Netspar

Network for Studies on Pensions, Aging and Retirement

Netspar THESES

Lara Evertsen

Backward Imputation of Financial Household Wealth

MSc Thesis 2013-022

Backward Imputation of Financial Household Wealth

L. Evertsen

Groningen, 27 August 2013

Masterthesis Econometrics
Supervisor: Prof. Dr. R.J.M. Alessie
Co-assessor: Prof. Dr. R.H. Koning
Project supervisor: R. van Ooijen MSc

Backward Imputation of Financial Household Wealth

Lara Evertsen

Abstract

This paper outlines a method to impute previous panel waves of checking and savings accounts and risky assets. The imputations are based on data on tax records, interest incomes and dividend returns for the waves to be imputed. Furthermore, future waves of the panel are used. In the imputation, household specific effects and autocorrelation in the error terms are taken into account. The imputation method is evaluated in two ways. First, the imputation is implemented for a wave where the actual values are known. The imputed values closely follow the actual distribution and the correlations between the imputed and actual variables are high. Next, the imputation is realized for the unknown waves. The distributions in each wave are compared with later known waves and external data. Cohort effects and ownerships rates are analyzed as well. The imputation method seems to work quite well.

Keywords: imputation, household wealth, panel data, backward prediction.

Contents

1 Introduction

Household wealth is defined as the sum of the market value of assets owned by household members minus the liabilities they own (Statistics Netherlands). The composition and distribution of household wealth receives much attention by policy makers. Household wealth can be used to smooth consumption over the life cycle, for instance to finance (early) retirement or other periods of low expected income. This can be either through the liquidation of assets or by the income streams generated from them. Moreover, the ownership of wealth can be used as a buffer against negative shocks which might lead to a reduction in income such as unemployment, illness or aging (Davies, Sandstrom, Shorrocks, and Wolff, 2006). In addition, wealth can be bequeathed to future generations. An extensive data set of household wealth can therefore provide insights into a number of relevant areas for policy makers. For example, the effects of tax policy or redistribution measures on the economic wellbeing and the degree of wealth concentration in the population can be examined.

Statistics Netherlands produces the administrative panel survey IPO (Inkomens Panel Onderzoek) where the same respondents are followed over time. IPO consists of two subpanels, IPO Income (IPO Inkomen) and IPO Wealth (IPO Vermogen). IPO Income exists since 1989 and it gives an overview of the yearly income distribution in the Netherlands. The primary sources of this panel are tax records and a database containing information on interest income and dividend returns. This database is supplied for all respondents in the IPO survey. In the Netherlands, wealth is taxed whenever its value exceeds a certain limit, a tax free allowance. IPO Income contains some variables on the size of wealth. However, they are based on tax records and, due to this allowance, only known for the upper percentiles of the sample.

Since 2005, the IPO Income panel is supplemented by the IPO Wealth panel. Where the IPO Income panel is primarily based on data from tax records, the IPO Wealth panel is based on administrative data from financial institutions. Household wealth components for the entire sample are reported, not just the upper percentiles. IPO Wealth gives an overview of the yearly distribution and composition of household wealth for the years 2005-2010.

In order to estimate the effects of tax policy or the financial crises for example, it would be useful to have a long panel of wealth data available. This study focusses on the backward imputation of full waves of wealth data. Wealth is correlated over the years. The size of household wealth in one year is strongly related to the size of household wealth in the next year. We will use this relation to predict backwards. Since the same respondents are followed over time, household specific effects can be estimated as well. Moreover, we have the IPO Income panel available for the years in which the IPO Wealth panel was not yet developed. As IPO Income contains information on wealth variables, this subpanel will be used in the prediction of household wealth as well. In addition, it is used as a benchmark to validate the quality of the imputations.

The assets which are reported in IPO Wealth are checking and savings accounts, risky assets (shares and bonds), the value of the primary residence, business equity and the value of other physical assets. Liabilities are divided into mortgage debt and other debt assets. Both are provided in IPO Income and used directly in IPO Wealth. In 2005-2010, the sum of checking and savings accounts, risky assets and the value of the primary residence make up 89.3% of the total value of the assets on average (source: Statline Statistics Netherlands). Business

equity and other assets are just a small proportion of the total value of household assets. Moreover, they are not closely related to any variable in IPO Income. Municipal authorities determine the WOZ ("Waardering Onroerende Zaken" or Valuation of Immovable property) values of residences. These are used by the Dutch Tax Authorities to calculate a so-called notional rental value on the house. The notional rental value is contained in IPO Income. We will focus in this study on the imputation of checking and savings accounts and risky assets.

We define gross financial household wealth as the sum of checking and savings accounts and the risky assets. Other debt assets consist of all debt assets except mortgage debt. We define net financial household wealth as gross financial household wealth minus other debt assets. Households with a taxable income in box 3 are obliged to report the value of other debt assets on tax records. This value is included in IPO Income. However, the value of other debt assets is unknown for households which are not box 3 liable. This means that the value of other debt assets is underreported. No proper imputation methods to correct for this have been developed yet (CBS, 2010b). Therefore, we will not consider net financial household wealth.

In 2001, the tax system in the Netherlands was reformed. This influenced the way wealth was taxed and IPO Income was composed. Before 2001, the wealth variables in IPO Income were different. We cannot infer the relation between these different variables and IPO Wealth. Hence, we will impute the wealth variables for the years 2001-2004.

In the past, there have been earlier attempts to predict wealth data based on income streams. The accuracy of these wealth estimates is questionable. Greenwood (1973) and Wolff (1983) predict household wealth based on tax records in the United States. They have data available on dividends and interest returns. They calculated average yields for each asset class. This yield was then used to capitalize the returns into asset values directly. Real values of the household wealth were not available, so they compared it with national balance sheet figures. Both of their methods produced lower estimates than reported in national balances sheets. This is a very simple method. It does not involve any statistical models or household specific effects. We could not find any literature on a more advanced imputation method regarding full waves of a panel.

In the remainder of this paper, we will discuss our imputation method. Section 2 provides a description of our data. In Section 3, we present the mathematical models which are used in the imputation method. We will carry out a within sample prediction of the IPO Wealth panel in 2005 without using this wave. This way, we can compare the imputations with the actual values. The 2005 imputation is illustrated in Section 4. The results of the full wave imputation for the years 2001-2004 will be discussed in Section 5. Section 6 concludes the paper.

2 Data

The data are taken from the panel survey IPO (Inkomens Panel Onderzoek). IPO consists of two subpanels, IPO Income (IPO Inkomen) and IPO Wealth (IPO Vermogen). IPO Income is developed on a yearly basis since 1989. In this study, we use the 2001-2010 sample of IPO Income. The IPO Wealth panel consists of the same respondents as IPO Income, however this subpanel is generated from 2005 onwards. We will use the 2005-2010 sample of IPO Wealth. The IPO panel is gathered by Statistics Netherlands. The Dutch Tax and

Customs Administration provides data on tax records to Statistics Netherlands. Banks and financial institutions provide yearly records of interest incomes and dividend payments and the corresponding account balances. Furthermore, information on the WOZ value ("Waardering Onroerende Zaken" or Valuation of Immovable property) of residences is provided. This information allows Statistics Netherlands to develop the two panels of IPO.

IPO is based on an administrative random sample of approximately 90,000 'key persons', supplemented by members of their households. This gives an approximate total sample size of 250,000 persons per year. The same key persons are followed over time, although the household composition can change. The number of key persons has increased in the past ten years. In Table 1, the sample size is displayed per year.

Table 1: Sample size
Source: IPO Income

Year	Number of households
2001	83941
2002	85094
2003	86437
2004	87700
2005	88618
2006	89810
2007	90704
2008	91938
2009	93863
2010	97877

Every year, the panel is cleansed such that only key persons living in the Netherlands are considered. Attrition is only due to death and emigration. This gives a very low attrition rate. New key persons are added from immigrants and newborns every year. The IPO panel does not depend on (voluntary) participation rates, which is a great advantage over surveys. Moreover, single person households and the elderly population have typically low participation rates in surveys (Knoef and De Vos, 2008). This effect is not present in the administrative IPO data. Including the elderly population is especially relevant when studying household wealth. Many households have accumulated substantial wealth at retirement and they keep considerable wealth holdings throughout old-age (Poterba, Venti, and Wise, 2012). A disadvantage of the IPO panel is that it does not contain relevant background variables such as education levels and health status. Individual characteristics (for example age, gender and marital status) and household characteristics (e.g. household composition and home ownership) are included in the IPO panel.

All financial variables considered in this study are measured in 2005 euros. We corrected the other years for inflation and currency (the 2001 sample was measured in the Dutch guilder). We used the yearly inflation rates as measured by consumer prices. These rates are taken from Statline, the online database of Statistics Netherlands.

2.1 Income Panel Survey (IPO Income)

The IPO Income is a panel dataset which gives a yearly overview of the composition and distribution of income of persons and households in the Netherlands (CBS, 2010a). The panel considers income data from 1 January until 31 December of a research year. The IPO Income is primarily based on data from the Dutch Tax and Customs Administration. In Appendix B, the Dutch Tax system is explained. In this study, we are interested in predicting wealth variables. We will introduce which variables are related to wealth in IPO Income in year t and their source. In Appendix C, the distributions of some of these variables are depicted.

- $bankteg_t$
  This is the value of checking and savings accounts as provided on tax records. This component is taxed in box 3, so it is only known for households with taxable income in box 3. We move any negative values to the variable other debt ($debt\_other_t$), since debit balances in bank accounts belong to $debt\_other_t$ by definition. This gives $bankteg_t \geq 0$. In Table 19, the distribution is displayed. It is highly skewed to the right, as the majority of the households is not box 3 liable.

- $risky\_assets_t$
  This is the value of bonds and shares that are not part of a substantial interest. According to the Dutch Tax and Customs Administration, a household has substantial interest if it owns at least 5% of one of the following items:

  - shares in a Dutch or or foreign company,
  - the profit-sharing certificates of a Dutch or foreign company,
  - the rights of enjoyment (also per class) of the profit-sharing certificates or shares in a Dutch or foreign company,
  - the voting rights in a cooperative or association organised on a cooperative basis.

  Furthermore, a household has substantial interest when a member owns options to to acquire at least 5% of the shares (also per class) in a Dutch or foreign company (Belastingdienst, 2012). In the remainder of the paper, we will not consider shares of substantial interest. When the term shares is mentioned, it refers to all shares that are not part of a substantial interest, unless explicitly stated otherwise.

  The variable $risky\_assets_t$ is the value as it is provided on tax records. This is only known for households which are box 3 liable. Its distribution is tabulated in Table 20.

- $taxinc3_t$
  This is the taxable income in box 3. On the basis of this variable, we classify whether a household is box 3 liable. We identify two groups, group 1 consists of all households for which $taxinc3_t > 0$ and group 2 consists of all households with $taxinc3_t = 0$. In Table 21 in Appendix D, the sample size per group per year is displayed. Next to sample size, we are interested in the transition probabilities between boxes. These are displayed in detail in Table 22 in Appendix D. The groups are fairly stable. The average yearly transition rates are displayed in Table 2.

Table 2: Average yearly transitions between groups (conditional on households being present in the sample in year $t+1$)
Source: IPO Income

	In group 1 in year $t+1$	In group 2 in year $t+1$
Households in group 1 in year t	89.6%	10.4%
Households in group 2 in year t	3.8%	96.2 %

This can be seen as a Markov system with transition matrix

$$P = \begin{pmatrix} 0.896 & 0.104 \\ 0.038 & 0.962 \end{pmatrix}. \quad (1)$$

The steady state distribution is the vector v for which $vP = v$ (Ross, 2009). It is given by

$$v = \begin{pmatrix} 0.268 & 0.732 \end{pmatrix}. \quad (2)$$

When comparing the fraction per group per year to the steady state distribution, we see that the fractions are quite close to the steady state distribution. The fraction of households which are box 3 liable is a bit smaller than in the steady state distribution.

- $interest_t$
  This denotes the interest income which is collected from checking and savings accounts. This variable is provided by banks and financial institutions. It is provided for all households independent of whether they are box 3 liable. When the interest income is less than 15 euro, it is not always provided by banks and financial institutions. Instead, a zero on this variable is observed. There is no information on the frequency of these small values of interest income. However, in 2009 and 2010, the information becomes more accurate and more smaller values are reported.

- $dividend\_othshares_t$
  This denotes the dividend return on shares that are not part of a substantial interest. Banks and financial institutions provide these values to Statistics Netherlands.

- $dividend\_substshares_t$
  This is the dividend return on shares of a substantial interest. The values are supplied by banks and financial institutions.

- $interest\_bonds_t$
  This is the interest income households received over the bonds they held. Again, this variable is issued by banks and financial institutions.

- $debt\_mortgage_t$
  This is the outstanding mortgage debt of the primary residence of a household. It is the amount of mortgage debt on which interest needs to be paid. Tax records are the source of this variable. In the Netherlands, many people have an endowment mortgage ("spaarhypotheek"). They pay interest on the mortgage while they do not pay off the mortgage debt. Instead, they pay an endowment premium. There is no information on this endowment in

IPO Income. For these households, mortgage debt is constant until the end of the endowment. At the maturity date, mortgage debt suddenly drops to zero. Since the endowments are unobserved, mortgage debt is effectively overreported.

- $debt\_other_t$
  This is the sum of all debts except for mortgage debt of the primary residence. It can include debit balances in bank accounts, debts incurred for consumer purchases or debts for financing a second residence. This variable is part of box 3 on a tax form, so it is only known for group 1. For households in group 2, the value of $debt\_other_t$ is unknown. Hence, this variable underreports the actual value of outstanding debt.

As IPO Wealth is measured on a household level, all relevant financial variables in IPO Income are aggregated to household level as well.

2.2 Income Panel Survey - Wealth (IPO Wealth)

The IPO Wealth panel provides an overview of the composition and distribution of household wealth (CBS, 2010b). Again, the measurement of wealth through administrative data has several advantages over other surveys. The value of assets and liabilities can be hard to report accurately. For example, the current market value of some of the assets may be unknown or respondents may forget to report some of their assets or debts. Moreover, wealth holdings are rather concentrated. An ordinary survey among random households is not likely to contain enough wealthy households to provide a correct representation of the distribution of household wealth. Moreover, the very wealthy are often reluctant to provide information about their wealth (Fries, Starr-McCluer, and Sundén, 1998).

Table 3 represents the asset and debt items which are observed in IPO Wealth. The values in the panel correspond to the market value of the items on 31 December of a research year. We will elaborate on some of the variables and describe their source.

- $balsav_t$
  This is the value of all Dutch checking and savings accounts held by a household. Banks and financial institutions provide these values, independent of whether a household has a taxable income in box 3. They are not obliged to report balances less than 500 euro. This implies that a value of zero can be observed on this variable, but the household actually owns an account worth less than 500 euro. From 2009 onwards, more small accounts are present in the panel. Households which previously had a zero as value, have a value less than 500 euro in 2009 and 2010. It may thus seem like the ownership of checking and savings accounts is increasing, but it is likely that those households had a small value on their bank accounts in previous years. This should be taken into acount when analyzing results.

  Similar to $bankteg_t$ in IPO Income, we move any negative values in $balsav_t$ to the item $Other\ debt_t$ in IPO Wealth. We will be imputing the values of $balsav_t$ for the years 2001-2004.

- $balshabon_t$
  We introduce the variable $balshabon_t$ as the sum of bonds and other shares, or the risky assets. Banks and financial institutions supply the total values of the sum of bonds and

Table 3: Items observed in IPO Wealth (CBS, 2010b)

Wealth

Assets

- Financial assets
  - Checking and savings accounts
  - Bonds
  - Shares
    - Shares, substantial interest
    - Shares, other
- Real estate
  - Primary residence
  - Other real estate
- Movable property ("Roerende zaken")
- Business equity

Debts

- Mortgage debt primary residence
- Other debt

other shares to Statistics Netherlands. Statistics Netherlands then divides this sum into separate values of bonds and other shares based on data on dividends and interest from bonds. Since the precise values of shares and bonds are unknown, we consider them together. Furthermore, in IPO Income, these two variables are not observed separately either. In this study, we will impute the value of $balshabon_t$ for the years 2001-2004.

Over the years 2005-2010, there are 22 negative values of $balshabon_t$. This is 0.004% of the sample. We believe that this is due to measurement errors and we set these values to missing.

- *Shares of a substantial interest*$_t$
  In Section 2.1, the definition of shares of a substantial interest as supplied by the Dutch Tax and Customs Administration is stated. The value of this variable is based on the taxable income in box 2 and the interest on shares of a substantial interest. Both are supplied in IPO Income. The tax rates on income from substantial interest were not equal in 2001-2010, they were lower in 2007. This caused a peak in income from substantial holdings in 2007. Many of the shareholders delayed their dividend returns until 2007. Due to this irregularity, Statistics Netherlands developed their own imputation method to arrive at the values of shares of a substantial interest (CBS, 2010b). The actual values of this variable are thus unknown. This is the reason we are not using any of the data on shares of a substantial interest and we will not impute this variable. In the definition of gross financial household wealth, we exclude shares of a substantial interest.

- *Mortgage debt primary residence*$_t$

This variable is directly taken from IPO Income. IPO Wealth does not contain any additional information on life insurances either.

- *Other debt*$_t$
  The values of this variable is provided in IPO Income. There is no new information about this variable in IPO Wealth, so it is only known for households which are box 3 liable.

