Erasmus University Rotterdam

Econometrics & Management Science

Master's Thesis

Macroeconomic drivers of household income growth

Exploring heterogeneity across subpopulations via latent class modeling

Supervisor: dr. W. Wang
Coreader: dr. A. Alfons
2017-10-19

Author:
Dennis Bogers

Abstract

Which macroeconomic factors drive household income growth and to what extend are these driving forces heterogeneous across different subpopulations, were the central questions of this research. To answer these questions, a latent class model has been estimated. Whereby new parameter values in the maximization step of the expectation-maximization algorithm where determined via the outlier robust generalized method of moments estimator of Lucas et al. (1997). Households were generally segmented into one group with average and stable incomes and one or two other groups with below average and volatile incomes. The effect of income inequality was generally negative for the latter group, while it was positive for the first group. Redistribution and gross domestic product growth usually had no significant effect. Hence, redistribution is a good tool to lower inequality and therewith support overall income growth.

KEYWORDS: Household income growth, Income inequality, Redistribution, Latent class model, Robust generalized method of moments estimation

Contents

1 Introduction

Which level of income inequality within a country would be optimal in the sense of social acceptability and economic growth, has been a frequently asked question from the very beginning of economic research onwards. Several economic models have been tried, ranging from communism with total income equality as its aim, up to liberalism where one was striving for a government as small as possible, which often resulted in a very high degree income inequality. Though, there was one economic assumption that almost became common belief for economic researchers in the past. This assumption stated that a higher level of income inequality would lead to a higher level of economic growth. It was based upon the idea that a higher inequality level gives more incentives to excellence and invest. Hence, economic papers were written on the question which level of income inequality would give an optimal trade-off between economic efficiency and social acceptability (Okun, 1975).

However, recent growth in data availability of macro-economic factors of most countries gave economic researchers the possibility to empirically check this assumption. Results of these empirical studies definitely put the assumption in doubt. A paper that has been written on behalf of the Organisation for Economic Co-operation and Development (OECD) concludes for example, 'drawing on harmonized data covering the OECD countries over the past 30 years, the econometric analysis suggests that income inequality has a negative and statistically significant impact on subsequent growth. In particular, what matters most is the gap between low income households and the rest of the population' (Cingano, 2014). A more global empirical research led by the International Monetary Fund (IMF) comes to similar conclusions, 'lower net inequality is robustly correlated with faster and more durable growth, for a given level of redistribution.' (Ostry et al., 2014) Moreover, the researchers also tried to answer the question if redistribution would be a good solution, which led to the following conclusion. 'Redistribution appears generally benign in terms of its impact on growth; only in extreme cases is there some evidence that it may have direct negative effects on growth. Thus the combined direct and indirect effects of redistribution, including the growth effects of the resulting lower inequality, are on average pro-growth.' (Ostry et al., 2014)

As these empirical studies are on a macro-economic level, they can only answer the second part of the question. 'Which level of income inequality within a country would be optimal in the sense of social acceptability and economic growth'. For example, what if we find an optimal level of income inequality in terms long term sustainable economic growth, that has a large negative effect on the household income growth of certain subpopulations. Probably, this level of income inequality cannot be seen as optimal from a social acceptability perspective. Therefore, my aim is to extend the macroeconomic empirical literature to a microeconomic scale. Hence, instead of modeling the gross domestic product (GDP) growth of a country, this research uses household level real disposable income growth as the dependent variable. Because reliable microeconomic data of developing countries is hardly available, I will focus on the household income trends of the industrialized world. Besides the level of income inequality, it is also possible to investigate the effects of other macro-economic factors and government policy related variables on household income growth within the proposed modeling framework. For example, by adding GDP growth as explanatory variable, one has

the possibility to investigate which households have an income that is more vulnerable to macro economic shocks.

Besides looking at the effect of income inequality on GDP growth, Ostry et al. (2014) also investigated if redistribution by the government would be a good solution to lower inequality and therewith establish a positive impact on subsequent GDP growth. Tools for redistribution might be direct transfers or using a progressive tax system. However, economists have pointed towards the fact that (too) progressive tax systems lead to economic inefficiencies and should therefore be avoided. On the other hand, some researchers doubt if the trade-off between equality (due to progressive taxes) and efficiency is really inevitable (Roed and Strom, 1999; van Ewijk et al., 2003). They point towards the fact that some common market failures are (partly) solved by charging progressive taxes. Hence, in order to investigate this issue empirically, I will also model the effect of the level of redistribution within a country on household income growth.

