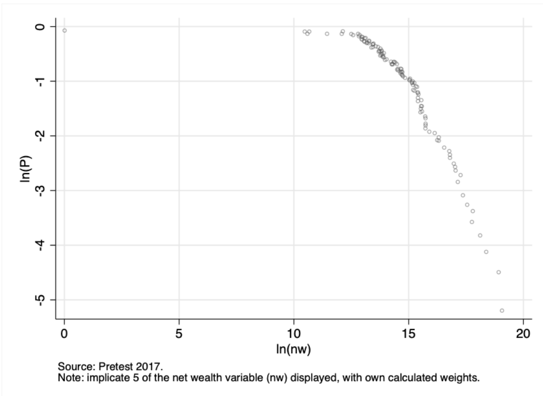
(a) Net Wealth Implicate 2



(b) Net Wealth Implicate 3

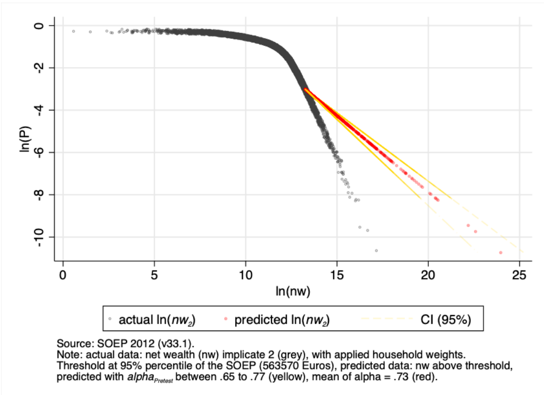


(c) Net Wealth Implicate 4

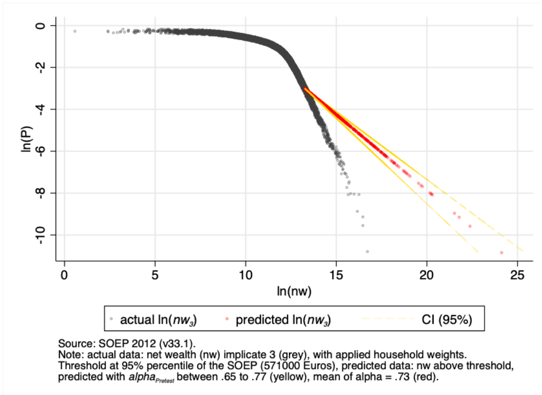


(d) Net Wealth Implicate 5

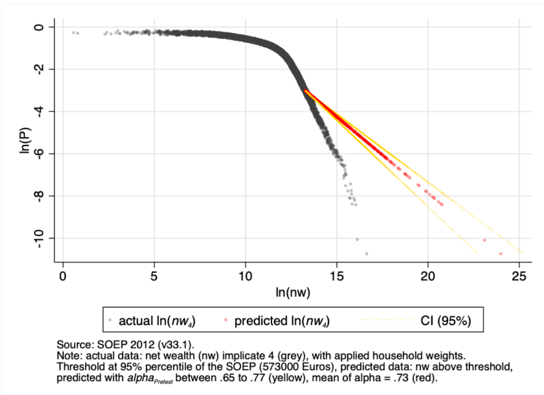
Figure A.2: Zipf Plots of Net Wealth Implicates of the Pretest