2.3 Financial assets in IPO

In this study, we focus on the imputation of the value of checking and savings accounts and risky assets. The variables $balsav_t$ and $balshabon_t$ are provided by banks and financial institutions. The variables $bankteg_t$ and $risky\_assets_t$ are the submitted values on tax records. There can be a difference between the values in IPO Wealth and IPO Income. It is possible that a household is in group 2 and $bankteg_t$ and $risky\_assets_t$ are not provided. Furthermore, people can report a different value on tax records than the actual value. We believe that in case of a difference, $balsav_t$ and $balshabon_t$ are the true values. Let $bankteg_{it}$ and $risky\_assets_{it}$ be the values for household i in year t of the variables in IPO Income. Similarly, let $balsav_{it}$ and $balshabon_{it}$ be the values for household i in year t in IPO Wealth. Let $y^*_{it} = balsav_{it} - bankteg_{it}$ denote the difference between $balsav_{it}$ and $bankteg_{it}$. We define the dummy

$$dum\_sav_{it} = \begin{cases} 1 & \text{if } y^*_{it} = 0 \\ 0 & \text{if } y^*_{it} \neq 0. \end{cases} \tag{3}$$

Likewise, let $z^*_{it} = balshabon_{it} - risky\_assets_{it}$ denote the difference for the value of risky assets. We define the dummy

$$dum\_shabon_{it} = \begin{cases} 1 & \text{if } z^*_{it} = 0 \\ 0 & \text{if } z^*_{it} \neq 0. \end{cases} \tag{4}$$

For 2005, we tabulate the distributions of y^*_{it} and z^*_{it} in Table 4. We distinguish between the cases where the variables in IPO Income equal zero and where they are known. For most households, the difference between the variables in IPO Wealth and IPO Income is positive. We expect a positive difference in the case where the variable in IPO Income equals zero. The value is unknown in IPO Income and IPO Wealth then provides the known value. However, we see that the difference is positive as well in the upper percentiles when there is a value available in IPO Income. This means that households underreport the value of their assets. In Table 5, we tabulated the number of observations for which the variables in IPO Income are larger than, equal to or smaller than the corresponding variable in IPO Wealth for the entire sample $t = \{2005, \ldots, 2010\}$. Especially in the case when a value for checking and savings accounts is reported in IPO Income, the number of positive differences is very high.

Table 4: Distribution of y^*_{it} and z^*_{it} in 2005

Distribution of y^*_{it}			Distribution of z^*_{it}		
Percentiles	$bankteg_t = 0$	$bankteg_t > 0$	Percentiles	$risky\_assets_t = 0$	$risky\_assets_t > 0$
10%	0	0	10%	0	0
25%	1114	0	25%	0	0
50%	5832	3863	50%	0	0
75%	16607	17633	75%	0	6330.5
90%	29745.67	40953	90%	3783	29999.32
Mean	11233.93	11882.62	Mean	2401.10	9907.77
Obs	63471	25147	Obs	75194	13424

It is surprising how many inequalities there are between IPO Income and IPO Wealth, even when $bankteg_t$ and $risky\_assets_t$ are known. For the variable checking and savings accounts, the trend that more small bank accounts are reported in 2009 and 2010 is visible. Where the small bank accounts previously had a zero on $balsav_t$, they have a positive value in 2009 and 2010. This causes that the number of observations with $y^*_{it} = 0$ is decreasing and the number of households with $y^*_{it} > 0$ and $bankteg_t = 0$ increases. There is no obvious reason why the number of inequalities where $y^*_{it} < 0$ and $z^*_{it} < 0$ increases dramatically in 2009 and 2010.

Table 5: Differences between IPO Income and IPO Wealth

(a) Checking and savings accounts

Year	$bankteg_t = 0$			$bankteg_t > 0$		
	Obs $y^*_{it} < 0$	Obs $y^*_{it} = 0$	Obs $y^*_{it} > 0$	Obs $y^*_{it} < 0$	Obs $y^*_{it} = 0$	Obs $y^*_{it} > 0$
2005	0	10355	53116	223	7528	17396
2006	0	10464	53034	136	7945	18231
2007	0	9591	53406	41	8296	19370
2008	0	9245	54554	39	8589	19511
2009	0	5483	60280	2221	6095	19784
2010	0	5099	62805	10182	7478	12313

(b) Risky Assets

Year	$risky\_assets_t = 0$			$risky\_assets_t > 0$		
	Obs $z^*_{it} < 0$	Obs $z^*_{it} = 0$	Obs $z^*_{it} > 0$	Obs $z^*_{it} < 0$	Obs $z^*_{it} = 0$	Obs $z^*_{it} > 0$
2005	0	63032	12162	639	7849	4936
2006	0	64741	11543	576	7748	5202
2007	0	66478	10809	510	7901	5006
2008	0	68669	10506	411	7527	4825
2009	0	69968	11361	1235	5307	5992
2010	0	74623	10771	5027	4955	2501

3 Econometric models

In this section, we will introduce the models which are used as building blocks in our imputation method. In the imputation of checking and savings accounts, the same generic models are used as in the imputation of risky assets. We focus here mainly on describing the prediction of checking and savings accounts. The method is a regression based imputation (see Appendix A for an overview of different imputation methods).

There are two possibilities, either $balsav_{it} = bankteg_{it}$ or $balsav_{it} \neq bankteg_{it}$. When the second equality holds, we need to predict $balsav_{it}$. In order to model this process, we use a two part model (Cameron and Trivedi, 2005). This allows us to model the existence of a difference separately from the mechanism which establishes the amount of a difference. The important assumption in a two part model is that the mechanism which determines whether there is a difference is independent from the mechanism which generates the amount of the difference.

In the first part of the model, we predict whether there is a difference between the variable in IPO Income and the wealth variable. We allow for an unobserved household effect which partly accounts for the difference. Respondents who are very accurate will provide the actual value on their tax records, whereas sloppy respondents might not know the actual value. Furthermore, households in group 2 do not need to report checking and savings accounts, but they may have a positive balance less than the tax-free allowance ('heffingvrij vermogen'). We can exploit our panel data and use a random effects probit model (see Section 3.1) to predict whether there is a difference using a predicted household effect.

When it is predicted that there is a difference, we predict the value of the wealth variable $balsav_{it}$ in the second part of the model. We assume that the balance on checking and savings accounts (or the value of the risky assets) of a household is partly caused by individual effects. The level of risk aversion, for example, will determine whether a household invests more in checking and savings accounts or in risky assets. Again, we will exploit our panel data and we will use a fixed effects model (see Section 3.2) for this purpose. We will use $balsav_t$ and not the difference $balsav_t - bankteg_t$ as a dependent variable. In Table 4, we see that the distribution of the difference is highly skewed, so we want to take the natural logarithm. However, the difference is negative for a few households. Therefore, we will use $balsav_t$ as dependent variable and include $bankteg_t$ as a regressor.

In this section, we will describe the mathematical models. Section 4 then discusses the application of these models in the imputation procedure. Since we are predicting the values of $balsav_t$ and $balshabon_t$ in $t \in \{2001, \ldots, 2004\}$, we end this section with reviewing some of the theoretical concepts behind prediction as described in Hayashi (2000).

3.1 Two Part Model Part 1: Random effects probit

A random effects model assumes that the dependent variable y_{it}^* depends on an individual specific effect c_i, as well as regressors $\boldsymbol{x}_{it}$ and $\boldsymbol{\beta}$ (Wooldridge, 2010). The model assumes that the individual specific effect c_i is a random variable with a specified distribution. In Section 4, we will discuss this model in the context of our imputation and define which regressors are used.

The latent model for the random effects probit model is given by

$$y_{it}^* = \boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i + u_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, T_i. \tag{5}$$

The observation rule in our imputation method is

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* = 0 \\ 0 & \text{if } y_{it}^* \neq 0. \end{cases} \tag{6}$$

Let $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}', \ldots, \boldsymbol{x}_{iT}')'$ denote the stacked vector of regressors. The errors terms are assumed to be serially uncorrrelated and to follow a standard normal distribution,

$$u_{it}|\boldsymbol{x}_i, c_i \sim \mathcal{N}ID(0, 1). \tag{7}$$

The main assumption of the random effects probit model is

$$P(y_{it} = 1|\boldsymbol{x}_{it}, c_i) = \Phi(\boldsymbol{x}_{it}'\boldsymbol{\beta} + c_i), \tag{8}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of a standard normal distribution.

The random effects probit model assumes that the individual specific effects are random independent drawings from a normal distribution,

$$c_i|\boldsymbol{x}_i \sim \mathcal{N}(0, \sigma_c^2). \tag{9}$$

Furthermore, Assumption (9) implies that $\mathrm{E}(c_i u_{it}) = 0$. Let $\epsilon_{it}^A = c_i + u_{it}$ be the total unobserved error term of the random effect probit regression. We have $\mathrm{E}(\epsilon_{it}^A) = 0$. Then

$$\mathrm{var}(\epsilon_{it}^A) = \sigma_c^2 + \sigma_u^2 = 1 + \sigma_c^2, \tag{10}$$

and

$$\mathrm{cov}(\epsilon_{it}^A, \epsilon_{i,t+1}^A) = \sigma_c^2. \tag{11}$$

3.2 Two Part Model Part 2: Fixed effects linear regression

A fixed effects model assumes that the dependent variable s_{it} depends on an unobserved household specific effect α_i, regressors $\boldsymbol{z}_{it}$ and parameter vector $\boldsymbol{\theta}$. The right hand side variables $\boldsymbol{z}_{it}$ of this regression in the imputation procedure will be discussed in Section 4. The fixed effects model is given by

$$s_{it} = \boldsymbol{z}_{it}'\boldsymbol{\theta} + \alpha_i + v_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, T_i. \tag{12}$$

The model allows for correlation between α_i and $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}', \ldots, \boldsymbol{z}_{iT_i}')'$,

$$\mathrm{E}(\alpha_i|\boldsymbol{z}_i) \neq 0. \tag{13}$$

We assume that the errors v_{it} are serially uncorrelated and we allow for heteroskedasticity,

$$v_{it}|\boldsymbol{z}_i \sim \mathcal{N}ID(0, \sigma_{it}^2). \tag{14}$$

Let $\epsilon_{it}^B = \alpha_i + v_{it}$ be the total error term of (12). We have that $\mathrm{E}(\epsilon_{it}^B) = 0$. The fixed effects regression is the second part of our two part model. Moreover, we assume that

$$\mathrm{E}\left(\epsilon_{it}^A \epsilon_{it}^B\right) = 0, \tag{15}$$

i.e. the error terms of the first and second part of the models are mutually uncorrelated. We have assumed that ϵ_{it}^A and ϵ_{it}^B follow a normal distribution with mean zero. This implies that the error terms of the first part model are independent of the error terms of the second part model. This is the important assumption of the two part model.

The parameter vector $\boldsymbol{\theta}$ is estimated by the within estimation procedure. First, (12) is averaged over t. This gives

$$\bar{s}_i = \bar{\boldsymbol{z}}_i'\boldsymbol{\theta} + \alpha_i + \bar{v}_i, \quad i = 1, \ldots, N, \tag{16}$$

where

$$\bar{s}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} s_{it}, \tag{17}$$

$$\bar{\boldsymbol{z}}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} \boldsymbol{z}_{it}, \tag{18}$$

$$\bar{v}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} v_{it}. \tag{19}$$

Next, (16) is subtracted from (12),

$$\ddot{s}_{it} = \ddot{\boldsymbol{z}}_{it}'\boldsymbol{\theta} + \ddot{v}_{it}, \tag{20}$$

where

$$\ddot{s}_{it} = s_{it} - \bar{s}_i, \tag{21}$$

$$\ddot{\boldsymbol{z}}_{it} = \boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i, \tag{22}$$

$$\ddot{v}_{it} = v_{it} - \bar{v}_i. \tag{23}$$

The within estimator $\hat{\boldsymbol{\theta}}$ is obtained by applying OLS on (20). The individual specific effect can be estimated through

$$\hat{\alpha}_i = \bar{s}_i - \bar{\boldsymbol{z}}_i'\hat{\boldsymbol{\theta}}. \tag{24}$$

When we would not allow for heteroskedasticity, we would have

$$\mathrm{Var}(v_{it}) = \sigma_v^2, \tag{25}$$

for all $i = 1, \ldots, N$ and $t = 1, \ldots, T_i$. Then,

$$\begin{aligned}
\mathrm{E}(\ddot{v}_{it}^2) &= \mathrm{E}\left((v_{it} - \bar{v}_i)^2\right) && (26)\\
&= \mathrm{E}(v_{it}^2) + \mathrm{E}(\bar{v}_i^2) - 2\mathrm{E}(v_{it}\bar{v}_i) && (27)\\
&= \sigma_v^2 + \frac{1}{T_i^2} T_i \sigma_v^2 - \frac{2}{T_i}\sigma_v^2 && (28)\\
&= \sigma_v^2\left(1 - \frac{1}{T_i}\right) && (29)\\
&= \mathrm{E}(\ddot{v}_{is}^2), && (30)
\end{aligned}$$

and, for $t \neq s$,

$$\begin{aligned}
\mathrm{E}(\ddot{v}_{it}\ddot{v}_{is}) &= \mathrm{E}\left[(v_{it} - \bar{v}_i)(v_{is} - \bar{v}_i)\right], \qquad t \neq s, && (31)\\
&= E(v_{it}v_{is}) - \mathrm{E}(v_{it}\bar{v}_i) - \mathrm{E}(v_{is}\bar{v}_i) + \mathrm{E}(\bar{v}_i^2) && (32)\\
&= 0 - \frac{1}{T_i}\sigma_v^2 - \frac{1}{T_i}\sigma_v^2 + \frac{1}{T_i}\sigma_v^2 && (33)\\
&= -\frac{1}{T_i}\sigma_v^2. && (34)
\end{aligned}$$

This gives

$$\begin{aligned}
\mathrm{cor}(\ddot{v}_{it}, \ddot{v}_{is}) &= \frac{-\frac{1}{T_i}\sigma_v^2}{\sqrt{\sigma_v^2\left(1 - \frac{1}{T_i}\right)}\sqrt{\sigma_v^2\left(1 - \frac{1}{T_i}\right)}} && (35)\\
&= \frac{-\frac{1}{T_i}\sigma_v^2}{\sigma_v^2\left(1 - \frac{1}{T_i}\right)} && (36)\\
&= -\frac{1}{T_i - 1}, && (37)
\end{aligned}$$

for $t \neq s$, $t, s = 1, \ldots T_i$.

3.3 The stochastic error component

So far, we have discussed models which can be used to linearly predict $balsav_t$ and $balshabon_t$. From Appendix A, it follows that a good imputation method involves a stochastic component as well. This could be implemented by randomly drawing residuals from the empirical distribution. This is consistent when the error terms are assumed to be independently and identically distributed and there is no heteroskedasticity. Yet, in assumption (14), we allow for heteroskedasticity. We will describe how we can model the heteroskedasticity. This approach is not a classical model, it is specified for the imputation of $balsav_t$ and $balshabon_t$. We will describe how we apply this in the context of the $balsav_t$ imputation.

After the fixed effects estimation of (58), there is a set of residuals $\widehat{\ddot{v}}_{it}$. Note that

$$\begin{aligned}
\widehat{\ddot{v}}_{it} &= \ddot{s}_{it}\hat{\boldsymbol{\theta}} - \ddot{\boldsymbol{z}}_{it} && (38)\\
&= s_{it} - \bar{s}_i - (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)'\,\hat{\boldsymbol{\theta}} && (39)\\
&= s_{it} - \boldsymbol{z}_{it}'\hat{\boldsymbol{\theta}} - (\bar{s}_i - \bar{\boldsymbol{z}}_i'\hat{\boldsymbol{\theta}}) && (40)\\
&= s_{it} - \boldsymbol{z}_{it}'\hat{\boldsymbol{\theta}} - \hat{\alpha}_i && (41)\\
&= \hat{v}_{it}, && (42)
\end{aligned}$$

so $\widehat{\ddot{v}}_{it} = \hat{v}_{it}$. Next, we square these residuals and regress them on the regressors $\boldsymbol{z}_{it}$. That is,

$$\hat{v}_{it}^2 = \boldsymbol{z}_{it}'\boldsymbol{\gamma} + e_{it} \qquad (43)$$

is estimated by OLS. By keeping the significant regresssors and discarding the insignificant regressors, the variables which account for the heteroskedasticity remain. Based on these regressors, we identify classes of households.

As an example, the continuous variable $balsav_{i,t+1}$ is considered. We would like to create classes based on a low, medium, high and very high amount of bank and savings accounts in $t+1$ respectively. Let $\boldsymbol{balsav}_{t+1}$ denote the $N \times 1$ stacked vector of observations of $balsav_{i,t+1}$, $i = \{1, \ldots, N\}$. We create the dummy $\delta_{balsav_{i,t+1}=LOW}$ based on whether the value of $balsav_{i,t+1}$ is less than the 25% quantile of $\boldsymbol{balsav}_{t+1}$. The dummy $\delta_{balsav_{i,t+1}=MED}$ equals 1 when the value of $balsav_{i,t+1}$ is greater than or equal to the 25% quantile and smaller than the 50% quantile of $\boldsymbol{balsav}_{t+1}$. The dummy $\delta_{balsav_{i,t+1}=HIGH}$ equals 1 when the value of $balsav_{i,t+1}$ is greater than or equal to the 50% quantile and smaller than the 75% quantile of $\boldsymbol{balsav}_{t+1}$. These dummies will be used in a regression, so the dummy for very high values, those above the 75% quantile of $\boldsymbol{balsav}_{t+1}$, need not be created. This way, the problem of perfect collinearity will be avoided.

A similar categorization can be done for all continuous regressors. Let the vector of dummy variables for household i, year t be denoted by $\boldsymbol{d}_{it}$. The regression

$$\hat{v}_{it}^2 = \boldsymbol{d}_{it}'\boldsymbol{\zeta} + w_{it} \tag{44}$$

is estimated by OLS.

3.4 Prediction

We will predict the values of $balsav_t$ and $balshabon_t$ for $t = \{2001, \ldots, 2004\}$. Therefore, we devote this subsection to some of the theoretical concepts behind prediction. It is mainly based on Hayashi (2000).

We consider a random variable y and a random vector $\boldsymbol{x}$. The joint distribution of $(y, \boldsymbol{x})$ and the value of $\boldsymbol{x}$ are first assumed to be known. We want to predict y based on this information. A predictor is defined as a function $f(\cdot)$ of $\boldsymbol{x}$, determined by the joint distribution of $(y, \boldsymbol{x})$. The forecast error is given by

$$y - f(\boldsymbol{x}), \tag{45}$$

and the mean squared error is defined as

$$\mathrm{E}\left((y - f(\boldsymbol{x}))^2\right). \tag{46}$$

Proposition 2.7 of Hayashi (2000) states that the best predictor of a random variable y, given observed vector $\boldsymbol{x}$, is $\mathrm{E}(y|\boldsymbol{x})$. This minimizes the mean squared error.