The main advantage of modeling household level data is that households can be segmented into subpopulations. Thereafter one can investigate if there are any significant differences in the effects of macroeconomic variables on income growth of household from distinct subpopulations. In contrast to most socio-economic literature, the segmentation of the households will be data driven and thus will not be based on household characteristics. By doing so, one is able to find optimal subpopulations. Optimal in the sense that the effects of the macroeconomic factors on household income growth are as homogeneous as possible within subpopulations and as heterogeneous as possible between subpopulations. To illustrate the downside of the traditional segmentation method via household background variables for this research. Suppose that there exist a subpopulation created via household characteristics, for which half of the subpopulation experiences a significant positive effect of the income inequality level within its country on its income growth, while the other half experiences a significant negative effect. Segmentation based on household characteristics would then probably yield an insignificant income inequality effect for this subpopulation. Which in turn could lead to the inaccurate conclusion that a certain level of income inequality is also optimal for this subpopulation from a social acceptability perspective.

Note that there are no country restrictions upon the data driven subpopulations. Hence, the homogeneous subpopulations of households in terms of the response of their income growth on macro economic factors, could be quite heterogeneous in terms of the countries in which they live. This is a conscious choice, since often reported results of recent economic globalization and automation are more heterogeneity between economic classes within countries and the contrary of more homogeneity between similar economic classes of distinct countries (Alderson and Nielsen, 2002; Kentor, 2001). From a labor income perspective, there is more heterogeneity between economic classes within countries. As there is a highly educated class who has benefited largely from these transformations, while low-skilled workers fear their job is taken over by a cheaper foreigner or a computer. On the other hand, labor income patterns of similar economic classes from different countries are getting more homogeneous due to heightened international competition and an increase in migration. Continuing with the second source of income for households, which are assets returns, we find a similar pattern. Firstly, the French economist Thomas Piketty showed in his book 'Capital in the 21st century' that the distribution of capital is heading towards very unequal pre-world-war lev-

els in Western countries. Hence, asset returns are also getting more heterogeneous across different subpopulations within a country. In contrast, globalization of the financial world has led to a higher degree of interdependence of asset returns from households originated in different countries. In conclusion, given the current state of the globalized economy, it is in my opinion far more interesting to allow for international population segments. By doing so, one is able to find global economic classes who respond similarly to changes in macroeconomic factors.

In order to model the heterogeneous effects of the macroeconomic variables on household income growth properly, I will make use of the finite mixture model, which will estimated by means of the expectation-maximization algorithm (EM algorithm) (Dempster et al., 1977). The maximization step of the EM algorithm will proven to be equal with estimating a weighted dynamic linear panel model, whereby the weights correspond to the segment probabilities for a certain household. This weighted dynamic linear panel model will be estimated via a generalized method of moments (GMM) estimator that is robust against anomalous observations. Besides heterogeneity in the estimated effects of the macroeconomic factors, the latent class model allows for heterogeneous household income dynamics across different subpopulations. In order to investigate global trends in the drivers of household income dynamics, data of the following five countries will be used, Germany, South-Korea, Switzerland, the United Kingdom (UK) and the United States (US). This yearly household panel data originally comes from national longitudinal household surveys. The corresponding household surveys are respectively SOEP, KLIPS, SHP, BHPS and PSID. Note that the corresponding full names of these abbreviated survey names are given in table 1. The data of all distinct surveys is harmonized by the Cross-National Equivalent File (CNEF). Hence, the eventual dataset contains equally defined variables that are based on previous research on cross-national comparable measures.

To summarize this introduction, the aim of this research is to answer the following question. Which macroeconomic factors drive household income growth, to what extend are these driving forces heterogeneous across different subpopulations and how are these subpopulations best characterized? The main focus will be on investigating the effect of the level of income inequality on household income dynamics. Next to this, I will examine if redistribution is a good tool for governments to lower income inequality. The vulnerability of household income to the general economic performance of a country will also be investigated by incorporating GDP growth to the model. Lastly, the effects of certain household characteristics on income growth will be investigated and these characteristics will be used to identify the different subpopulations. This paper is organized as follows, in section 2, a review on the existing literature of this topic will be given. Then the properties and quality of the CNEF data will be discussed in further detail in section 3. Thereafter, I will give the required methodology for answering the research questions in section 4. In the following section, results of the latent class models will be presented. Lastly, the results will be summarized and discussed in section 6 and 7.