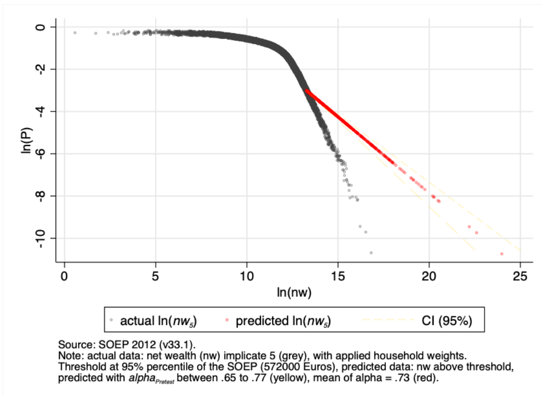
(a) Net Wealth Implicate 2



(b) Net Wealth Implicate 3

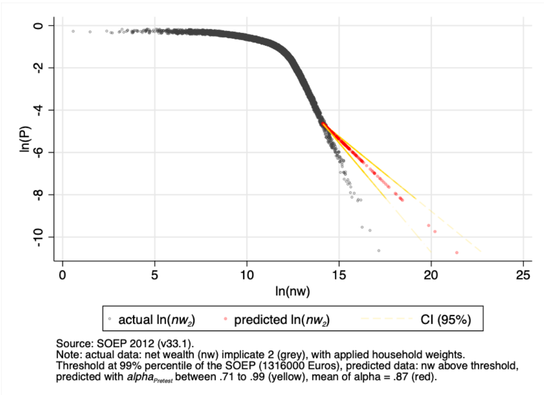


(c) Net Wealth Implicate 4

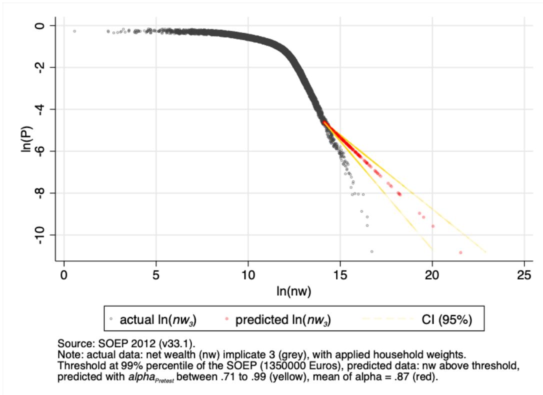


(d) Net Wealth Implicate 5

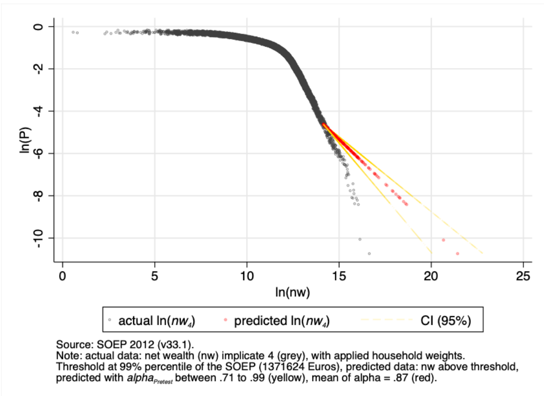
Figure A.3: Actual vs. Predicted Top Net Wealths of the SOEP ($\hat{\alpha}_{PT}$, p95)