In order to calculate $\mathrm{E}(y|\boldsymbol{x})$, the joint distribution of $(y, \boldsymbol{x})$ should be known. When the predictor is restricted to be linear, the least squares projection of y on $\boldsymbol{x}$, denoted by $\widehat{\mathrm{E}}^*(y|'vx)$, is considered. It is defined by

$$\widehat{\mathrm{E}}^*(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}^*, \tag{47}$$

where $\boldsymbol{\beta}^*$ is defined such that

$$\mathrm{E}(\boldsymbol{x}\boldsymbol{x}')\boldsymbol{\beta}^* = \mathrm{E}(\boldsymbol{x}'y) \tag{48}$$

is satisfied. Proposition 2.8 of Hayashi (2000) states that the least squares projection $\widehat{\mathrm{E}}^*(y|\boldsymbol{x})$ is the best linear predictor of y in that the mean squared error is minimized. The joint distribution of $(y, \boldsymbol{x})$ need not be known, instead only the second moments of the joint distribution

of $(y, \boldsymbol{x})$ are needed.

Next, we assume that there is a random sample available and $\mathrm{E}(\boldsymbol{x}\boldsymbol{x}')$ is nonsingular. Hayashi (2000) states that, under these assumptions, the OLS estimator is always consistent for the projection coefficient vector $\boldsymbol{\beta}^*$ that satisfies the orthogonality condition (48). For example, consider the case where we would like to include lead dependent variables in a regression. Then, exogeneity assumptions on the regressors are not satisfied. Hayashi (2000) ensures that OLS provides a consistent estimator to optimally linearly predict the dependent variable as long as there is a random sample.

4 Imputation of Checking and Savings accounts for 2005

In 2005, the values of both IPO Income as well as IPO Wealth are known. It is possible to impute the items in IPO Wealth for $t = 2005$ without considering the actual values. Then, we can compare the imputed values with the actual values. This is known as *within sample prediction* and it gives an indication of the performance of the imputation procedure. In this section, we will explain the imputation procedure for the variable checking and savings accounts. We describe how we apply the models from Section 3. We use the same procedure for the imputation of risky assets. In Section 4.7, these outcomes are briefly discussed.

For households which are box 3 liable (group 1), a value of $bankteg_t$ is provided. In Appendix E, the number of households within each group is schematically displayed together with the relation of $bankteg_t$ to $balsav_t$ for $t = 2005$.

The total value of savings and investments are taxed in box 3 when this value exceeds a threshold. Households in group 1 are therefore wealthier than those in group 2. The distribution of checking and savings accounts is very different in both groups (see Table 8). Due to this large difference in distribution of bank and savings accounts and since the transition rates between groups are stable, we decided to model the account balances separately for both groups.

Although the groups are fairly stable, the composition can change over time. Since one of our interests is fixed effects modelling, we decided to form samples based on the status in the imputation year. That is, when we are imputing bank and savings accounts in 2005, we label which households have a taxable income in box 3 in 2005. These households are selected as the group 1 sample in $t = \{2006, \ldots, 2010\}$, regardless of whether they belong in group 1 in these other years. Similarly, the households which do not have a taxable income in box 3 in 2005 are labelled. These households make up the group 2 sample in $t = \{2006, \ldots, 2010\}$, irrespective of whether they have a taxable income in box 3 in these years. Based on these samples, the probit models are estimated.

4.1 Probit models with random effects

We are interested in predicting whether $balsav_{i,2005}$ will equal $bankteg_{i,2005}$. A random effects probit model is estimated based on the dependent variable dum_sav_{it} (as defined in (3)) in the years $t = \{2006, \ldots, 2010\}$.

The probit models for group 1 and group 2 are estimated separately. The same right hand variables are used in the estimation of both probit models, they are listed in Table 23 in

Appendix F. [1] The lead value of the dependent variable is used in the regression, so our regressors are not strictly exogenous. However, our interest is in predicting probabilities, not in causal inference. Therefore, we apply the principle of linear projections (see Subsection 3.4). The estimated parameters cannot be interpreted in a causal way. However, since our sample is randomly drawn, the predictions are consistently estimated.

The regression results are displayed in Table 24. Based on the z-values, we can identify the variables with the highest explanatory power in the probit regressions. They are

- dum_sav_{t+1}, the value of dum_sav in year $t+1$;
- $dum_{interest_t=0}$, a dummy whether the interest income in year t is zero;
- $dum_{balsav_{t+1}=0}$ a dummy whether $balsav$ is zero in year $t+1$;
- $I_{contrsav_t>0} * \ln(contrsav_t)$, the logarithm of contractual savings ("spaarloon") in year t;
- $I_{balsav_{t+1}>0} * \ln(balsav_{t+1})$, the value of $balsav$ in year $t+1$.

Estimates $\hat{\boldsymbol{\beta}}^j$ and $\hat{\sigma}^2_{c^j}$ for groups j ($j = \{1, 2\}$) are obtained. It is estimated that

$$\hat{\sigma}^2_{c^1} = 2.01e^{-6} \quad (49)$$
$$\hat{\sigma}^2_{c^2} = 9.12e^{-6} \quad (50)$$

for the subsample of households which are box 3 liable ($j = 1$) and for the households without a taxable box 3 income ($j = 2$). These values are very small, so we test the hypothesis

$$H_0 = \sigma^2_{c^j} = 0. \quad (51)$$

In this test, the pooled probit estimator is compared with the random effects panel estimator by means of a likelihood-ratio test. The p-values are 0.482 and 0.478 for groups 1 and 2 respectively. There is no evidence to reject H_0. A possible explanation for this is that $balsav_{t+1}$ and $balsav_{t+2}$ are included in the regressors. This can already correct for a household specific effect. Since the random effect is not significant in 2005, we will not include any random effect in our imputation. For the imputation of the other years, a probit model with random effects will be estimated. If the random effect is significant, we include it in our prediction. A household effect $\hat{c}_i$ is then randomly drawn from the normal distribution $\mathcal{N}(0, \hat{\sigma}^2_{c^j})$. In the imputation of bank and savings accounts in 2001-2004, the random effect is never significant. For risky assets, it is sometimes significant, but it is still small. It is around 0.4 in most years. The probability for a household in group j is estimated through

$$P(balsav_{it} = bankteg_{it}|\boldsymbol{x}_{it}) = \Phi(\boldsymbol{x}_{it}\hat{\boldsymbol{\beta}}^j + \hat{c}_i), \quad (52)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution and $\hat{c}_i = 0$ when the random effects are not significant.

[1] In the discussion of the imputation results (Section 5), cohort graphs are given. The cohorts are based on the age of the principle income earner of a household. There are strong cohorts effects present for both bal_sav_t and bal_shabon_t. We used the variable 'age of the principle income earner of a household' in the regressions. This was not significant, so we did not include it in the final models.

In the random effects probit model, we assumed

$$u_{it}|\boldsymbol{x}_i, y_{i,t+1}, c_i \quad \sim \quad \mathcal{N}ID(0,1). \tag{53}$$

In estimating the probit model, we also included the regressor y_{t+2}. This variable was not significant. This is evidence that the errors are serially uncorrelated over time. If this variable were significant,

$$\mathrm{E}\left(u_{it}|\boldsymbol{x}_i, y_{i,t+1}, c_i\right) \quad \neq \quad 0, \tag{54}$$

could be true. The autocorrelation should then be taken into account in the prediction. Since y_{t+2} was not significant, the following holds

$$P(y_{it}=1|\boldsymbol{x}_{it}, y_{i,t+1}, y_{i,t+2}) \quad = \quad P(y_{it}=1|\boldsymbol{x}_{it}, y_{i,t+1}). \tag{55}$$

In other words, conditioning on the lead value y_{t+2} does not provide additional information.

4.2 Monte Carlo simulation

In order to determine whether or not the value of $balsav_{it}$ needs to be imputed or is set equal to $bankteg_{it}$, Monte Carlo simulations are applied (Kelton and Law, 2007). A random value from the uniform distribution is drawn for every household i in each year t. We denote this value by U_{it}. Suppose we have estimated that, for household i, $P(dum\_sav_{it}=1|\boldsymbol{x}_{it}, \hat{c}_i)=\lambda$, for some $\lambda \in [0,1]$. Since $P(U_{it} \leq \lambda)=\lambda$, we can simulate dum_sav_{it} using U_{it}. When the random draw U_{it} is smaller than the predicted probability, $balsav_{it}$ is assigned to equal $bankteg_{it}$. When the random draw U_{it} is higher than the predicted probability, $balsav_{it}$ will be predicted.

In Table 6, a cross-tabulation of dum_sav_{it} is shown for the imputation in $t=2005$. In 5421 households, it is predicted that $balsav_t = bankteg_t$ but there is a difference in the actual sample. In 5973 households, $balsav_t = bankteg_t$ is true in the actual sample but a difference is predicted. For 75530 households, which is 86.9% of the sample, the correct decision is made. Verbeek (2004) describes a goodness-of-fit measure based on a cross table like Table 6. He

Table 6: Cross-tabulation of predicted and actual outcome in differences checking and savings accounts ($t=2005$)

		$\widehat{dum\_sav}_t$		
		0	1	Total
dum_sav_t	0	64242	5421	69663
	1	5973	11288	17261
	Total	70215	16709	86924

compares the proportion of incorrect predictions,

$$wr_1 \quad = \quad \frac{5421+5973}{86924} = 0.131, \tag{56}$$

with the proportion of incorrect predictions based on a model with an intercept only. The proportion of ones in our sample is $\frac{17261}{86924}=0.199$. This is less than 50%, so a zero would

be predicted for each observation in the intercept-only model. The proportion of incorrect predictions is therefore $wr_0 = 0.199$. The goodness-of-fit measure is obtained as

$$R_p^2 = 1 - \frac{wr_1}{wr_0} = 1 - \frac{0.131}{0.199} = 0.339, \tag{57}$$

which is not very high. The correlation between dum_sav_{2005} and $\widehat{dum\_sav}_{2005}$ is 0.5833. In order to compare it with (57), we should square the correlation. This gives 0.340, which is equal to the pseudo R^2 developed by Verbeek (2004).

4.3 Amount regressions

In the imputation of $balsav_{it}$, we would again like to distinguish between households which are box 3 liable and those which are not. For a fraction of the households, $balsav_{it}$ is predicted to equal $bankteg_{it}$. We define group 1^* to consist of the households in $t = \{2006, \ldots, 2010\}$ which are box 3 liable in 2005 and are predicted to have $\widehat{balsav}_{i,2005} \neq bankteg_{i,2005}$. Similarly, subsample group 2^* is defined to consist of the households in $t = \{2006, \ldots, 2010\}$ which are not box 3 liable in 2005 and are estimated to have $\widehat{balsav}_{i,2005} \neq bankteg$. We will estimate two separate fixed effects models based on subsamples group 1^* and group 2^*.

The fixed effects model for bank and savings accounts for a household in group j ($j = \{1^*, 2^*\}$) is

$$s_{it} = \boldsymbol{z}_{it}'\boldsymbol{\theta}^j + \alpha_i + v_{it}, \tag{58}$$

where $s_{it} = \ln(balsav_{it})$ is the log of $balsav_{it}$, $\boldsymbol{z}_{it}$ is the vector of observations for the explanatory variables, $\boldsymbol{\theta}^j$ the parameter vector for group j, α_i is the unobserved household specific effect and v_{it} is the error term. Since $\ln(balsav_t)$ is the dependent variable, our panel can be unbalanced. When $balsav_t = 0$, $\ln(balsav_t)$ is not defined and that year will not be considered in the fixed effects regression.

The right hand side variables are the same for both subsamples, they are outlined in Table 25 in Appendix G. Since checking and savings accounts are measured at the household level, only household characteristics are included in the regression.

We estimate (58) for the two samples. The results are summarized in Table 26. Based on the z-values, we identify which regressors have the highest z-values and are statistically most significant in the prediction. These are the regressors with the highest prediction power:

- $I_{bankteg_t>0} * \ln(bankteg_t)$, the logarithm of *bankteg* (provided in IPO Income) in year t;
- $dum_{bankteg_t=0}$, a dummy whether *bankteg* is zero in year t;
- $I_{interest_t>0} * \ln(interest_t)$, the logarithm of interest income in year t whenever this was positive;
- $hhsize_t$, the household size in year t;
- $I_{grossinc_t>0} * \ln(grossinc_t)$, the logarithm of gross household income in year t.

The R^2 values are displayed in Table 7. There is a large difference between the households which are box 3 liable and have a value for $bankteg_{it}$ in IPO Income (group 1^*) and the households for which $bankteg_{it}$ is not supplied in IPO Income (group 2^*). The R^2 are much higher for group 1^*, the households with a taxable box 3 income. The variable $bankteg_t$ is a strong predictor of $balsav_t$.

Table 7: Values of R^2 for fixed effect regressions

R^2	Group 1^*	Group 2^*
within	0.5824	0.2003
between	0.8691	0.5241
overall	0.8195	0.4623

We have estimates $\hat{\boldsymbol{\theta}}^j$ and $\hat{\alpha}_i$ based on $t = \{2006, \ldots, 2010\}$. Furthermore, the estimation produces residuals $\hat{v}_{it}$, for $t = \{2006, \ldots, 2010\}$.

4.4 Estimation of heteroskedasticity

The estimated residuals $\hat{v}_{it}$, for $t = \{2006, \ldots, 2010\}$, are squared and regressed over regressors $\boldsymbol{z}_{it}$ using clustered standard errors. For group 1^*, we find that the variables $dum_{interest_t=0}$, $interest_t$, $bankteg_t$, $dum_{bankteg_t=0}$, $dum_{bankteg_{t+1}=0}$, $balsav_{t+1}$ and $taxinc3_t$ (the taxable income in box 3) account for heteroskedasticity. For group 2^*, these are $dum_{interest_t=0}$, $interest_t$, $bankteg_t$, $dum_{balsav_{t+1}=0}$, $dum_{bankteg_t=0}$, $balsav_{t+1}$ and $dum_{bankteg_{t+1}=0}$. We create classes for these variables, as described in Section 3.3. There are no interaction effects between the dummies included. Consequently, the model is not saturated. This implies that, theoretically, it is possible that negative outcomes will be predicted. The squared residuals are regressed on the dummies and we delete the insignificant dummies. Let $\boldsymbol{d}_{it}$ denote the vector of dummies for household i, year t. We run the regression

$$\hat{v}_{it}^2 \quad = \quad \boldsymbol{d}_{it}'\boldsymbol{\zeta}^j + w_{it} \tag{59}$$

for $t = 2006, \ldots, 2010$ and groups $j = \{1^*, 2^*\}$ to obtain an estimate $\hat{\boldsymbol{\zeta}}^j$ for the parameter vector. We predict

$$\hat{\sigma}_{it}^2 \quad = \quad \boldsymbol{d}_{it}'\hat{\boldsymbol{\zeta}}^j \tag{60}$$

for households i in groups j. We find that only positive values are produced. These will be used as an estimate of σ_{it}^2, the variance of the residual in (58) of household i in year t.

With the estimated variance of the residuals $\hat{\sigma}_{it}^2$, we can standardize the residuals of the fixed effects regressions. If the normality assumption (14) of the second part model holds, the standardized residuals should (approximately) follow a standard normal distribution. The distributions are displayed in Table 27. For both groups, the distribution is symmetric around zero and standard deviations are equal to 1.00 and 0.99 respectively. The skewness for box 3 liable households is -0.27, while the skewness for households which are not box 3 liable equals -0.62. These skewness values are not equal to zero (which they would have been in case of a standard normal distribution), but they are still small. The kurtosis for the first group is 13.7

and the kurtosis is 11.6 for the second group. The standard normal distribution has a kurtosis of 3, so the residuals have fatter tails than the standard normal distribution. Still, the value of the kurtosis is not extremely large. For these reasons, we are comfortable with assuming normality in the second part model.

4.5 Autocorrelation in the fixed effects regression

We started with the assumption that the error terms are serially uncorrelated over time. In this subsection, we will investigate this assumption. When the residuals would have been homoskedastic, we derived (see Section 3.2)

$$\text{cor}(\ddot{v}_{it}, \ddot{v}_{is}) = -\frac{1}{T_i - 1}. \tag{61}$$

In order to find out whether this holds, we define

$$v_{it} = \rho_i^* v_{i,t+1} + e_{it} \tag{62}$$

$$= \left(-\frac{1}{T_i - 1} + \rho\right) v_{i,t+1} + e_{it} \tag{63}$$

where

$$\rho = \rho_i^* + \frac{1}{T_i - 1}. \tag{64}$$

From this, it follows

$$v_{it} + \frac{1}{T_i - 1} v_{i,t+1} = \rho v_{i,t+1} + e_{it}. \tag{65}$$

When there is no autocorrelation, we have $\rho = 0$. We can test whether this is true based on the obtained residuals. However, we need to correct for the heteroskedasticity first. We standardize the residuals by

$$\frac{\hat{v}_{it}}{\hat{\sigma}_{it}} \tag{66}$$

where $\hat{\sigma}_{it}$ is predicted as in (60). We will again analyze the group of households with box 3 income ($j = 1^*$) separately from the group of households without box 3 income ($j = 2^*$). There is a positive correlation between the right hand side variables of the fixed effects regression and the residuals. We include them too in the regression for group j

$$\frac{\hat{v}_{it}}{\hat{\sigma}_{it}} + \frac{1}{T_i - 1} \frac{\hat{v}_{i,t+1}}{\hat{\sigma}_{i,t+1}} = \rho_j \frac{\hat{v}_{i,t+1}}{\hat{\sigma}_{i,t+1}} + \boldsymbol{z}_{it}' \boldsymbol{\theta}^j + \kappa_{it}. \tag{67}$$

We use clustered standard errors. This produces estimates

$$\hat{\rho}_{1^*} = 0.337 \tag{68}$$

$$\hat{\rho}_{2^*} = 0.140. \tag{69}$$

The null hypothesis that

$$H_0 = \rho_j = 0, \tag{70}$$

was rejected for both groups with p-values of 0.0000 in both groups. This implies that there is autocorrelation in the errors terms and we need to take this into account in the prediction. As, for a household in group j,

$$\begin{aligned} \sigma_{it}^2 &= \text{Var}(v_{it}) = \text{Var}(\rho_j v_{i,t+1} + e_{it}) && (71)\\ &= \rho_j^2 \text{Var}(v_{i,t+1}) + \text{Var}(e_{it}) && (72)\\ &= \rho_j^2 \sigma_{i,t+1}^2 + \text{Var}(e_{it}), && (73) \end{aligned}$$

and since we have estimates $\hat{\sigma}_{it}^2$ and $\hat{\rho}_j$, we can estimate for households in group j

$$\widehat{\text{Var}}(e_{it}) = \hat{\sigma}_{it}^2 - \hat{\rho}_j^2 \hat{\sigma}_{i,t+1}^2. \quad (74)$$

For the households without box 3 income, this produced only positive estimates for $\widehat{\text{Var}}(e_{it})$. In the group of households which are box 3 liable, a small negative value was predicted in 0.5% of the cases in 2005. We decided to set those negative values equal to the median of $\widehat{\text{Var}}(e_t)$ of the subsample $\widehat{\text{Var}}(e_t) > 0$.