2 Relevant Literature

While there has been put quite some effort into modeling labor income, the literature on modeling household income dynamics directly is quite scarce. Though, there is quite some literature on modeling measures derived from household income, such as income mobility measures and income groups to model poverty dynamics. As noted by Stephen Jenkins in his summary paper on household income dynamics, a reason could be that researchers fear the impossibility to model all underlying forces of household income dynamics. Because in labor earnings dynamics one can focus on homogeneous subgroups (e.g. prime age men), which makes the assumption of a simple model structure more plausible. Moreover, in earnings dynamics one models a single income source, so accounting for the combination of different income sources and household composition changes is not necessary, while it is for modeling household income (Jenkins, 2000). The issues addressed by Jenkins can be split into two parts. Firstly, the impossibility to propose a model that is both parsimonious and is capable of capturing all the underlying forces and heterogeneity of household income dynamics. To overcome this issue, I will make use of robust estimation techniques to estimate the parameters. By doing so, households with aberrant income dynamics, which cannot be explained by the model, will be down weighted. A second reason for using a robust estimation method is the stylized fact that the distribution of income growth is fat-tailed, even for labor earnings of homogeneous subpopulations (Guvenen et al., 2015). Hence, observations with inexplicable extreme changes in annual household income will have no or little influence on the estimated parameters. Which is a good property in this setting, since the purpose of this research is to find general effects of the explanatory macroeconomic variables on the income shift of household segments. In contrast, when one would investigate poverty dynamics, these extreme observations obviously cannot be ignored. The second issue is that changes in household composition play a major role in household income dynamics. DiPrete and McManus (1999) and DiPrete and McManus (2000) have put effort into modeling these household composition changes explicitly. The underlying goal was to study the differences in the average impact of a certain household composition change on household income in different countries. A more advanced way of modeling was proposed by Burgess and Propper (1998), as they used hazard models for the probability of a certain household composition change. However, note that the variability of the impact of a household composition change, such as a divorce, on household income is quite large. Moreover, the changes in income can get quite extreme in case of a household composition change. Again, as the purpose is to find general effects of macroeconomic variables, I will threat a household before and after a household composition change as two separate households. Two exceptions to this rule are when a child is born or leaves the house, since the impact on household income changes is generally low in these cases. This topic will be further discussed in section 3.

As just has been mentioned, the number of academic papers where the income growth of households is modeled is not large. Most of the papers for which income growth was modeled focused on finding drivers of income mobility, which is defined as the ability of a family to improve (or lower) their income. Fields et al. (2003) were one of the first to investigate the relationship between household income growth and some household characteristics. They estimated separate models for Indonesia, South Africa, Spain and Venezuela and compared results across these countries. Woolard and Klasen (2005) and Aristei and Perugin (2015)

both use the same model specification that was proposed by Fields et al. (2003). Woolard and Klasen (2005) extended the existing literature by looking at the differences in income dynamics between rural and urban households. Aristei and Perugin (2015) used data of 25 European countries, which they classified into six capitalistic models, such that possible differences in the drivers of income growth could be found across these capitalistic models. Next to this, they added some job characteristics as covariates. However, these three papers are less useful for this research, as they all make use of a cross-sectional model instead of a panel model. Hence, they model the income growth of only one time-period. Aaberge et al. (2002) used panel data of Scandinavian countries to find some potential drivers of income mobility and eventually compared results with the US. Though, the dependent variable they use is relative income change, which is slightly different. Next to this, the number of covariates they investigate is quite low. Note that I could not find any paper which was focusing on heterogeneity in the drivers of household income growth across population segments, whereby the populations segments are created by some cluster algorithm. However, cross-country household panel datasets are used quite often for such research. For example, Clark et al. (2005) use a finite mixture model to show that the effect of income on reported well-being is quite heterogeneous across segments. Next to this, I was not able to find any paper that investigates the relationship between macroeconomic variables and household income growth. Although the idea of investigating the vulnerability of households to macroeconomic shocks by including GDP as explanatory variable was obtained from Glewwe and Hall (1998). These researchers examined this vulnerability of households in Peru by modeling their consumption growth instead of their income growth.

Two interesting papers have been written on the dynamics of household income level instead of income growth. Jalan and Ravallion (2002) used panel data of rural China to show that the dynamics of household income exhibit non-linearity. More precisely, there research revealed that the first lag and its squared and cubed value were all significant explanatory variables. Lokshin and Ravallion (2004) applied this non-linear approach to Hungarian and Russian panel data during their economic transition in the nineties. They extended the literature by using different type of estimators. Jalan and Ravallion (2002) estimated the dynamic linear panel model via the regular first differenced GMM estimator proposed by Arellano and Bond (1991). Lokshin and Ravallion (2004) on the other hand used the system GMM estimator of Blundell et al. (2000). This estimator has the advantage that time-invariant covariates can be added to the model, though some extra assumptions are required. However, there main contribution was to estimate the dynamic linear panel model by means of the Semi-Parametric Full Information Maximum Likelihood method (SPFIML). By doing so, they were able account for possible attrition bias in the estimated parameters.