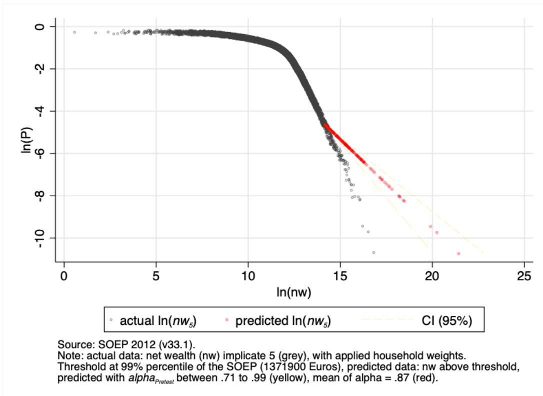
(a) Net Wealth Implicate 2



(b) Net Wealth Implicate 3

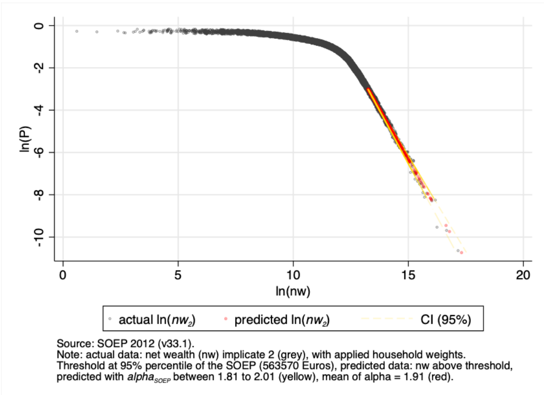


(c) Net Wealth Implicate 4

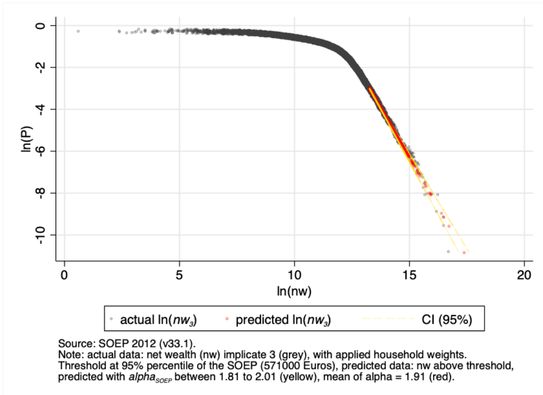


(d) Net Wealth Implicate 5

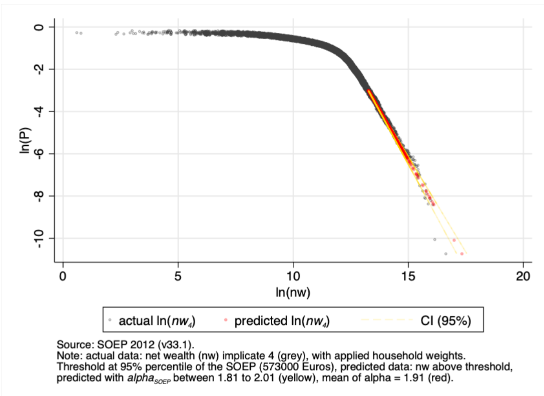
Figure A.4: Actual vs. Predicted Top Net Wealths of the SOEP ($\hat{\alpha}_{PT}$, p99)