For years $t = \{2006, 2007, 2008\}$, we now have a set of residuals $\hat{u}_{it}$, $\hat{u}_{i,t+1}$ and estimates $\hat{\rho}_j$ and $\widehat{\text{Var}}(e_{it})$. From this, we can derive the values of $\hat{e}_{it}$ and standardize them. That is, we calculate for households in group j

$$\frac{\hat{u}_{it} - \hat{\rho}_j \hat{u}_{i,t+1}}{\sqrt{\widehat{\text{Var}}(e_{it})}} \quad (75)$$

for $t = \{2006, 2007, 2008\}$. The resulting distributions for both groups are tabulated in Table 28. The distributions are symmetric around zero and the standard deviations equal 1.12 and 1.02 respectively. This is very similar to the standard normal distribution. The skewness for box 3 liable households is -0.14. For households without taxable box 3 income, the skewness equals -0.80. These values are close to zero. The kurtosis for the first group is 16.9 and it is 12.9 for the second group. The tails are fatter than the tails of a standard normal distribution. However they are not extremely fat-tailed, so we are comfortable with assuming that e_{it} is normally distributed.

We randomly draw an error $\tilde{e}_{it}$ from $\mathcal{N}(0, \widehat{\text{Var}}(e_{it}))$. The stochastic error component which can be added to the linear predictions, for household i in group j, is then predicted by

$$\hat{v}_{i,2005} = \hat{\rho}_j \hat{v}_{i,2006} + \tilde{e}_{i,2005}. \quad (76)$$

For some households, the value of $\hat{v}_{t+1}$ is not known. This occurs when the household is not included in the sample in year $t+1$ or when the value of the $balsav_{t+1}$ or $balshabon_{t+1}$ is zero. Then, the logarithm cannot be taken and that observation is not included in the regression sample. In the imputation of risky assets, the latter happens quite often. Many households do not own any risky assets. Then, the stochastic error component is calculated as if there were no autocorrelation. In these cases, $\hat{v}_{it}$ is a random drawing from $\mathcal{N}(0, \hat{\sigma}_{it}^2)$.

4.6 Comparing imputations with actual values

The final value of the imputation is predicted by

$$\widehat{balsav}_{i,2005} = \exp\left(\boldsymbol{x}'_{i,2005}\hat{\boldsymbol{\beta}}^j + \hat{c}_i + \hat{v}_{i,2005}\right) \quad (77)$$

where the appropriate parameters vector is selected depending on whether the household has a taxable income in box 3 ($j = 1^*$) or not ($j = 2^*$). The exponent is taken since the dependent variable in (58) is the logarithm of $balsav_t$. Together with the households which are assigned $balsav_t = bankteg_t$, we now have completed the imputation for $balsav_t$. The distribution of the imputed variable, together with the distribution of the actual $balsav_t$ is displayed in Table 8. We only considered observations for which an imputation was predicted, so we can compare both distributions.

For the first group, the distribution of the imputed variable is very close to the actual distribution. The predictions are a little smaller than the actual values until the 50% quantile. In the upper quantiles, the predictions are somewhat larger than the actual values. The variation within the imputations is larger than in the actual values. For the second group, the imputation has higher values for nearly all quantiles. The values of the bank and savings accounts are slightly overestimated. Furthermore, the imputed variable has a far higher kurtosis than the original variable.

Table 8: Distributions of $balsav_t$ and imputed $balsav_t$ in 2005

	Variable			
	$balsav_t$ Group 1	Imputed $balsav_t$ Group 1	$balsav_t$ Group 2	Imputed $balsav_t$ Group 2
Percentiles				
5%	8270	7958	0	0
10%	17374	15896	0	0
25%	38033	37666	1360	1779
50%	71242	70891	6723	7557
75%	123481	123692	18596	18494
90%	209176	212596	33983	33580
95%	301247	308788	43734	46212
Mean	106806	107677	12930	13624
Std. Dev.	167189	176459	18967	20531
Variance	2.80e+10	3.11e+10	3.60e+08	4.22e+08
Skewness	14.524	18.052	8.816	16.056
Kurtosis	419.3541	684.797	239.084	1149.542
Observations	19462	19462	67462	67462

Other measures describing the performance of the imputation procedure, are the correlation and Spearman's ρ of the real value $balsav_t$ and the imputed value of $balsav_t$. Since Spearman's ρ uses ranks instead of values, it is more robust to outliers. The correlations are depicted in Table 9. The correlations are quite high.

Table 9: Correlations actual and imputed value of *balsav*

	correlation	Spearman's ρ
Households with taxable income in box 3	0.8991	0.9215
Households without taxable income in box 3	0.6488	0.7798
Total sample	0.9083	0.8696

4.7 Imputation of risky assets

The imputation of the value of stocks and bonds in IPO Wealth, $balshabon_t$, is implemented in a similar fashion as the imputation of $balsav_t$. The value of risky assets as provided in IPO Income is denoted by $risky\_assets_t$. The distribution of $balshabon_t$ is very different for households with taxable income in box 3 compared to households without taxable income in box 3. In 2005, 63.5% of households with a taxable income in box 3 have a positive value of stocks and bonds in IPO Wealth. For the households without a taxable income in box 3, the ownership rate is 18.9%.

First, probit models are estimated for the two groups to estimate whether $balshabon_t = risky\_assets_t$. The results after the Monte Carlo simulations are listed in Appendix H. The results are quite good, the pseudo R^2 as developed by Verbeek (2004) is 0.504. The correlation between the estimated vector of differences and the vector of actual differences is 0.689. When $balshabon_t$ is assigned to be different from $risky\_assets_t$, fixed effects regressions are run. A stochastic error term is added to the linear predictions. We correct for heteroskedasticity and take autocorrelation into account as described in previous sections. The resulting distributions and the original distributions of $balshabon_t$ are depicted in Appendix H. It should be noted that there are a few extreme outliers predicted. These outliers influence the means, variances and correlations strongly. Therefore, trimmed statistics are calculated as well.

For the first group, the distribution of the imputed variable closely follows the actual distribution. The 95 % quantile is overestimated, and both the mean and the trimmed mean are higher for the imputed variable. The standard deviation is also much higher for the imputed variable. The difference in standard deviations is smaller in the distribution of $balshabon_t$ conditional upon (predicted) ownership. For the households which are not box 3 liable, the distribution are again quite close to the actual distributions. Again, the 95% quantile (conditional upon ownership) is overestimated. The standard deviation and mean are lower for the imputed variable than the actual variable. We have calculated correlations with and without the outliers. Furthermore, Spearman's ρ, which is robust to outliers, is calculated as well. The values are displayed in Table 10. The correlations are very high.

Table 10: Correlation $balshabon_t$ and imputed variable

	Correlations	Correlations *without outliers*	Spearman's ρ
Households with taxable income in box 3	0.4984	0.8635	0.9579
Observations	19134	19127	19134
Households without taxable income in box 3	0.0334	0.6305	0.8130
Observations	66109	66108	66109
Total sample	0.1888	0.8644	0.8930
Observations	85243	85235	85243

Only a small percentage of households own risky assets. In Table 11, we cross tabulated the predicted ownership versus the actual ownership. In 94.0 % of the households, the correct ownership status was assigned. However, for 2632 households (3.1 %) ownership was assigned but they did not own any risky assets. The other 2.9% of the households were assigned that they did not own any risky assets, while they did have a positive balance of risky assets in IPO Wealth.

Table 11: Cross tabulation of ownership of risky assets

		Predicted ownership		
		0	1	Total
Actual ownership	0	57638	2632	60270
	1	2496	22477	24973
	Total	60134	25109	85243

4.8 Conclusion

Based on the 2005 imputation, we conclude that the imputation method works quite well. The probit models combined with the Monte Carlo simulations produce good results and the correlations are high. Distributions look similar, especially for the households with taxable income in box 3. We will now continue with the imputation of the IPO Wealth data for 2004 until 2001. The same regressors will be used. We start with the imputation of checking and savings accounts in 2004. The imputation in 2005 showed that the actual values were overestimated. Therefore, we will not use the imputed values in the imputations of other years. This means that in 2003, we use the 2005 values for the right hand side variable $balsav_{t+1}$ instead of the 2004 imputed values. [2] This procedure is mathematically incorrect. For example, in 2003 we estimate models with regressor $balsav_{t+1}$. However, in the linear prediction in 2003, we substitute the value of $balsav_{t+2}$ in the regressor $balsav_{t+1}$. In 2001, we thus use the value of $balsav_{t+4}$ for the regressor $balsav_{t+1}$. If we were to use $balsav_{t+4}$ as a regressor in the modelling of 2001, then the sample size reduces greatly from five years to two years. Therefore, we decided to base our models on $balsav_{t+1}$.

[2] In an earlier attempt, we used the 2004 predictions in the 2003 imputation. However, this gave rise to very high means in 2001 and 2002. Every imputed year, the values were somewhat overestimated. In order to stop this snowball effect, we decided to use 2005 values only.

A similar procedure for the imputation of risky assets was implemented. [3] Again, when a value $balshabon_{t+1}$ was required in the imputation, we use the value of $balshabon_t$ in 2005. In the imputation of risky assets, the imputed values of $balsav\_t$ are used in the regressions. We also produced imputations without using the imputed values of checking and savings accounts and the difference was negligible.

5 Results of the Imputation of Financial Wealth Components 2001-2004

This section describes the results of the imputation of the checking and savings accounts and the risky assets. We will first discuss them separately, finally we will briefly consider the gross financial wealth of households.

5.1 Checking and savings accounts

In Table 12, the distribution of checking and savings accounts is displayed. The values for the years 2001-2004 are imputed. The values for 2005-2010 are supplied in IPO Wealth. In order to assess our imputations, we use an external data source. The Dutch Central Bank, the DNB (De Nederlandsche Bank), produces a yearly overview of financial assets and liabilities of Dutch households. The DNB collects data on checking and savings accounts from banks and financial institutions. The aggregated amounts for all households in the Netherlands are published. Statistics Netherlands reports the total number of households in the Netherlands per year in Statline, an online database. We combine these two variables. For every year, we divide the aggregated amount of checking and savings accounts by the number of households in the Netherlands. This gives an average value of checking and savings accounts per household. We again use the inflation rates from Statline (Statistics Netherlands) to express the numbers in 2005 euros. This way, we can compare them directly to our imputations. We denote the resulting numbers by DNB averages.

[3] In previous attempts, we also used the 2004 predictions in the 2003 imputation and so forth. This gave rise to very high assigned ownership rates of risky assets in 2001. The mean ownership rate was around 0.8 for the box 3 liable households and 0.3 for not box 3 liable households. As a check, we compared this to the ownership rates of returns on risky assets. These rates were much lower and we concluded that the predicted ownership was far too high. This was probably caused by the fact that we predicted backwards and the variable $balshabon_{t+1}$ is a strong predictor of ownership in year t. Once a household is assigned to own risky assets, the household is very likely to be assigned to own risky assets in the previous year as well. In every imputed year, a fraction of "incorrect" ownerships will be predicted, and due to this backward prediction, the number of "incorrect" ownerships will accumulate. The outcomes indicated that this was occurring.

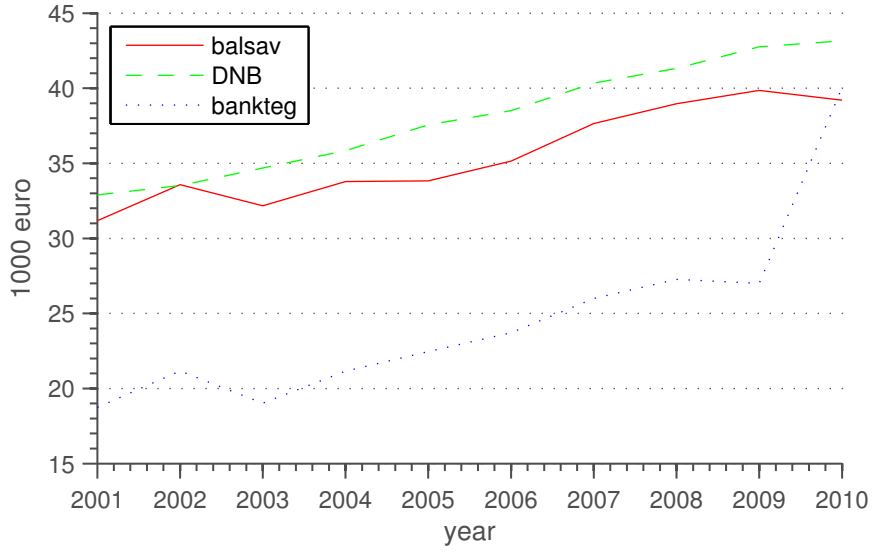
Table 12: Distribution of checking and savings accounts

year	q10	q25	q50	q75	q90	Mean	DNB Averages	Std Deviation
2001	0	3088,38	11966,06	30699,65	75394,65	31172,62	32894,7	77435,68
2002	0	3133,96	12262,13	32111,53	82279,85	33574,83	33519,3	82626,46
2003	0	2948,73	11933,05	31415,44	77511,31	32165,01	34688,2	88084,4
2004	0	2943,46	11676,89	32817,3	82569,36	33782,87	35835,6	96915,78
2005	0	2320	11378,5	34084,5	83810	33820,73	37563,9	89362,12
2006	0	2309,601	11399,61	34985,67	86827,92	35139,02	38509,3	93971,92
2007	0	2648,03	12236,42	37048,07	92867,93	37643,28	40342,7	103933
2008	0	2771,04	12436,68	37766,9	95648,66	38962,07	41322,0	111114,4
2009	338,81	2906,66	12752,88	38599,33	98900,53	39854,24	42750,1	109532,5
2010	298,33	2597,88	12522,44	39066,38	95695,76	39201,05	43159,1	108823,1

The median values are higher in the imputed years than in 2005. After 2005, the median values are increasing for almost every year. Based on this trend, we would expect the median values to be lower in the imputed years than in 2005. We observe that the standard deviations are small in 2001-2003. The values are smaller than the standard deviations in the later years. The standard deviation in 2004 is somewhat higher but it is still lower than the standard deviations in 2007-2010. We conclude that the variation in our imputations is reasonable.

In Figure 1, the mean of $balsav_t$ (IPO Wealth), the DNB average and the mean of $bankteg_t$ (IPO Income) are graphically displayed over time. The DNB average increases quite steadily. In 2005-2010, the mean of $balsav_t$ differs from the DNB average with an approximately fixed amount of 3000-4000 euro. The imputations are much closer to the DNB averages.

Figure 1: Mean $balsav_t$, DNB Average and mean $bankteg_t$



The mean value of $balsav_t$ does not increase strongly in 2001-2004. However, when we compare it to the mean value of $bankteg_t$, we see that this mean does not increase strongly in those years either. The trend in $bankteg_t$ is different from the trend in the DNB averages. Therefore, we conclude that the imputated means of $balsav_t$ seem to capture the overall trend quite well.

We expect the value of checking and savings accounts to differ greatly per age group. Younger households tend to have smaller accounts than older households. Furthermore, there might be differences in cohorts. Therefore, it is interesting to examine cohort effects more closely next to analyzing the overall sample means. In IPO Income, there is a variable which classifies the age of the principle income earner of a household. Cohorts are generated using the age class of the principle earner in 2001. The different classes and the number of observations in 2001 per class are given in Table 13.