Sample attrition can be defined as: 'partial response in the sense of response in the initial periods but nonresponse in later periods (incomplete participation)' (Cameron and Trivedi, 2005). When this sample attrition is a non-random process, this could lead to biased parameter estimates, which is referred as attrition bias. A general example of non-random sample attrition is that lower income households tend to have a higher probability of sample attrition than households with a higher income. Several researchers have investigated the potential problems of non-random sample attrition in the US dataset. This quote obtained from Lillard and Panis (1998) summarizes the findings of all researchers pretty decently. 'While we found significant evidence of selective attrition, it appears that this nonrandom censoring

introduces only very mild biases in substantive results. No substantive conclusions regarding the processes that generate household income, adult mortality, marriage formation and marriage dissolution would change if attrition is ignored.' More recently, Fitzgerald (2011) investigated the potential problem of attrition bias in intergenerational models of health and found little-to-no evidence of biased estimates. For the German, UK and Swiss datasets, the potential issue of attrition bias has been examined as well. Behr et al. (2005) find minimal attrition effects for income regressions and income mobility analysis on the German and British data. Lastly, Lipps (2007) was not able to find any significant attrition effects for the Swiss data. Because none of these researchers could find any significant attrition bias, I will not make use of the SPFIML method, which accounts for possible attrition bias in the estimated parameters.

As was already mentioned in the introduction, the finite mixture model will be estimated by means of the EM algorithm (Dempster et al., 1977). Because some but not all parameters are segments specific, the finite mixture model will be non-linear. However, the big advantage of using the EM algorithm is that the obtained model in maximization step is actually linear. I will follow Jalan and Ravallion (2002) and Lokshin and Ravallion (2004) in the sense that lagged values of household income growth will be included as explanatory variables. Hence, the eventual model that needs to be estimated in the maximization step of the EM algorithm will be a dynamic linear panel model. The most frequently used estimator for such models is the first differenced GMM estimator of Arellano and Bond (1991). However, this estimation method is not outlier robust, as was required.

By my knowledge, there are two outlier robust estimation methods for dynamic linear panel models. Firstly, Lucas et al. (1997) extended the first differenced GMM estimator of Arellano and Bond (1991) by including observation weights that account for outlying model errors as well as outlying instrumental variables. These observation weights are estimated iteratively with the other model parameters. Secondly, Dhaene and Zhu (2009) derived an outlier robust estimation method based on a linear transformation of the median ratio of adjacent first-differenced data pairs. Later on, the estimator of Dhaene and Zhu (2009) was extended for better efficiency and robustness properties by Aquaro and Cizek (2013). From a robustness perspective, it would be better to use the latter estimation method as it is globally robust, while the estimator of Lucas et al. (1997) is only locally robust. However, three other major downsides of the estimator of Dhaene and Zhu (2009) have led to decision to use the outlier robust GMM estimator of Lucas et al. (1997) in this paper. Firstly, Dhaene and Zhu (2009) only derived an estimator for the basic dynamic linear panel model with one lag. Extending their estimator to higher order or non-linear dynamics (which is necessary according to Jalan and Ravallion (2002)), is very difficult or may be even impossible. For certain, it is not within the scope of this research to perform such an extension. Secondly, a necessary assumption for this estimator is time-stationary of the sequence of income growth values for each household. However, Dynan et al. (2012) show in their paper that the standard deviation of income growth of US households rose with 25 percent between 1970 and 2000, which violates this assumption. The last and probably most important reason is the fact that the estimator of Dhaene and Zhu (2009) is asymptotically biased towards zero in case of independent additive outliers and biased upward when there are patched additive outliers. The major downside of the outlier robust GMM estimator of Lucas et al. (1997), which is the lacking of global robustness properties, arises from the fact that the observations weights

and the model parameters are calculated iteratively. This makes the eventual parameter estimates dependent on the chosen starting observation weights. In order to overcome this issue, I will try multiple starting conditions and compare the resulting parameter estimates.