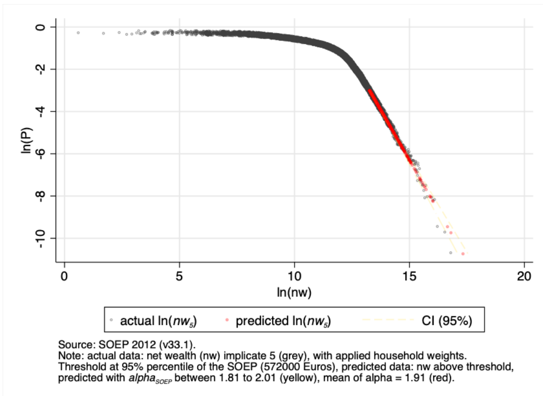
(a) Net Wealth Implicate 2



(b) Net Wealth Implicate 3

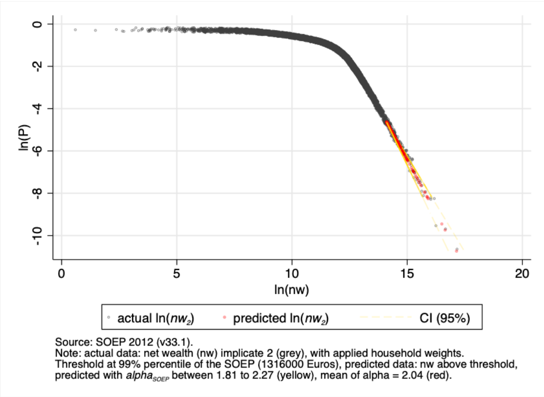


(c) Net Wealth Implicate 4

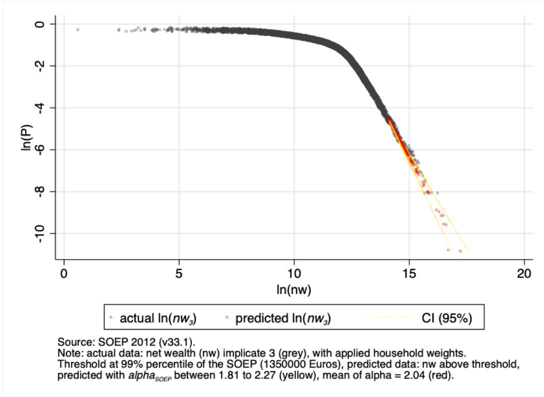


(d) Net Wealth Implicate 5

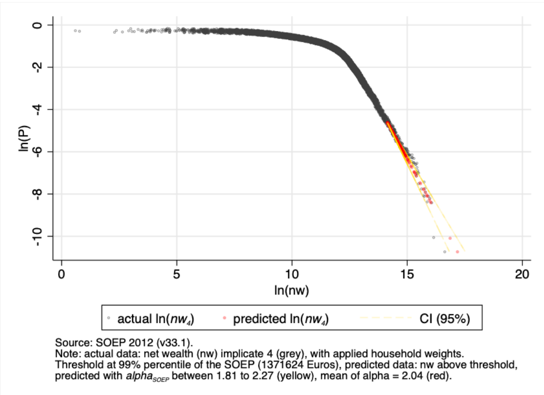
Figure A.5: Actual vs. Predicted Top Net Wealths of the SOEP ($\hat{\alpha}_{SP}$, p95)



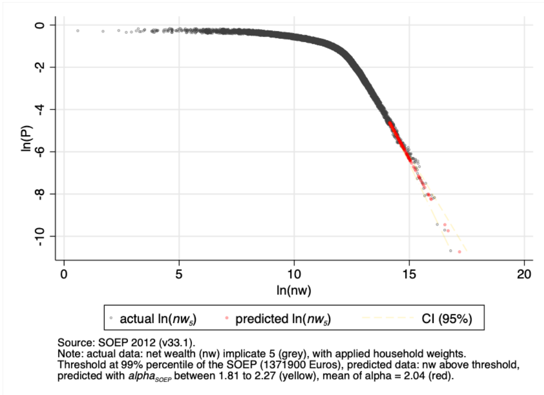
(a) Net Wealth Implicate 2



(b) Net Wealth Implicate 3

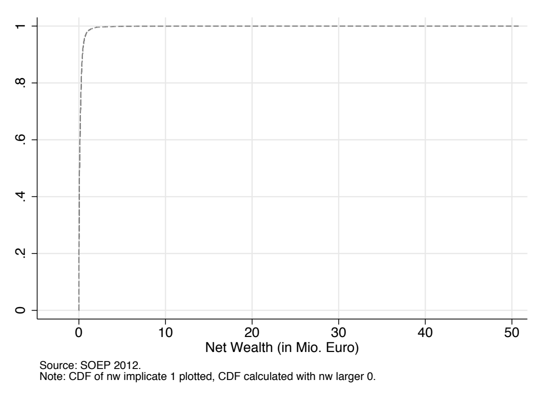


(c) Net Wealth Implicate 4

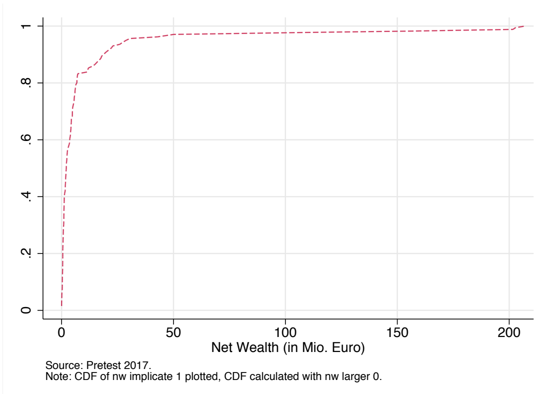


(d) Net Wealth Implicate 5

Figure A.6: Actual vs. Predicted Top Net Wealths of the SOEP ($\hat{\alpha}_{SP}$, p99)



(a) CDF of the SOEP



(b) CDF of the Pretest

Figure A.7: CDFs of SOEP and Pretest