Table 13: Age class of principle income earner
Source: IPO Income

Cohort	Ages	Number of observations in 2001
1	..-24	2381
2	25-29	5385
3	30-34	9772
4	35-39	11384
5	40-44	9122
6	45-49	7339
7	50-54	7520
8	55-59	6481
9	60-64	4777
10	65-69	3786
11	70-74	3073
12	75 and older	4099

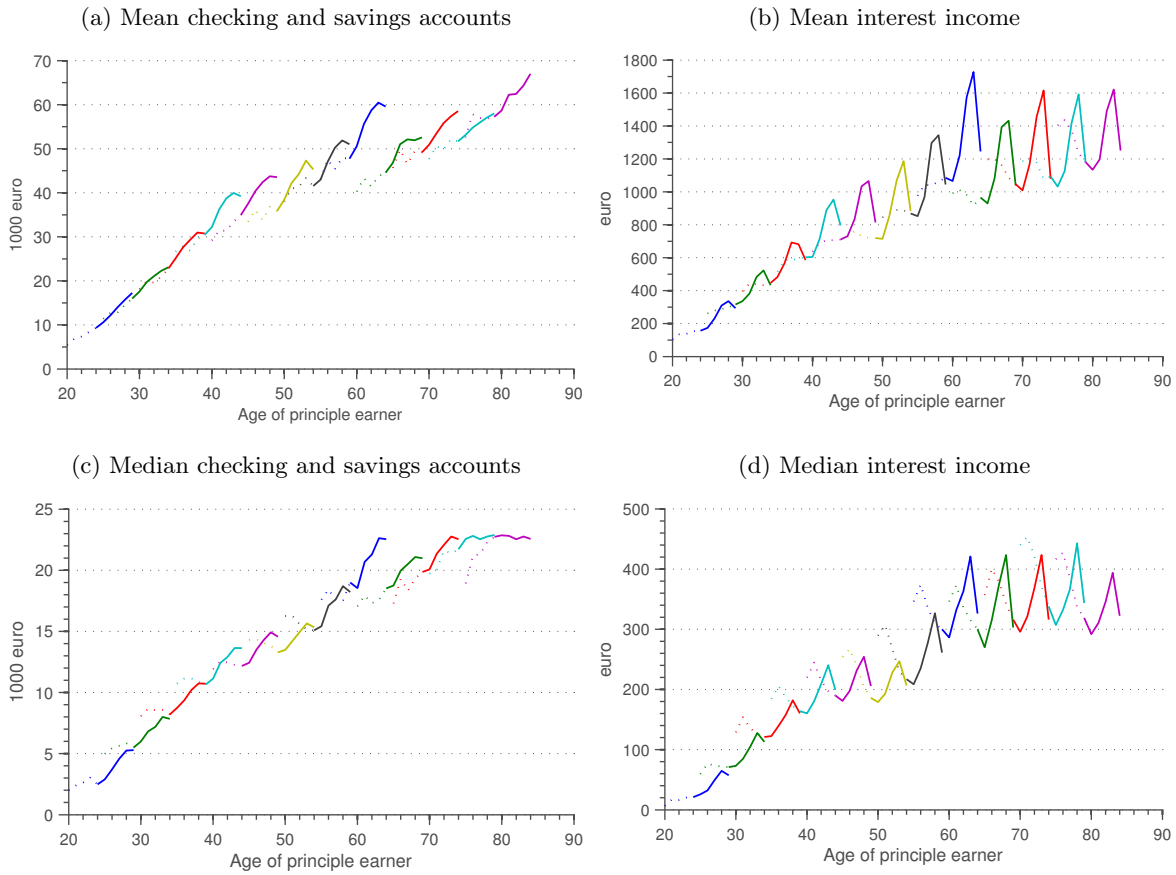
We want to analyze the mean and median of checking and savings accounts per cohort. These are shown in Figures 2a and 2c respectively. The imputations are predicted values. One of the variables which is highly related to checking and savings accounts is the interest income generated from these accounts. This variable is supplied in IPO Income, so we have information for all years. Figure 2b represents the mean interest income from checking and savings accounts for each cohort group and Figure 2d the median values. The solid lines represent 2005-2010 and the dashed lines represent the years 2001-2004. As we expected, there are large differences between cohorts.

We observe that the positive and negative peaks are more extreme for interest incomes than for checking and savings accounts. This holds for all cohorts in all years. Yet, these peaks occur mostly in the later years, for which the actual values of checking and savings accounts are known. The business cycle patterns are more strongly visible in the interest incomes than

in checking and savings accounts. It should be noted that the trends in interest incomes and the amounts people deposit on checking and savings accounts can be different due to business cycles. When the economy is strong, the interest percentages on checking and savings accounts are high. In economic downturns, the interest percentages are low. However, people might be more willing to save money on checking and saving accounts when the economy sours. The accounts could serve as a buffer. In good times, consumption could be much higher. So even though the two variables are strongly related, this (potential) difference between the two variables should be taken into account.

When comparing the means and medians of checking and savings accounts to the means and medians of interest incomes, we see similar trends for most cohort groups. The exception is the median checking and savings accounts of the two oldest cohorts. The medians of $balsav_t$ of this group are increasing in the first years. However, the interest incomes for these groups are strongly decreasing in the first years. The means of $balsav_t$ of these cohorts do follow approximately the same trend as the mean interest incomes. Furthermore, in the years for which the actual values are known, the medians of $balsav_t$ also behave differently from the median interest incomes in these cohorts. This might be due to a discrepancy between the interest incomes and $balsav_t$, for reasons we mentioned above.

Figure 2: Cohort graphs of checking and savings accounts and interest income from checking and savings accounts



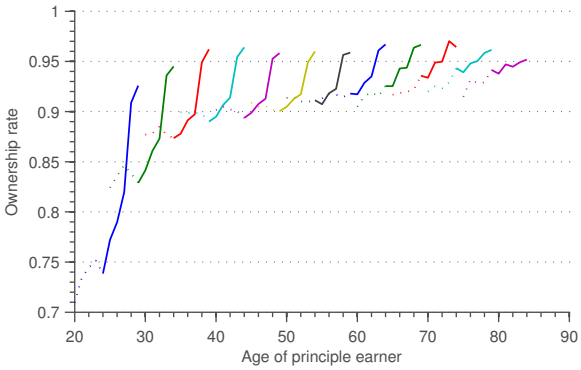
Finally, we look at the ownership rates of the checking and savings accounts. Since the imputations are predictions, we do not know the real ownership rates. However, we do have information available on the ownership of interest income from checking and savings accounts. This is provided in IPO Income. We define a household as owning checking and savings accounts in year t when the (imputed) value of $balsav_t$ is greater than zero. Likewise, when the interest income is positive, a household is said to own interest income. The mean ownership rates per cohort are depicted in Figure 3.

The patterns in both graphs are very similar for the first six cohorts. However for cohorts 6-12, ownership rates of checking and savings accounts increase over the years. The ownership rates of interest incomes decrease in first years. We cannot tell whether the imputation method assigns too few owners or that this deviation is due to business cycle patterns. The ownership of checking and savings accounts is higher than the ownership of interest incomes. This holds for all cohorts in all years, even for the six cohorts where the trends in ownership rates are different. This indicates that, even in the case that the assigned ownership would be too low, the difference is not extremely large.

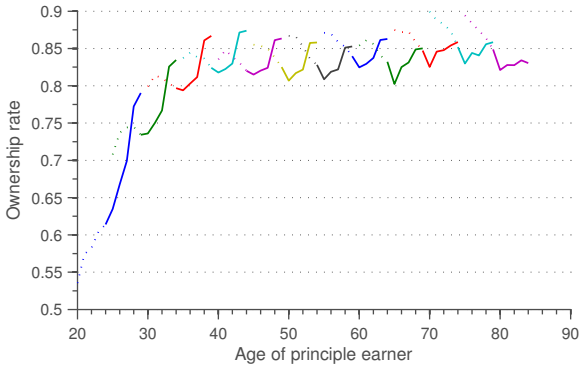
There is an increase in both ownership of interest income and ownership of checking and savings accounts in the last years. This is caused by the increased accuracy of reporting small account balances and small interest incomes.

Figure 3: Ownership of checking and savings accounts and interest income

(a) Ownership checking and savings accounts



(b) Ownership interest income



5.2 Risky assets

In the distribution of risky assets, there are several outliers which influence the means and standard deviations in a year strongly. We decided to discard all observations with a value larger than 10 million euro. This concerned 52 households in total, or 0.006% of the total sample. 17 of these outliers were predicted in the years 2001-2004 and 36 of these values occured in the years 2005-2010.

The distribution of risky assets is depicted in Table 14. The mean ownership rates of risky assets are included as well. Moreover, the MSCI World Index prices at the end of the year are reported (source: MSCI - `http://www.msci.com`). This is a large and mid cap representation across 24 developed countries. The index covers approximately 85% of the free float-adjusted market capitalization in each country. We corrected the index for inflation, so it is measured in 2005 euros. This gives an indication of the performance of the stock markets in our time frame. We have no external information on bonds.

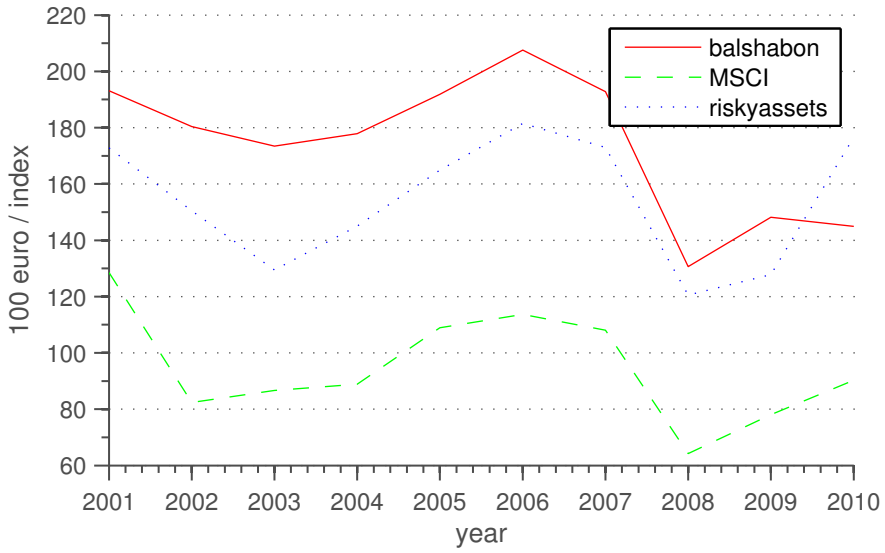
We see that the values of the 75% quantile decrease strongly over time in the known years. In 2005, this quantile was around 1900 euros while it is zero in 2010. The 75% quantiles in the imputed years are higher than in 2005, especially in 2004. The 90% quantiles of the imputed values are lower than the actual values in 2005 and 2006. After 2006, the 90% quantiles are decreasing again.

Table 14: Distribution of risky assets

year	q10	q25	q50	q75	q90	Mean	Std Deviation	Ownership	MSCI
2001	0	0	0	2307,14	26769,32	19310,29	143177,3	0,286	128,488
2002	0	0	0	2307,97	25347,96	18036,22	135126,58	0,288	82,358
2003	0	0	0	2069,87	23443,41	17345,82	145486,6	0,285	86,685
2004	0	0	0	2407,29	25037,45	17786,47	131827,89	0,298	88,900
2005	0	0	0	1916	28693	19184,91	133322,84	0,289	108,885
2006	0	0	0	1328,88	29469,84	20759,01	147982,14	0,279	113,635
2007	0	0	0	711,66	25869,88	19282,42	143996,96	0,267	108,030
2008	0	0	0	37,04	16545,52	13063,8	101001,27	0,253	64,206
2009	0	0	0	27,22	18665,68	14815,59	111352,47	0,253	78,050
2010	0	0	0	0	17733,95	14493,22	114605,49	0,238	90,271

In Figure 4, the MSCI World Index is displayed, together with the (rescaled) means of $balshabon_t$ (IPO Wealth) and $risky\_assets_t$ (IPO Income). Based on the trends in 2005-2010, we conclude that the MSCI World Index is a good external benchmark to evaluate the performance of the mean of the imputed $balshabon_t$. The trends in both lines are very similar for these years. The imputed mean $balshabon_t$ has a very similar trend as the MSCI Index as well. There is a difference in 2002-2003, the MSCI Index increases while the mean $balshabon_t$ decreases. However, the mean of $risky\_assets_t$ decreases in this period as well. We conclude that the quality of the imputed mean $balshabon_t$ is good.

Figure 4: Mean $balshabon_t$, MSCI Index and mean $risky\_assets_t$



The distribution of risky assets is highly skewed to the right. In Table 15, the distribution of risky assets is displayed, conditional upon ownership. We see that, for all statistics, the imputed values in 2001 are the highest and it is decreasing in each imputed year. In the

10% quantile, the decreasing trend continues in 2005-2010. The other quantiles are higher in 2005 than in 2004. Still, the mean value in 2001 is higher than the mean value in 2005. A similar trend occurs in the MSCI index. The MSCI index in 2001 is higher than the price in 2005. However, the MSCI index in 2005 is much higher than the MSCI indices in the years 2002-2004. 2001 was before the financial crises, so the values of risky assets could very well have been higher in 2001 than in 2005. The standard deviations in the imputed years are very similar to the standard deviations in the known years.

Table 15: Distribution of risky assets conditional upon ownership

year	q10	q25	q50	q75	q90	Mean	Std Deviation	obs
2001	1908,69	4583,51	13361	46311,05	134810,19	67517,19	261578,68	21882
2002	1804,27	4388,97	12438,81	43470,25	126666,55	62664,12	246260,2	22825
2003	1718,93	4260,39	12066,81	40696,28	115814,74	60957,98	267820,96	22908
2004	1552,08	3874,62	11403,03	40485,6	115659,88	59670,39	236229,43	25086
2005	1348	4324	14268	46221	133885	66493,75	241788,43	25567
2006	1280,91	4463,4	15231,95	51659,76	148697,38	74414,26	272935,62	25056
2007	1183,82	4109,31	14360	49276,7	143107,71	72206,47	271710,06	24221
2008	856,72	2720,23	9815,23	36097,15	104592,93	51626,53	195765,74	23263
2009	884,1	3115,01	11011,42	40035,3	116211,9	58588,95	215570,38	23734
2010	983,93	3641,11	12056,41	40765,56	120135,18	61027,44	229030,93	23247

Analyzing the ownership of risky assets can be insightful regarding the quality of the imputations. Only a small fraction of households owns risky assets. In Table 14, we see that the ownership rates in the imputed years are very similar to the ownership rates in 2005. This is probably due to the fact that predictions are based on the values in 2005 and ownership in future years is a strong predictor of ownership in earlier years. In 2005-2010, the ownership rate of risky assets is decreasing.

We will now further investigate which households are assigned ownership. The group of households which are imputed owners of risky assets but do not receive any dividend should be carefully dealt with. A possible explanation could be that some stocks do not pay out dividend. Stock owners are rewarded with a so-called 'stock dividend'. Owners receive new stocks instead of dividends. This is not measured in the return of risky assets in IPO Income. Before the tax reforms in 2001, returns on risky assets were taxed. The stock dividends were not taxed, so this might have been an incentive to invest in stocks which paid out stock dividends. Furthermore, some bonds do not pay out interest.

The ownership in 2001-2004 is predicted. It is likely that some households did not own any risky assets in year t, but are predicted to be owners. Likewise, there will be some households which are predicted to hold no assets, but did in fact own assets. This will likely occur for owners which are not box 3 liable, which is a small group. Owners of risky assets are generally box 3 liable and a value of $risky\_assets_t$ is known in IPO Income. In the imputation, we will predict whether $balshabon_t = risky\_assets_t$. If not, $balshabon_t$ is predicted to have a different value. The outcome of the imputations for box 3 liable households will thus always predict ownership when $risky\_assets_t > 0$.

In Table 16, we cross tabulated the ownership rates of risky assets and returns on risky assets. For the group with $balshabon_t > 0$ and zero returns, we also tabulated whether $risky\_assets_t$ in IPO Income equals zero. The fraction of households with $balshabon_t > 0$, $dividends_t = 0$ and $risky\_assets_t = 0$ is much higher in the imputed years. This indicates that 'wrong' ownerships are assigned. Furthermore, we see that the fraction of households with $balshabon_t = 0$ and $dividends_t > 0$ increases in the imputed years. The increase is of roughly the same amount as the increase in the fraction of households with $balshabon_t > 0$, $dividends_t = 0$ and $risky\_assets_t = 0$. This explains why the mean ownership rates are constant over time. The fraction of 'wrong' owners almost completely offsets the fraction of actual owners which are predicted to own nothing. In Table 17, we investigate the group of 'wrong' owners further.

Table 17 shows the total number of households which are (imputed) owners, but receive no

Table 16: Cross table ownership risky assets and returns on risky assets

year	$balshabon_t > 0$ $dividends_t > 0$	$balshabon_t = 0$ $dividends_t = 0$	$balshabon_t = 0$ $dividends_t > 0$	$balshabon_t > 0$ $dividends_t = 0$ $risky\_assets_t > 0$	$balshabon_t > 0$ $dividends_t = 0$ $risky\_assets_t = 0$
2001	0,215	0,687	0,027	0,020	0,049
2002	0,221	0,689	0,023	0,021	0,045
2003	0,221	0,697	0,018	0,020	0,043
2004	0,235	0,690	0,012	0,019	0,044
2005	0,236	0,703	0,009	0,018	0,034
2006	0,231	0,713	0,008	0,018	0,029
2007	0,218	0,723	0,010	0,020	0,029
2008	0,205	0,738	0,009	0,021	0,027
2009	0,198	0,742	0,005	0,021	0,033
2010	0,191	0,755	0,008	0,021	0,025

dividends and have a value $risky\_assets_t = 0$ in IPO Income. We then split this number into box 3 liable households and not box 3 liable households. The increase in (imputed) ownership without dividends occurs mainly in the group of households which are not box 3 liable. The number of these households which are box 3 liable is rather constant in 2001-2010.

Next, we examine whether these households did own risky assets in 2005-2010. If they did own risky assets in 2005-2010, then the imputed ownership in 2001-2004 could be true as well. We tabulated the numbers of the households which are not owners in 2005-2010. This number is quite low.

In Table 32, the distribution of $balshabon_t$ for all (predicted) owners with $risky\_assets_t = 0$ and no returns is displayed. When we compare this to Table 15, we see the values are relatively small. We conclude that there is reason to believe that a number of wrong ownerships are predicted. This mainly concerns households without a taxable box 3 income. The number of wrong ownerships is small and the assigned values are relatively small as well.