3 Data

3.1 General information

As was already mentioned in the introduction, the cross-national equivalent file (CNEF) has harmonized yearly household panel data of distinct countries. In total, household surveys of eight countries participated with this project. Hence, the data of three countries (Australia, Canada and Russia) has not been used in this research. There are distinct underlying causes, such as data quality issues or simply data access restrictions. The CNEF contains harmonized demographical, employment, income and medical variables, whereby variables of the first three categories are exploited in this research. Table 1 gives some general information of the household panel data of all five countries that are used in this research. Frick et al. (2007) have written a paper in which they discuss the CNEF data in further detail. Next to this, extensive codebooks of each country can be found on the website of CNEF. In these codebooks, one can find basic properties of each variable and which algorithm is used to construct them from the original household panels.

	Original household panel survey	Period	Households p. year
Germany	Socio-Economic Panel (SOEP)	1983-2013	8501
South-Korea	Korean Labor & Income Panel Study (KLIPS)	1997-2013	4679
Switzerland	Swiss Household Panel (SHP)	1998-2013	3605
UK	British Household Panel Survey (BHPS)	1990-2014	9201
US	Panel Study of Income Dynamics (PSID)	1969-2012	6302

Table 1: General information of household panels of five countries

In order to use the data at hand optimally, four separate latent class models will be estimated. First, I will only use all the American data, which already lasts for over half a century. Then, it will be extended to a cross-country analysis by estimating a model for West-German and US households for the period 1983-2012. In the third analysis, which runs from 1991 onwards, East-German data can be added to the list due to the German reunification. Next to this, the data of the United Kingdom will be included. Lastly, data of all five countries from 1999 until 2012 will be used in the fourth analysis. There are multiple reasons for analyzing these four situations separately. The data at hand is used optimally in the sense that for each country most of the data available is used in one of the four distinct cases. Next to this, it is useful to estimate a model for one country separately. As it is then possible to examine the differences between this model and the cross-country models. By investigating these differences one could encounter if the underlying model is capable of capturing the cross-country heterogeneities. Besides the fact that the US panel is the longest running household panel of all five panels, the US is an interesting country to have a closer look at in the sense that that the level of income inequality has grown drastically in the last few decades. The second analysis also adds value due to some explanatory variables that are

only available for the US and German dataset (e.g. education years and working hours). Hence, incorporating this model makes it possible to analyze the effects of these variables in a cross-country setting. Another reason for zooming in on Germany, UK and US before looking at all five countries is the fact that the underlying household panel surveys (SOEP, BHPS and PSID) are more reliable. For example, lots of previous research on household panel data quality (e.g. panel attrition) has been based on one of these three household panel surveys. Lastly, the estimation of four distinct models gives the possibility to compare the discovered effects of the macroeconomic variables. Thereafter, one is able to give more powerful conclusions if a certain effect is identified in multiple models. In the remainder of this paper I will refer to these four distinct models via analysis 1, 2, 3 and 4. Whereby the same order is used as above, so in each successive analysis the total number of countries will be larger and the time frame will be shorter.

3.1.1 Macroeconomic variables

As was mentioned in the introduction, the effects of three distinct macroeconomic indicators on household income growth will be investigated. The country averages per year of these macroeconomic indicators are plotted in figure 1 up to figure 3. Because the CNEF data also has a variable for the regions within a country where the households are living, regional macroeconomic indicators will be used whenever possible. Regional macroeconomic indicators are preferred, as these indicators give a better reflection of the current economical environment for a certain household than nation-wide indicators would give. Especially for large countries these indicators can be quite divergent among different regions. For example, historical data of regional GDP growth rates collected by the US Bureau of Economic Analysis confirm this finding. The first two macro-economic indicators, which were the level of income inequality and the level of redistribution within a certain region, are measured in a similar fashion as Ostry et al. (2014) did in their paper. Hence, the level of income inequality is determined via the Gini-coefficient of household post-government income. While the level of redistribution is measured by taking the difference in Gini-coefficients of pre-government and post-government income. As historical regional macro-economic data is not available for most of the reviewed countries, the pre- and post goverment Gini-coefficients will be calculated via the Cross-National Equivalent file dataset. In order to ensure a sufficient number of observations per year to calculate reliable Gini-coefficients, Germany, the UK and the US have been divided into ultimately 5 different economical regions. Dividing Switzerland into separate regions was unnecessary, given the size of the country. While information on the province of residence was missing for the Korean data. Unfortunately, the income variables available in the CNEF data are not sufficient for calculating regional GDP-growths. Therefore, national historical GDP-growth rates obtained from the Worldbank database will be used. This is a bit unfortunate though, since it is expected that the variation in GDP-growth levels between different regions within a country will be larger than the variation in Gini-coefficients of these regions.

3.1.2 Microeconomic variables

The dependent variable of the latent class model will be constructed by first differencing the logarithm of the CNEF variable Household Post-Government Income, taking inflation into account. Summary statistics of the real income growth variable can be found in table 9,

while summary statistics of the other microeconomic variables are given per country in table 10 up to 14 of the appendix. The historical levels of inflation within a country are derived from consumer price indexes, which in turn are also gathered from the Worldbank database. A few households characteristics, or transformations of these characteristics will be added to the model as explanatory variables. Please note that I will allow for a fixed household specific effect in the eventual panel model. Hence, the effects of time-invariant household characteristics will be subsumed by the fixed effect parameter and cannot be investigated. As the main focus of this research is on the effects of the macroeconomic indicators, this is not really an issue.

The time-varying household characteristics that will be included in the model are age, child born dummy, child left dummy, education years, education level up dummy and income growth due to working hours difference. Whereby the latter three variables are only included in the first two models with US and German data. Except for the dummy variables of child events, all these variables are originally on a personal level. Different algorithms are used to transform these personal variables to household level variables, which will be explained shortly. In general, the relative importance of each household member for the construction of the eventual household level variable is determined by their labor income. The underlying reasoning is that these characteristics mainly influence labor income growth, while their influence on the growth of other income sources is rather limited. In the methodology section this assumption will be discussed in further detail. In practice, it means that labor income is used as weighting factor for creating the household level variables. The variables age and education years are simply a weighted average over all household members. Whereby the education years are first divided by the average of all working persons in the corresponding country and time period, such that the variable is equivalent across countries and time. If one of the household members who earns at least 15 percent of the total labor income reached a higher educational degree, the variable educational level up dummy will be one for the next two years. Lastly, the variable income growth due to working hours difference is constructed in the following manner, whereby M equals the total number of household members. Note that household members who either started working or ended working in the current period do not contribute to this variable. Because their labor income might be replaced by some governmental income source (e.g. public retirement income or public transfers).

$$\Delta\text{Income due to } \Delta\text{work hours}_t = \frac{\sum_{i=1}^{M}(\Delta\text{work hours}_{it}/\text{hourly wage}_{it-1})}{\text{total household labor income}_{t-1}} \tag{1}$$

3.2 Household weight and household identifier modification

The first thing to note is that using correct household weights in the eventual estimation procedure is of high importance in this case. In the US dataset for example, poorer households are oversampled, probably due to the oversampling of poorer neighborhoods, such that researchers are better able to study poverty dynamics. Hence, an unweighted income growth regression would lead to biased results. Next to this, each household needs three observations to initialize the estimation procedure. This is because the data is first differenced twice, once for the dependent variable and once for estimation. Additionally, one observation is required as instrument for the GMM estimator, Due to this fact, households with more observations

are relatively more influential than households with less observations. To illustrate this, take for example a household with ten observations, eventually seven out of these ten observations will actually be modeled. This 70 percent is a lot more, compared to 25 percent of the observations of another household with only four observations. Assuming that the eventual number of observations of each household is uncorrelated with personal characteristics that influence income growth is unsuitable in this case. For example, individuals who change the composition of their household more often have less observations than individuals who stay with the same partner. The individuals who stay with the same partner are probably older in general, and age influences income growth almost certainly. Therefore, all household weights will be inflated with the following household specific correction factor (cor_i).

$$cor_i = \frac{\sum \text{household weights}_i}{\sum \text{household weights}_i \text{ excluding the 3 non-modeled observations}} \tag{2}$$

After adding this correction factor, the sum of the corrected household weights excluding the three initialization observations will be equal to the sum of the original household weights over all observations for each household, which solves the issue. The only problem that remains is the fact that households with three or less observations will automatically be discarded from the eventual model. Unfortunately, to my knowledge there does not exist any approach to correct for the possible bias induced by the removal of households with less than four observations. In order to do a proper cross-country analysis, all household weights will be multiplied with a population factor after the inflation with the correction factor. This population factor is determined for each year specifically by dividing the countries total population at that time with the sum of the household weights of a country.

As has been explained in section 2, the impact of a household composition change on household income can be quite large and quite heterogeneous among different households. Moreover, the purpose of this research is to find the general effects of macroeconomic variables. Therefore, I will define a household after a composition change as a new household. Whereby a composition change due to the birth of a child or a child who left the house were two exceptions exceptions to this rule. Stated differently, if we remove the observations corresponding to children from the sample, each adjusted household identifier (ID) correspond to a household that consists of the exact same persons over time. Note that this was not necessarily the case for the original household ID's. Due to this adjustment of the household ID variable, the average number of observations per household will become less. Which is a downside of this modification, as more households will have less than four observations and therefore be discarded.