Finally, we can look at the cohort trends in ownership rates of risky assets and the ownership rates of the returns on risky assets. The mean values are shown in Figure 5. We distinguish

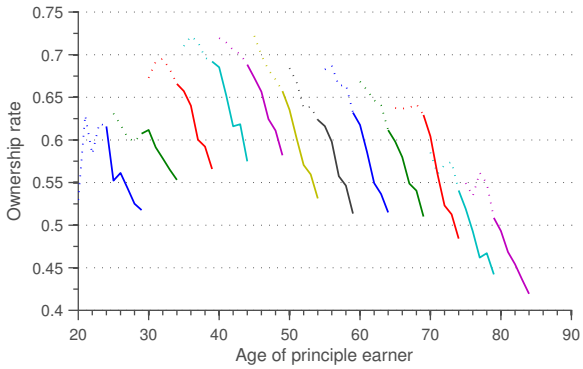
Table 17: (Imputed) Owners of risky assets without returns on risky assets and $risky\_assets_t = 0$ in IPO Income

year	Total Households	Households Box 3 liable	Of which hold no risky assets in 2005-2010	Households Not box 3 liable	Of which hold no risky assets in 2005-2010
2001	3786	439	80	3347	560
2002	3543	364	91	3179	653
2003	3490	385	102	3105	657
2004	3727	496	177	3231	1018
2005	2984	369	-	2615	-
2006	2577	354	-	2223	-
2007	2602	420	-	2182	-
2008	2468	413	-	2055	-
2009	3140	614	-	2526	-
2010	2466	477	-	1989	-

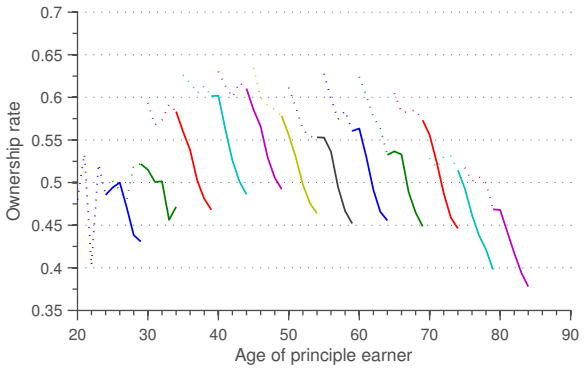
between households which are box 3 liable and households which are not box 3 liable. We see that the ownership of risky assets is higher than the ownership of the returns on risky assets in both groups for all cohorts in all years. For both box 3 liable households and not box 3 liable households, the trends in ownership of the returns look very similar to the trends in ownership of the assets. This indicates that our imputation method captures the mean ownership trend in cohorts very well.

Figure 5: Ownership of risky assets and interest and dividend returns

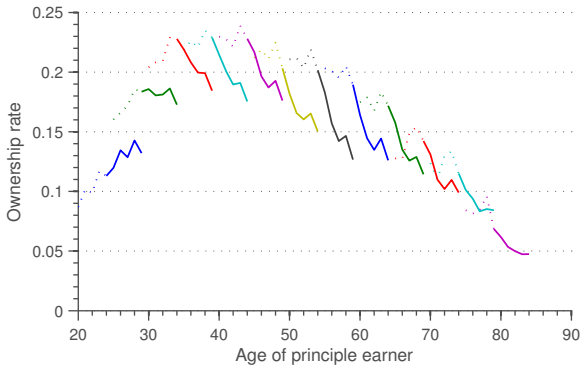
(a) Ownership risky assets (box 3 liable)



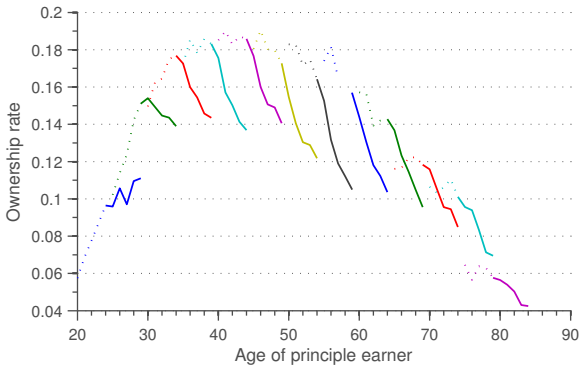
(b) Ownership returns on risky assets (box 3 liable)



(c) Ownership risky assets (not box 3 liable)



(d) Ownership returns on risky assets (not box 3 liable)



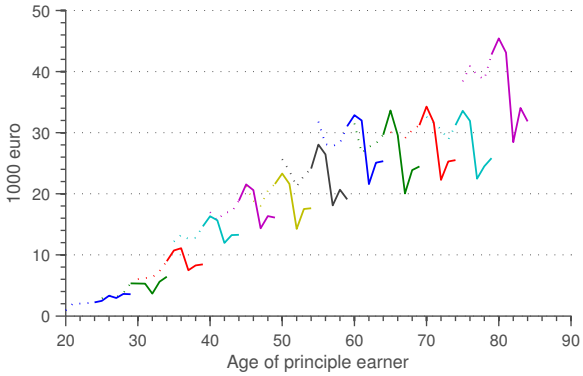
Next to ownership rates, the cohort effects in the mean of risky assets and in the mean and median values conditional upon ownership are insightful. We can compare these values to the mean returns on risky assets and to the mean and median values of the returns conditional upon ownership of risky assets respectively. This is displayed in Figure 6. There are large differences between cohorts. The means and median of the youngest cohort is very close to zero while these values are much higher for the older cohorts.

We can compare the (imputed) value of risky assets to the known value of returns on risky assets. For the unconditional means (Figure 6a and 6b), we see some small differences in cohorts 3-5. The mean risky assets increase in the first years while the mean returns on risky assets drop strongly.

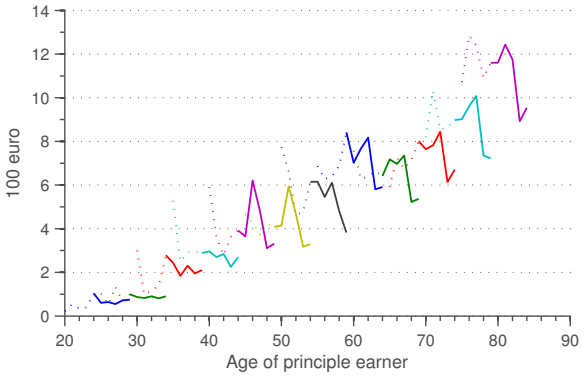
The conditional means (Figure 6c and 6d) show very similar cohort patters in the risky assets and returns on risky assets. The only difference is that mean returns increase and decrease more sharply than the risky asset means. This holds in both the imputed and actual years. It might be due to the cycle of financial markets. In good years, returns can increase more strongly but they can decrease heavily in bad years. In the cohort graphs of conditional medians (Figure 6e and 6f), we see the sharper increase in mean returns for the older cohorts as well. The trends in both graphs are again very similar.

Figure 6: Cohort graphs of risky assets and returns on risky assets

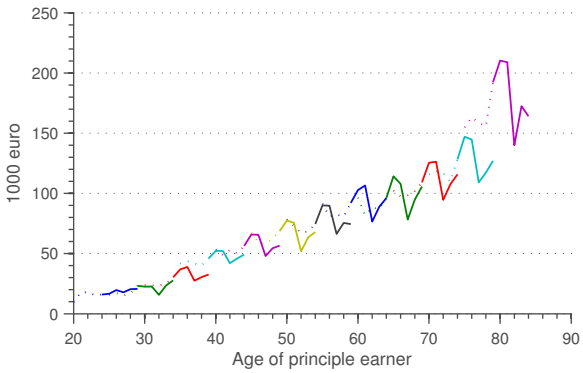
(a) Mean risky assets



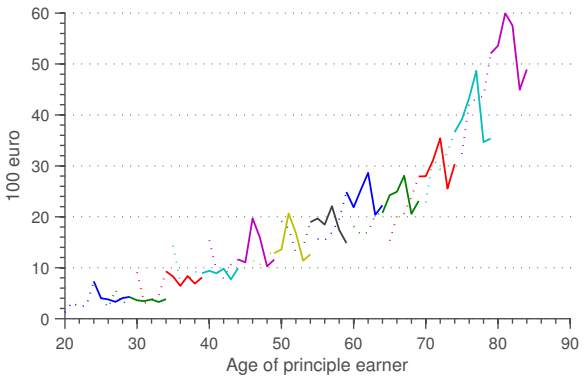
(b) Mean returns



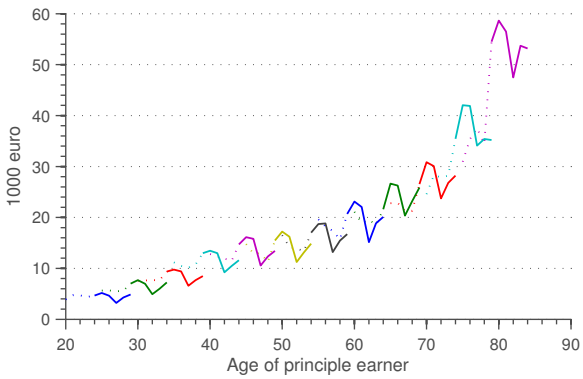
(c) Mean risky assets conditional upon ownership



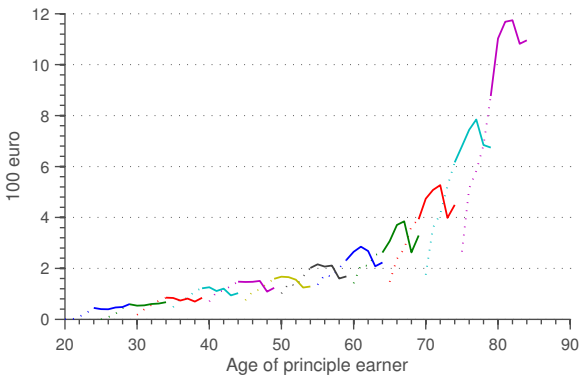
(d) Mean returns conditional upon ownership risky assets



(e) Median risky assets conditional upon ownership



(f) Median returns conditional upon ownership risky assets



5.3 Gross financial wealth

As we now have imputations for the checking and savings accounts and risky assets, we can look at the distribution of gross financial wealth. This is reported in Table 18. In the 10%

quantile, the effect of more accurate reporting of small values in IPO Wealth in later years is visible again. In 2002 and 2003, the values of the 10% quantile are extremely low. This might be caused by a 'wrong' prediction of ownership of checking and savings accounts or risky assets. The first part model predicts ownership while this may not have been true. The right hand side variables in the second part model generate very low predictions, but they are positive.

The mean is generally increasing over the years, with the exception of 2003 and 2008. The standard deviation in the imputed years is comparable to the standard deviation in the observed years.

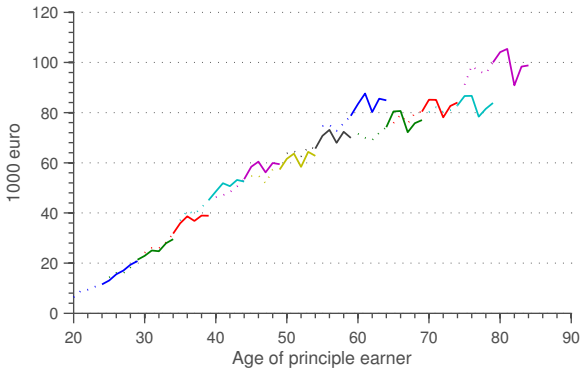
Table 18: Distribution of gross financial wealth

year	q10	q25	q50	q75	q90	Mean	Std Deviation
2001	0	3804,4	14470,72	39671,48	112943,24	50482,91	178945,03
2002	6,65	3841,08	14654,12	40964,44	119054,12	51611,06	176371,27
2003	0,4	3601,48	14279,67	40017,29	108870,7	49510,83	188060,55
2004	0	3574,93	14039,83	42365,07	115468,45	51569,34	183379,38
2005	0	2854	14138,5	43317	119509	53005,64	182605,97
2006	0	2734,92	14071,22	44220,16	125034,66	55898,03	200160,13
2007	0	3066,65	14674,17	45398,24	127533,74	56925,71	205746,42
2008	134,87	3087,8	14203,78	43633,66	119911,43	52025,87	177089,36
2009	428,91	3254,86	14578,34	45015,07	127010,74	54669,83	181707,06
2010	352,99	2937,9	14420,35	45147,86	120942,36	53694,28	183833,16

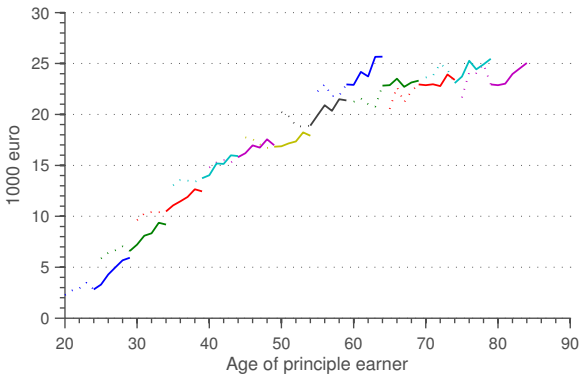
Finally, we can show the cohort effects of gross financial household wealth. The mean and median values are displayed in Figure 7. In both graphs, the trends are similar.

Figure 7: Cohort graph of gross financial wealth

(a) Mean gross financial wealth



(b) Median gross financial wealth



6 Conclusion

We finished the imputation of the values of checking and savings accounts and risky assets for the years 2001-2004. The imputations are based on future values of these variables and data on tax records. A priori, we believed that the values on tax records would be the same as the values in IPO Wealth. This turned out not to be true, as was shown in Table 5. This complicated the imputations.

The quality of the imputations of checking and savings accounts is high. In the 2005 imputation, we concluded that the distribution of the imputed variable was very similar to the distribution of the actual variable. This was true for households with a taxable box 3 income and for households without a taxable box 3 income. The upper quantiles for both groups were slightly overestimated and this resulted in larger means. The overall correlation between the actual and the imputed variable was 0.91, which is very high.

In order to assess the quality of the 2001-2004 imputation, we compared the imputed means to the average value of checking and savings accounts published by the DNB and the variable $bankteg_t$ which is provided in IPO Income. We conclude that the imputed means capture the trend in checking and savings accounts very well.

The cohort trends in mean and median interest incomes from checking and savings accounts are in general very similar to the cohort trends in imputed mean and median checking and savings accounts. The ownership rates per cohort look very reasonable as well when we compare them to the ownership rates of interest incomes. There is a discrepancy for the last six cohort groups. The ownership rates of checking and savings accounts are increasing in the imputed years, whereas the ownership rates of interest incomes in these years are decreasing. This could be caused by business cycle effects or it might be that the imputation method assigns too few owners. The ownership rates of checking and savings accounts are higher than the ownership rates of interest incomes. This holds for all cohorts in all years, even for the cohorts where the trends in both rates differ from each other. From this follows that, even if too few owners are predicted, the predicted ownership rate could not be substantially different from the actual rate. External data on ownership rates of checking and savings accounts could serve here as a final check.

Before 2009, there is no information on small values of checking and savings accounts in IPO Wealth. Therefore, we were not able to impute the small values in 2001-2004. Since 2009, banks and financial institutions report these small values more accurately. As the IPO Wealth panel is extended each year with another wave, more information on small balances will be available. This makes it possible to develop an imputation method for these small values in the future.

The imputations of risky assets look good as well. The 95% quantile was overestimated in the 2005 imputation of both box 3 liable households and not box 3 liable households. This resulted in that the imputed means were higher than the actual means. The other quantiles were very close to the actual values. The overall correlation between the actual and the imputed variable was 0.86.

The 2001-2004 imputations follow roughly the same trends as the MSCI World indices. The dissimilarities between the MSCI and the imputations can be explained by trends in

$risky\_assets_t$ (IPO Income). Therefore, we conclude that the quality of the imputed means in 2001-2004 is high.

When we look at the cohort trends in the mean risky assets, we see many similarities with the cohort trends in the mean returns of risky assets. This holds for both the unconditional and the conditional means. The mean ownership rates in 2001-2004 are comparable to the ownership rate in 2005. However, in 2005-2010 there is a strong decreasing trend in the ownership rate. When comparing the cohort trends in ownership of risky assets to the ownership of returns on risky assets, we concluded that the cohort trends are very similar in both graphs. This indicates that the imputed ownership rate per cohort of risky assets is very close to the actual values.

There is a strong indication that a small proportion of households which are not box 3 liable, are predicted to be owners of risky assets while they probably did not own any stocks. This occurs in every imputed year. Likewise, a proportion of households in the not box 3 liable group are predicted to hold no risky assets, but they do receive income from risky assets. They probably did own risky assets in 2001-2004. In the 2005 imputation, the actual and predicted ownership was different for a small group as well. The group which was predicted to hold no assets but actually did own assets was of almost the same size as the group which was predicted to hold assets but actually did not own assets. This yielded similar predicted and actual ownership rates. The distribution of risky assets of the 'wrong' owners are relatively small as compared to the full distribution of owners of risky assets.

The imputation method which is developed in this paper does not cover other components of household wealth. The values of secondary residences, shares of a substantial holding, business equity and movable property ('roerende zaken') are measured on a household level in IPO Wealth. There are no variables closely related to these components in IPO Income. Furthermore, the group of people owning any of these other components will be very small. The values of these components will be large. In IPO Income, the variable $debt\_other_t$ is reported for households which are box 3 liable. These other debt asets could have been used to finance the other components of household wealth. Therefore, we did not consider net financial household wealth of box 3 liable households in this paper. A topic for future research could be to impute all components of household wealth. External data related to these components in the years 2001-2004 are necessary to develop a proper method.

Overall, the quality of the imputations of checking and savings accounts and risky assets in 2001-2004 seems to be high. The ownership rates of checking and savings accounts of the last six cohorts have different patterns than the ownership rates of interest incomes of these cohorts. External data could be used here to find out what the actual trends were and whether the imputation method predicts reasonable rates or not. Furthermore, the ownership rates of risky assets have some flaws in households which are not box 3 liable. Future research can refine the imputation method.

References

Andridge, R.R. and R.J.A. Little (2010). A review of hot deck imputation for survey non-response. *International Statistical Review 78*(1), 40–64.

Belastingdienst (2012). Explanatory notes tax return Form C. `www.belastingdienst.nl`.

Bethlehem, J.G. (2008). Wegen als correctie voor non-respons. Centraal Bureau voor de Statistiek.

Cameron, A.C. and P.K. Trivedi (2005). *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

CBS (2010a). Documentatierapport inkomenspanelonderzoek (IPO) V1. Centraal Bureau voor de Statistiek.

CBS (2010b). Documentatierapport inkomenspanelonderzoek met vermogen (IPO Vermogen) V2. Centraal Bureau voor de Statistiek.