3.3 Handling missing data

There are two type of missing values that need to be imputed before estimating any statistical model. Obviously, the missing variables within a wave due to item nonresponse should be imputed. Next to this, it could be the case that a full wave of data is missing for a respondent. A missing full wave is only imputed for a certain household if it is in-between, or more precisely, if a household has participated in the panel before and after the missing year. Hence, missing waves due to a later start or earlier dropout will not be imputed. Note that in-between missing waves do not necessarily need to be imputed in all cases. However, as already mentioned in section 2, I will use a robust version of the first-differenced GMM

estimator of Arellano and Bond (1991), which cannot be estimated if in-between waves are missing. Next to this, the dependent variable income growth is based on fist-differencing as well and can therefore not be determined if in-between waves are missing.

The missing values will be imputed by means of the k-Nearest-Neighbors algorithm (kNN). The idea of kNN is to find k households (donors) that are most similar to household i and have an observed value for the current variable. Then estimate the missing value of household i based on values of the donors. The interested reader is referred to Kowarik and Templ (2016) for a more comprehensive explanation of the kNN algorithm. Before imputing missing values via kNN, missing values of deterministic variables (e.g. age) will already be imputed via simple logic whenever possible. There are two main reasons for choosing the donor based kNN algorithm and not one of the model based imputation methods. Firstly, for model based imputation one has to assume a certain model specification for each variable that contains missing values. As I am not an expert in this field of research, the risk of inaccurate imputations due to a model misspecification is too high. Next to this, it is possible with kNN to impute only a certain part of the columns of the data matrix. Which is very convenient in this case, since rolling windows of five years will be used to impute the variables corresponding to the third year of the window. In contrast, for model based imputation one has to impute all columns iteratively. Where after statistical models are updated with the new information coming from the imputed values. Hence, imputed values of earlier and later years influence the statistical models and therewith the imputed values of interest (of the third year). This is another downside of model-based imputation compared to kNN, as these imputed values of earlier and later years could be from households that did not participate to the panel anymore at that point in time.

The imputation via kNN will be performed for each country separately to speed up the imputation procedure. The eventual results will be very similar to imputing the full dataset, since it seems plausible that enough good donors will be found within one country. Next to this, rolling windows of five years of data will be used to impute the missing values for the third year of the window. The basic idea behind these rolling windows is that the process of finding the most similar households (donors) is only based on data of the surrounding five years. The missing values of households that have a missing in-between wave at the third year of the window should be based on reliable donors. Therefore, the requirement have been set that these households should have data available in at least one of the two preceding years and in at least one of the two succeeding years. This requirement is only met for a certain household, if it has no more than two consecutive missing in-between years. Hence, households with three or more consecutive missing years are split up into two separate households before the kNN-algorithm is performed. To illustrate the procedure more clearly, suppose there are H_t households that either participated in year t or have a missing in-between wave in year t and that there are k variables. Then, a $H_t \times 5k$ wide-format data matrix will be constructed, with all variables of the periods $t-2, ..., t+2$ as columns. Distance calculation for finding the best donors will be based on all columns, while only the values of the middle k columns that match year t will be imputed. Eventually, this process will be executed for each year. Lastly, note that the CNEF data can be split up into household level and personal level variables. The imputation via kNN will be done separately for these two types of variables. The variables of persons that have a missing wave at the current year of interest, will only be imputed if they correspond to a household

that have a missing in-between wave in the current year.

To obtain standard errors of the model parameters that take into account the uncertainty of the imputed values one has to make use of bootstrapping. Hence, the idea is to take R bootstrap samples from the original (non-imputed) data, impute all these datasets via kNN and then estimate the latent class model for each imputed dataset separately. Note that a bootstrap sample of the households will be taken, such that the estimation process remains feasible. The eventual model parameter estimates will be set equal to the mean of the bootstrap replicates, while the standard errors of the model parameters are simply determined by means of the standard errors of the bootstrap replicates. In order to do proper significance testing, one has estimate a latent class model for at least a few hundred bootstrap replicates. This is the major downside of using kNN compared to model based imputation. Because model based imputation allows the researcher to perform multiple imputation for obtaining correct standard errors, which requires a lot less replicates.