Davies, J.B., S. Sandstrom, A. Shorrocks, and E.N. Wolff (2006). The world distribution of household wealth. Commissioned paper, World Institute for Development Economics Research, Helsinki.

Dempster, A.P., N.M. Laird, and D.B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society 39*(1), 1–38.

European Central Bank (2013). The eurosystem household finance and consumption survey - results from the first wave.

Fries, G., M. Starr-McCluer, and A.E. Sundén (1998). The measurement of household wealth using survey data: An overview of the survey of consumer finances. Federal Reserve Board of Governors.

Greenwood, D. (1973). An estimation of U.S. family wealth and its distribution from microdata. *Review of Income and Wealth 29*(1), 23–44.

Hayashi, F. (2000). *Econometrics*. Princeton: Princeton University Press.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics 2*(4), 303–314.

Kalton, G. and D. Kasprzyk (1982). Imputing for missing survey responses. *Proceedings of the Survey Research Methods Section*, 22–31. American Statistical Association.

Kelton, W.D. and A.M. Law (2007). *Simulation Modeling and Analysis* (4 ed.). New York: McGraw-Hill.

Knoef, M. and K. De Vos (2008). Representativeness in online panels: how far can we reach. `www.lissdata.nl`.

Lepkowski, J.M. (1989). The treatment of wave nonresponse in panel surveys. In G. Kasprzyk, G. Duncan, G. Kalton, and M.P. Singh (Eds.), *Panel Surveys*, New York. J.W. Wiley and Sons.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review 54*(2), 139–157.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economics Statistics 6*(3), 287–296.

Mittinty, M.N. and E. Chacko (2005). Imputation by propensity matching. *American Statistical Association: Proceedings of the Survey Research Methods CD-ROM*, pp 4022–4028.

Poterba, J.M., S.F. Venti, and D.A. Wise (2012). Were they prepared for retirement? Financial status at advanced ages in the HRS and AHEAD cohorts. NBER Working Paper No. 17824.

Ross, S.M. (2009). *Introduction to Probability Models* (tenth ed.). New York: Academic Press.

Rubin, D.B. (1987). *Multiple Imputation in Sample Surveys and Censuses*. New York: John Wiley.

Verbeek, M.J.C.M. (2004). *A guide to modern econometrics* (second ed.). Chichester: John Wiley and Sons.

Wolff, E.N. (1983). The size distribution of household disposable wealth in the United States. *Review of Income and Wealth 29*(2), 125–146.

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.

A Imputation methods

The problem of missing values is categorized in unit nonresponse and item nonresponse. Unit nonresponse occurs when no data are collected for a respondent. Item nonresponse refers to the situation when only a subset of information is missing for a respondent. In panel data, a third category of missing data arises as a combination of unit nonresponse and item nonresponse. Wave nonresponse occurs when one or more waves are missing for a respondent that has provided data for at least one other wave (Lepkowski, 1989).

When values of one variable are missing, this is known as univariate nonresponse. In the case that values of multiple variables are missing, the problem is referred to as multivariate nonresponse. In this section, strategies to compensate for unit nonresponse, item nonresponse and wave nonresponse will be discussed respectively. First, we discuss the problem of sample selection and why there is a need to correct for the missing values.

In Wooldridge (2010), sample selection is defined as a nonrandom sample. Several selection mechanisms result in nonrandom samples. Some are due by the sample design, others are related to the behaviour of the respondents. Nonresponse on survey questions or attrition from social programs are examples of a nonrandom sample due to the behaviour of the respondents. The nonresponse can be selective. This occurs when the nonrespondents behave differently with respect to the research variables than the respondents. In other words, the reason why observations are missing is correlated with the dependent variables. Estimates will be systematically too low or too high (Bethlehem, 2008). An example of this occurs in surveys of household wealth. The very rich are often reluctant to provide data about their wealth (European Central Bank, 2013). Total wealth will be underestimated.

A.1 Unit nonresponse

Unit nonresponse reduces the sample size. This leads to an increase in estimated standard errors. This is in itself not a problem. When the nonresponse is selective, the set of complete observations is systematically different from the set of nonrespondents. A standard method to correct for this problem, is the assignment of a set of weights to the complete cases in the data set. Typically, the sample weights for complete cases are divided by the (estimated) response rate in a subclass of the sample (Little, 1986).

A.2 Item nonresponse

Imputation is a general strategy used to compensate for item nonresponse. Imputation is the assignment of values at the microlevel. Several procedures will be discussed in the next paragraphs. According to Kalton and Kasprzyk (1982), imputation has three advantages. First of all, imputation can reduce bias caused by sample selection. Moreover, the data set can be used as if it were complete. This makes it easier to analyze the data and present the results. Third, the results obtained from different analyses are bound to be consistent. This need not be true with an incomplete data set. However, there are several downsides to imputation. Although imputation can reduce the bias, it does not necessarily lead to less biased estimated than those produced from the incomplete data set. Furthermore, analysts should not treat the imputed data set as an actual complete data set. This would overstate the precision of the estimates.

Let the imputed value of variable y for non-respondent i be denoted by y_i. Then,

$$y_i \quad = \quad f(x_{1i}, \ldots, x_{pi}) + \epsilon_i, \tag{78}$$

where $f(\boldsymbol{x})$ is a function of observed auxiliary variables $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})$ and ϵ_i is an error term. An imputation process is called stochastic when the errors are randomly added to the predicted value, i.e. $E(\epsilon_i) = 0$. A deterministic imputation procedure involves no error term, $\epsilon_i = 0$. In the following paragraphs, we will describe several imputation procedures.

- *Logical (or deductive) imputation*
  This method can only be used when there is an exact relation between the missing value and the observed values. The missing value can be deduced from the information which is provided. For example, in the Dutch tax system, a tax is levied on home ownership. A percentage of the actual value of the house ("WOZ Waarde"), is derived. This is the so-called notional rental value or "eigenwoningforfait". Once this value is known, the value of the residence can be calculated directly.
  This method is preferred over any other method, if logical imputation is possible, it should be executed.

- *Mean imputation*
  For this method, the total sample is divided into several imputation classes according to auxiliary variables. For each class, the imputed value is equal to the mean of all complete responses in that class. In the most extreme case, the mean of all complete responses is assigned to the missing responses. This strategy is easy to implement and it does not change the sample means in the classes. However, the distributions and associations between the variables are distorted. The variance of the variable with missing values, y, is strongly underestimated.

- *Cold deck imputation*
  In cold deck imputation, the missing values are replaced by data from an external source. For example in the IPO Income panel, information on the checking and savings accounts for households which are not box 3 liable is not provided. The Dutch Central Bank produces annual values on the total amount of checking and savings accounts for all households. The average value of checking and savings accounts can be calculated and these could be used as imputed values for households which are not box 3 liable.

- *Hot deck imputation*
  Andridge and Little (2010) define hot-deck imputation as the replacement of missing values of one or more variables for a nonrespondent (which they call recipient) with observed values from a respondent (the donor). The donor is similar to the recipient with respect to characteristics observed for both respondents. A simple method is the creation of classes by ordering the observed categorical variables. A donor which is in the same class as the recipient is then selected. When this selection is random, it is referred to as *random hot deck imputation*. In *deterministic hot deck methods*, a single donor is selected based on a so-called "nearest-neighbour" principle. If there is no donor available in a class, categorical variables can be dropped until a suitable donor is found.

  There are many methods to determine which donor is the closest to the recipient. Metrics, such as minimum distance (where the variables should be scaled so they can be compared

with each other) or the Mahalonobis distance function, can be considered. Once a metric is defined, a group of donors can be formed by setting maximum distances.

Hot deck imputation is widely used in practice. Its advantages are that it imputes real values and it does not need strong parametric assumptions. Furthermore, covariate information can be incorporated. In order to keep the associations in case of multivariate nonresponse, matching each non-respondent to a single donor is preferred to finding different donors for all missing variables. A drawback of hot deck imputation is that some donors can be used multiple times. Certain methods therefore impose a limit d as to how many times a donor can be used. The optimal choice of d is not yet researched. Another downside is that hot deck imputation requires good matches of donors to recipients. Finding a good match may be problematic in a small sample.

The hot deck imputation procedure has many different forms, they differ in the classification and what further rules are imposed on the donors. It is therefore difficult to evaluate the hot deck method consistently. According to Andridge and Little (2010), the existence of at least some respondents with complete responses to all items that are related to missingness, is critical for the hot deck imputation method to be consistent.

- *Regression based imputation*
  Regression based imputation uses respondent's data to regress the variable y, for which values are missing, on observed variables $\boldsymbol{x}$ for the complete cases. The imputations are computed as the predicted values from the regression equation. The regression can be done through Ordinary Least Squares, but, for example, probit regressions could be considered as well. The obtained estimate is a conditional mean. The direct use of these values affects the distributions and associations between variables. Alternatively, a stochastic error term can be added to the predicted values. These errors can be determined in different ways, depending on the distributive assumptions. For example,

  - The error terms are assumed to follow an I.I.D. normal distribution. Then, the errors can be drawn from a normal distribution with mean zero and variance equal to the residual variance from the regression.
  - The errors are assumed to follow the same distribution, which is unknown. Then, the error terms can be drawn from the empirical distribution of the respondent's residuals.
  - A hot-deck method can be applied. The recipients can be classified in different groups and residuals can be (randomly) selected from a donor.

  Using this modification of the regression imputation preserves distributions and correlations between variables (Kalton and Kasprzyk, 1982).

- *Predictive Mean Matching*
  This method is an application of hot deck imputation. It was first proposed in the context of statistical matching. In Little (1988), it is extended to handle nonresponse. Similar to the regression based imputations, the predicted mean of the missing variable is estimated. Next, each nonrespondent is matched to a donor with the closest predicted mean. The donor's value is imputed directly to the nonrespondent's missing value. Its advantages over regression imputation are that only feasible values are imputed. Moreover, since the model is used only to select a match, it is less sensitive to model misspecification. When comparing it to another hot deck method, such as using the nearest neighbour in the Mahalonobis

distance metric, the predictive mean matching is preferred by Little (1988). According to him, variables which are not associated with the missing y are not used in the predictive mean calculation. However, they might influence the Mahalonobis distance strongly. Critics of this method argue that it heavily depends on the prediction model. The match is done based on the predictions, not on the predictors, so there is a great deal of faith in the model that one is using. In an ordinary hot-deck procedure, the match is done based on the predictors.

Instead of predicting means, in some applications an estimated probability of response is calculated and the match is based on this probability. Mittinty and Chacko (2005) introduce this imputation as *Propensity score matching*.

- *Imputation using the Expectation-Maximization (EM) Algorithm*
  Originally, the EM algorithm was developed by Dempster, Laird, and Rubin (1977) to obtain maximum likelihood estimates from incomplete data. The algorithm consists of two steps. In the expectation step, the expected complete data loglikelihood is calculated based on the observed data. In the maximization step, the parameter values are updated by calculating the maximum of the expected complete data loglikelihood. This process is repeated until convergence. In the E-step, the expected values of the missing data are calculated. These expectations can be used as imputations, this is a deterministic method. Stochastic imputations can be obtained by random drawings from the specified distribution. The maximum likelihood estimates are used as parameters. A drawback of the EM algorithm is that a distribution needs to be specified, so it is prone to model error. Furthermore, the EM algorithm can be computationally exhaustive and it can take a long time to develop it.

- *Multiple Imputation*
  After an imputation procedure has finished, a data set looks complete and data analysis can start. However, when doing inference, the computed standard errors do not take the imputation effects into account. In order to correct for the imputation, Rubin (1987) introduced the notion of multiple imputation. Each missing value is replaced by a set of imputed values which represents the uncertainty in imputation. This is not possible for deterministic techniques. Here, every repetition leads to the same imputed value. In hot deck imputation, it can be accomplished by finding K nearest respondents in the metric and sample from this set. To simulate the distribution correctly, sampling with replacement is required (Little, 1988).

- *Weighting*
  Instead of imputing the missing items, one could also consider a weighting strategy. The incomplete cases are discarded and the complete cases are assigned a set of weights. Kalton (1986) discusses the advantages and disadvantages of both imputation and weighting. The main drawback of weighting with respect to imputation, is that incomplete cases are discarded which implies a loss of information. However, it does have several advantages over imputation. With imputation, data are fabricated to some extent (except in the event of logical imputation). This increases the uncertainty in the data set and it can cause an attenuation in some of the covariances between variables. Weighting does not lead to additional uncertainty or this attenuation problem. However, it is possible that different sets of weights need to be used for different types of analyses (Lepkowski, 1989). A set of weights which provides reasonably accurate predictions for one purpose may not work well for an-

other. This may lead to inconsistent results. Another difference is that weighting is a global strategy, which treats all variables at the same time. The selection of imputation procedure can be item specific. The choice whether to impute missing items or use a weighting strategy depends heavily on the auxiliary information available. Imputation is more efficient than weighting when it is based on a model with high predictive power. However, when the auxiliary variables are weakly correlated to the missing variables, weighting tends to be preferred.

A.3 Wave nonresponse

When dealing with wave nonresponse, there is more information available about a particular nonrespondent in other waves. Many survey items are repeated over time. Responses to some of the items may be stable over time. A simple and effective imputation procedure is to use the responses from the closest available wave. This procedure is called *carry-over* or *direct substitution imputation*. This may produce much better imputations than standard cross-sectional imputation procedures.

When responses to repeated items are highly correlated over waves, the value of an item on one wave will be a strong predictor of a missing value of the same item on another wave. A hot deck or a regression based imputation procedure can be used. Next to cross-sectional variables, responses on other waves can be used as auxiliary variables.

Since wave nonresponse is a form of unit nonresponse in a particular wave, weighting is another option to be considered. The choice of adjustment procedure for wave nonresponse is not straightforward. Several articles have paid attention to whether weighting or an imputation procedure is more appropriate. According to Kalton (1986), the number of missing waves, the missing data pattern and the correlation with earlier waves and auxiliary variables highly influence which method is preferred.

B Dutch income tax system

In the Netherlands, taxes are paid over income generated from labour, business activities and savings and investments. In 2001, a so-called box system was introduced where income is categorized in three different boxes.

In box 1, income from labour and home ownership is taxed. Income from labour includes wages, profits, pension payments and social benefits. Home ownership is taxed through a notional rental value of the residence. The WOZ value ("Waardering Onroerende Zaken" or Valuation of Immovable property) is determined by municipal authorities. The notional rental value is a progressive percentage of the WOZ value. The tax rate in box 1 is progressive and it differs per income category and age. For people under the age of 65, the rate currently varies from 34% to 52%. For people over the age of 65, the rate presently differs between 16 and 52%. There are several deductible expenditures, such as travel expenses for public transport, deduction of mortgage interest and expenditure on income insurance.

In box 2, taxes on income from substantial shareholding are levied. In Section 2.1, the definition of substantial shareholding as maintained by the Dutch Tax Authorities is provided. Taxes are levied on the dividends of these shares or on the profit made from selling them. The

tax rate is currently 25%.

In box 3, income from savings and investments is taxed. In this box, the value of the assets are taxed using a fictitious return. The actual returns are not taxed in the box system. In other words, the tax in box 3 is basically a wealth tax. The fictitious return is calculated using the average value of assets minus liabilities over the year. The part of this value which exceeds the tax free allowance ("heffingvrij vermogen") is taxed using a fixed return of 4%. On this return, a 30% tax rate is imposed. The tax free allowance was 19.522 euro per person in 2005.

C Distribution of financial wealth variables in IPO Income

Table 19: Distribution of $bankteg_t$

year	q10	q25	q50	q75	q90	Mean	Std Deviation
2001	0	0	0	0	56544,97	18743,92	107620
2002	0	0	0	0	66453,24	21165,96	76643,56
2003	0	0	0	0	58064,52	19018,8	72412,58
2004	0	0	0	3341,62	63349,68	21146,45	82534,23
2005	0	0	0	7682	66498	22452,47	86933,39
2006	0	0	0	10216,62	69069,76	23712,66	91610,01
2007	0	0	0	13879,27	74306,41	26001,27	102545,5
2008	0	0	0	14791,24	77288,08	27275,34	107742,6
2009	0	0	0	13639,8	77438	27005,23	104949,7
2010	0	0	0	20883,08	120916,2	40047,25	142083,8

Table 20: Distribution of $risky\_assets_t$

year	q10	q25	q50	q75	q90	Mean	Std Deviation
2001	0	0	0	0	17611,47	17282,76	165359,1
2002	0	0	0	0	16053,48	15049,28	144101,7
2003	0	0	0	0	12835,03	12946,84	109977,1
2004	0	0	0	0	14534,34	14504,64	144816,3
2005	0	0	0	0	17418	16492,73	159072,6
2006	0	0	0	0	17911,48	18147,78	197102
2007	0	0	0	0	15265,11	17298,28	200927,5
2008	0	0	0	0	8006,81	12050,86	179731,4
2009	0	0	0	0	7936,29	12778,09	180689,6
2010	0	0	0	0	8920,25	17663,3	190199,8

D Classification in groups based on taxable income in box 3

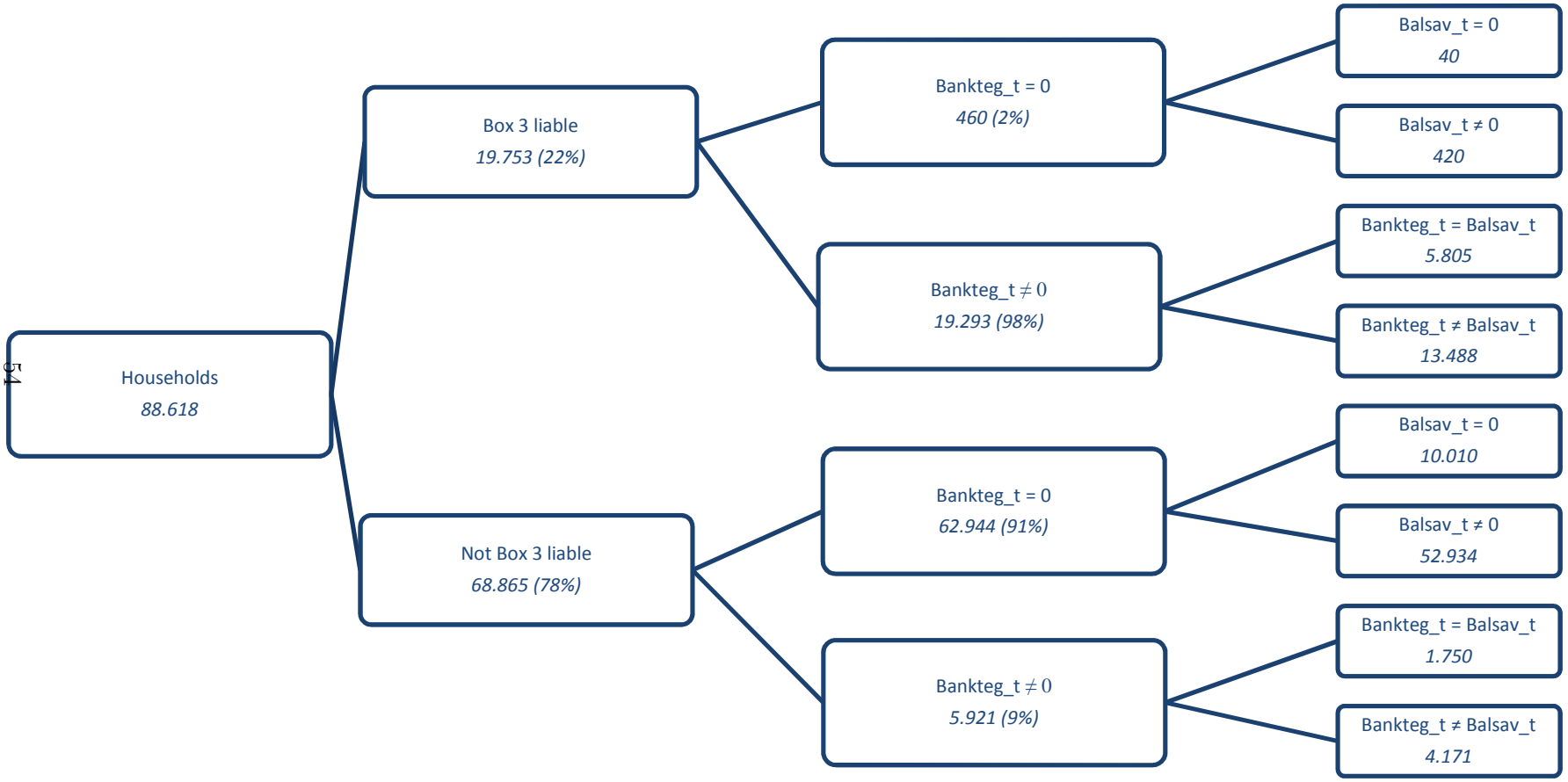
Table 21: Sample size per group
Source: IPO Income

Year	Households in group 1 *(Taxable income in box 3)*	Fraction of total	Households in group 2 *(No taxable income in box 3)*	Fraction of total	**Total**
2001	16874	0,201	67067	0,799	83941
2002	17440	0,205	67654	0,795	85094
2003	17411	0,201	69026	0,799	86437
2004	18503	0,211	69197	0,789	87700
2005	19753	0,223	68865	0,777	88618
2006	20530	0,229	69280	0,771	89810
2007	21432	0,236	69272	0,764	90704
2008	21657	0,236	70281	0,764	91938
2009	21822	0,232	72041	0,768	93863
2010	23345	0,239	74532	0,761	97877

Table 22: Transition rates between groups (conditional on a household being present in year $t+1$)
Source: IPO Income

	Households in group 1 in year t		Households in group 2 in year t	
Year	In group 1 in $t+1$	In group 2 in $t+1$	In group 1 in $t+1$	In group 2 in $t+1$
2001	85.6 %	14.4 %	4.5 %	95.5 %
2002	86.2 %	13.8 %	3.6 %	96.4 %
2003	89.2 %	10.8 %	4.3 %	95.7 %
2004	90.2 %	9.8 %	4.5 %	95.5 %
2005	90.9 %	9.1 %	3.7 %	96.3 %
2006	91.0 %	9.0 %	4.0 %	96.0 %
2007	89.7 %	10.3 %	3.5 %	96.5 %
2008	91.9 %	8.1 %	2.7 %	97.3 %
2009	90.5 %	9.5 %	4.0 %	96.0 %

E Household classification in 2005



F Probit modelling of dum_sav_t

Table 23: Right hand side variables used in probit regressions

Name of variable	Description of variable	Source
dum_sav_{t+1}	value of dum_sav in year $t+1$	IPO Wealth
$dum_{t=\alpha}$	time dummies $\alpha = \{2006, \ldots, 2009\}$	
$dum_{bankteg_t=0}$	dummy for $bankteg_t = 0$	IPO Income
$dum_{bankteg_{t+1}=0}$	dummy for $bankteg_{t+1} = 0$	IPO Income
$dum_{balsav_{t+1}=0}$	dummy for $balsav_{t+1} = 0$	IPO Wealth
$dum_{interest_t=0}$	dummy for $interest_t = 0$	IPO Income
$I_{interest_t>0} * \ln(interest_t)$	natural logarithm of interest income from checking and savings accounts in year t	IPO Income
$I_{bankteg_t>0} * \ln(bankteg_t)$	natural logarithm of $bankteg_t$ in year t	IPO Income
$I_{assets_t>0} * \ln(assets_t)$	natural logarithm of the average value of assets in year t	IPO Income
$I_{risky\_assets_t>0} * \ln(risky\_assets_t)$	natural logarithm of the value of stocks and bonds in year t	IPO Income
$I_{balsav_{t+1}>0} * \ln(balsav_{t+1})$	natural logarithm of $balsav_{t+1}$	IPO Wealth
$I_{balsav_{t+2}>0} * \ln(balsav_{t+2})$	natural logarithm of $balsav_{t+2}$	IPO Wealth
$I_{contrsav_t>0} * \ln(contrsav_t)$	natural logarithm of contractual savings ("spaarloon") in year t	IPO Income
cons	constant term	

Table 24: Estimation results of Probit regressions
z statistics in italics

	Group 1	Group 2
$dum_{t=2007}$	-0,027318	-0,067956
	-1,325279	*-5,445085*
$dum_{t=2008}$	0,36511	0,053438
	17,79967	*4,204676*
dum_sav_{t+1}	2,304557	1,896811
	125,1321	*80,97265*
$dum_{bankteg_t=0}$	-0,923693	0,119666
	-8,082284	*1,62062*
$dum_{bankteg_{t+1}=0}$	0,530016	0,513599
	15,05355	*21,03136*
$dum_{balsav_{t+1}=0}$	-1,943399	-2,154497
	-15,171	*-53,9064*
$dum_{interest_t=0}$	1,33586	1,42734
	20,24502	*56,2431*
$I_{interest_t>0} * \ln(interest_t)$	0,029024	-0,041445
	3,778134	*-7,455372*
$I_{assets_t>0} * \ln(assets_t)$	0,032638	0,023697
	5,290939	*4,66199*
$I_{risky\_assets_t>0} * \ln(risky\_assets_t)$	-0,007847	-0,008601
	-4,804054	*-3,445487*
$I_{bankteg_t>0} * \ln(bankteg_t)$	0,057286	0,214437
	5,380841	*26,76124*
$I_{balsav_{t+1}>0} * \ln(balsav_{t+1})$	-0,089696	-0,175334
	-7,640936	*-40,42946*
$I_{balsav_{t+2}>0} * \ln(balsav_{t+2})$	-0,031856	-0,072412
	-4,025364	*-37,36608*
$I_{contrsav_t>0} * \ln(contrsav_t)$	-0,231083	-0,135258
	-47,28592	*-35,99172*
cons	-1,033922	-0,714782
	-10,82107	*-9,662065*
$\ln(\sigma_u^2)$	-13,11846	-11,60495
Number	55509	191773
Number of groups	18835	65081
σ_u	0,001417	0,00302
ρ	2,01E-06	9,12E-06

G Fixed effects modeling of $balsav_t$

Table 25: Right hand side variables used in fixed effects regressions

Name of variable	Description of variable	Source
$hhsize_t$	household size in year t	IPO Income
$hhearners_t$	number of earners in household in year t	IPO Income
$hhtype_t$	household type (e.g. family with children in year t)	IPO Income
$housetype_t$	type of house (e.g. rental or owned) in year t	IPO Income
$hhincsrc_t$	main source of household income (e.g. wages or pensions) in year t	IPO Income
$dum_{interest_t=0}$	dummy for $interest_t = 0$	IPO Income
$I_{interest_t>0} * \ln(interest_t)$	natural logarithm of interest income from checking and savings accounts in year t	IPO Income
$dum_{bankteg_t=0}$	dummy for $bankteg_t = 0$	IPO Income
$dum_{bankteg_{t+1}=0}$	dummy for $bankteg_{t+1} = 0$	IPO Income
$I_{bankteg_t>0} * \ln(bankteg_t)$	natural logarithm of $bankteg_t$	IPO Income
$I_{bankteg_{t+1}>0} * \ln(bankteg_{t+1})$	natural logarithm of $bankteg_{t+1}$	IPO Income
$dum_{balsav_{t+1}=0}$	dummy for $balsav_{t+1} = 0$	IPO Wealth
$I_{balsav_{t+1}>0} * \ln(balsav_{t+1})$	natural logarithm of $balsav_{t+1}$	IPO Wealth
$I_{contrsav_t>0} * \ln(contrsav_t)$	natural logarithm of contractual savings in year t	IPO Income
$I_{taxinc3_t>0} * \ln(taxinc3_t)$	natural logarithm of taxable income in box 3 in year t	IPO Income
$I_{grossinc_t>0} * \ln(grossinc_t)$	natural logarithm of gross household income in year t	IPO Income
$I_{assets_t>0} * \ln(assets_t)$	natural logarithm of the average value of assets in year t	IPO Income
dum_{box3t}	dummy whether a household is box 3 liable in year t	IPO Income
$I_{risky\_assets_t>0} * \ln(risky\_assets_t)$	natural logarithm of the value of stocks and bonds in year t	IPO Income
cons	constant term	

Table 26: Estimation results of fixed effects regressions
z statistics in italics

	Group 1∗	Group 2∗
$hhsize_t$	0,143531	0,148840
	23,065100	*28,708550*
$hhearners_t$	0,054883	0,105782
	9,228548	*19,708730*
$hhtype_t$	0,004828	0,016919
	2,790291	*13,011570*
$housetype_t$	-0,010409	-0,009512
	-1,522228	*-1,771293*
$hhincsrc_t$	-0,004788	-0,001940
	-3,173438	*-1,442785*
$dum_{interest_t=0}$	-0,269904	0,111077
	-10,044540	*9,383400*
$I_{interest_t>0} * \ln(interest_t)$	0,058508	0,149495
	23,869790	*71,652100*
$dum_{bankteg_t=0}$	5,306561	3,393776
	139,216000	*76,380820*
$dum_{bankteg_{t+1}=0}$	0,678245	0,504322
	17,987040	*12,320890*
$I_{bankteg_t>0} * \ln(bankteg_t)$	0,586921	0,422226
	174,472100	*92,188000*
$I_{bankteg_{t+1}>0} * \ln(bankteg_{t+1})$	0,077121	0,064588
	20,623130	*15,897570*
$dum_{balsav_{t+1}=0}$	0,075709	0,257641
	1,373269	*11,426850*
$I_{balsav_{t+1}>0} * \ln(balsav_{t+1})$	0,028361	0,037506
	6,654372	*15,399480*
$I_{taxinc3_t>0} * \ln(taxinc3_t)$	0,024287	0,055967
	7,389548	*10,822290*
$I_{grossinc_t>0} * \ln(grossinc_t)$	0,035977	0,069405
	10,556450	*21,777510*
$I_{assets_t>0} * \ln(assets_t)$	-0,009139	-0,022030
	-3,968957	*-8,409291*
$I_{risky\_assets_t>0} * \ln(risky\_assets_t)$	-0,006127	-0,004922
	-6,874344	*-3,031179*
dum_{box3t}	-0,102073	-0,352682
	-4,509907	*-10,640750*
$I_{contrsav_t>0} * \ln(contrsav_t)$	0,014631	0,024603
	11,505820	*19,438340*
cons	2,274610	2,622996
	34,171910	*36,934130*
Number	53569	2039261
Number of groups	13675	54350
σ_u	0,391663	0,957630
σ_e	0,369390	0,657827
ρ	0,529241	0,679405
R^2 overall	0.8195	0.4623

Table 27: Distributions of standardized residuals $\hat{u}_{it}$ for both groups

	Residuals	
	Group 1*	Group 2*
Percentiles		
1%	-2.878326	-2.929222
5%	-1.526546	-1.508967
10%	-1.0292	-1.018744
25%	-0.4208967	-0.4350456
50%	0	0.0053503
75%	0.400721	0.4577858
90%	0.9831455	1.041862
95%	1.471349	1.517764
99%	2.82042	2.705814
Mean	-0.0141001	0.0010183
Std. Dev.	1.001264	0.9890561
Variance	1.00253	0.978232
Skewness	-0.2658475	-0.6212872
Kurtosis	13.72305	11.57247
Observations	53569	203926

Table 28: Distributions of standardized $\hat{e}_{it}$ for both groups

	Residuals	
	Group 1*	Group 2*
Percentiles		
1%	-3.234861	-3.038912
5%	-1.709214	-1.601776
10%	-1.139521	-1.068497
25%	-0.4406936	-0.4618963
50%	0.0104306	-0.0053519
75%	0.4370695	0.4457946
90%	1.088138	1.03301
95%	1.671358	1.516015
99%	3.265885	2.744972
Mean	-0.0039636	-0.0233714
Std. Dev.	1.124784	1.017409
Variance	1.265138	1.035122
Skewness	-0.1430486	-0.8029932
Kurtosis	16.8786	12.86124
Observations	39794	147220

H Imputation Results of $balshabon_{2005}$

Table 29: Cross-tabulation of predicted and actual outcome in differences risky assets

		$\hat{y}_{i,2005}$		
		0	1	Total
$y_{i,2005}$	0	13676	4229	17905
	1	4654	62684	67338
	Total	18330	66913	85243
Pseudo R^2 (as developed in Verbeek (2004))		0.504		
Correlation $\hat{y}_{i,2005}$ and $y_{i,2005}$		0.689		

Table 30: Distributions of households with taxable box 3 income

	Variable			
	$balshabon_t$	Imputed $balshabon_t$	$balshabon_t$ given > 0	Imputed $balshabon_t$ given > 0
Percentiles				
5%	0	0	1614	1641
10%	0	0	3807	3774
25%	0	0	12937	12037
50%	10703	10496	40602	37967
75%	60741	60147	103367	101546
90%	163862	163768	240085	243024
95%	289970	310509	429798	446742
Mean	78459	85728	122841	132992
Trimmed mean	*75305*	*79705*	*117921*	*123663*
Std. Dev.	331804	633463	408562	785009
Trimmed Std. Dev.	*267078*	*324737*	*326610*	*397719*
Variance	1.10e+11	4.01e+11	1.67e+11	6.16e+11
Trimmed Variance	*7.13e+10*	*1.05e+11*	*1.07e+11*	*1.58e+11*
Skewness	19.711	76.481	16.208	62.167
Trimmed Skewness	*12.328*	*14.916*	*10.169*	*12.269*
Kurtosis	606.520	8136.062	405.993	5337.152
Trimmed Kurtosis	*226.187*	*315.031*	*152.965*	*211.747*
Observations	19134	19134	12221	12334
Trimmed Observations	*19129*	*19130*	*12216*	*12330*

Table 31: Distributions of households without taxable box 3 income

	Variable			
	$balshabon_t$	Imputed $balshabon_t$	$balshabon_t$ given > 0	Imputed $balshabon_t$ given > 0
Percentiles				
5%	0	0	542	522
10%	0	0	907	1070
25%	0	0	2416	2293
50%	0	0	6585	5312
75%	0	0	15654	13205
90%	6123	5022	31153	29329
95%	15086	12731	42380	47078
Mean	2877	6026	14913	31183
Trimmed mean	*2877*	*2650*	*14913*	*13713*
Std. Dev.	24249	868365	53564	1975249
Trimmed Std. Dev.	*24249*	*23293*	*53564*	*51539*
Variance	5.88e+08	7.54e+11	2.87e+09	3.90e+12
Trimmed Variance	*5.88e+08*	*5.43e+08*	*2.87e+09*	*2.66e+09*
Skewness	65.629	256.835	30.886	112.899
Trimmed Skewness	*65.629*	*99.168*	*30.886*	*46.726*
Kurtosis	6575.146	66011.91	1403.918	12755.62
Trimmed Kurtosis	*6575.146*	*15033.7*	*1403.918*	*3205.87*
Observations	66109	66109	12752	12775
Trimmed Observations	*66109*	*66108*	*12752*	*12774*

I Distribution of risky assets

Table 32: Distribution of the (imputed) value of risky assets when $risky\_assets_t = 0$ and returns are zero

year	q10	q25	q50	q75	q90	Mean	Std Deviation	Observations
2001	962,89	2042,39	4180,97	9060,58	18402,55	8802,57	20998,8	3786
2002	1001,97	1966,73	3894,79	8359,96	16463,47	7719,42	13433,52	3543
2003	854,11	1824,48	3798,15	7876,9	16557,21	10603,53	111726,1	3490
2004	860,19	1729,78	3415,92	7157,22	15515,98	8339,36	57469,42	3727
2005	577	1331	3612	8776,5	17826	8036,17	22779,31	2984
2006	214,64	1153,31	3829,87	10431,26	22889,23	9760,35	27877,32	2577
2007	179,13	1036,82	3868,36	10096,58	23914,04	9255,46	15346,58	2602
2008	85,48	773,14	2513,65	7399,42	17633,04	7280,65	14184,52	2468
2009	59,13	615,21	2550,95	8104,76	20951,97	9776,7	45352,65	3140
2010	64,85	599,44	2667,83	8153,11	20846,03	7974,77	15212,28	2466