The last missing values that need to be imputed before any statistical analysis can be performed are the uneven years of the US dataset from 1997 onwards. Because the PSID has only conducted surveys for even years from 1996 onwards. Note that another possibility would be to transform everything to a biennial setting instead of the annual setting that is now used. The main problem of using biennial data is the fact that a lot of households will be discarded from the model. As has been explained in section 3.2, households with less than four observations cannot be modeled. Hence, if one would transform to a biennial setting, all households with less than eight years of data would be discarded from the model. Besides the high efficiency loss one would get from discarding all this data, it would give more biased results compared to the annual setting as well.

Because columns of the US wide format data matrix corresponding to the uneven years from 1997 onwards are completely filled with missing values, it is not possible to perform kNN to impute the missing values. Therefore, linear interpolation will be used to impute the missing uneven years of the US dataset. If one would simply use the mean value of the two surrounding even years, the volatility level of each variable would become much lower after 1997, which is quite unrealistic. In order to ensure that the volatility levels are equally high over all time periods, a random error term will be added to the linearly interpolated values. Generating distinct random errors for each bootstrap sample will ensure that the randomness of the imputed values will automatically be reflected in the standard errors of the model parameters. Note that this random error term will not be added to the deterministic variables, such as age. Next to this, most of the income variables of a particular source are mixed in the sense that they are discrete at the value zero. The linear interpolation algorithm of the mixed variables takes this feature into account.

The variable pre-government income is missing for the Korean data of 1999, 2000 and 2001 due to missing tax information. Next to this, post-government income values are missing for UK households of 2006 and 2007. These post-government income values are imputed via linear interpolation as well. For the Korean data this was not possible though, as 1999 is the first year for which there is Korean data. Therefore, a model for the total amount of tax that has to be paid is estimated, using labor income with different tax boxes and all other income sources as explanatory variables. Then this model is used to predict the amount of

tax paid for each household in 1999, 2000 and 2001, from which household pre-government income can be deducted. Lastly, note that the Gini-coefficients will be calculated after imputing the missing values. Because it has been regularly highlighted in the literature that the probability of a missing value for an income variable is higher for households with a low income. Hence, calculating the coefficients before imputing the missing income variables would yield a biased result.

4 Methodology

4.1 Basic model

A finite mixture model will be estimated in order to find possible heterogeneity in the macroeconomic drivers of household income growth. More precisely, the following model will be estimated, whereby the set $\{1, 2, ..., S\}$ are the segments of the mixture model.

$$\Delta Inc_{it} = \mu_i + \alpha_{s_i} \Delta Inc_{it-1} + x'_{it-1}\beta_{s_i} + v'_{it-1}\lambda + \varepsilon_{it} \tag{3}$$

$$\text{With } s_i \in \{1, 2, ..., S\} \text{ and } P(S_i = s) = p_s \text{ for } s = 1, ..., S \text{ with } \sum_{s=1}^{S} p_s = 1 \tag{4}$$

where

ΔInc_{it} = real growth in disposable income of household i at time t ($\log(Inc_{it}) - \log(Inc_{it-1}) - \text{inflation}_t$).
μ_i = Household specific time-invariant fixed effect parameter.
x_{it-1} = Macro-economic and government policy related variables that might be heterogeneous across different segments.
v_{it-1} = Socio-economic and demographic characteristics of household i

Recall that the panel data set is unbalanced. Hence, the model given above runs from $i = 1, ..., N$ and $t = T_i^F, ..., T_i^L$, with T_i^F and T_i^L respectively the first and last year for which there is data available of household i. Please notice that T_i^F corresponds to the fourth observation of household i in the original dataset. This is due to the fact that the first three observations are required for calculating income growth (first differencing), including one lag of income growth in the model and using a first differenced estimator. Furthermore, note that a household specific fixed effect parameter is added to model. By adding this parameter, one allows for an unexplainable part in the average income growth of a certain household. Hence, this parameter could be seen as the average growth of 'skills' of the household members. Other fixed effects parameters that could have been added to this multi level panel data model are a time specific effect υ_t, a country specific effect ω_j and the cross-product terms $\mu_i\upsilon_t$, $\mu_i\omega_j$ and $\upsilon_t\omega_j$. There are different reasons for leaving these parameters out of the model. Firstly, the time and country specific effect parameters and its cross-product term are already reflected by the variable GDP-growth. Since the viewed households stay within one country over time, the cross-product $\mu_i\omega_j$ is equivalent with the household specific parameter μ_i. Hence, adding this cross-product would not give any benefits to the model. Lastly, the cross-product $\mu_i\upsilon_t$ does not make much sense for this model

specification. Because adding this cross-product term means that one would assume that all household specific fixed effect parameters experience an equal multiplicative transformation at each point in time.

To estimate the model properly, first differences of the model equation will be used as input. Hence, in order to simplify notation later on, it is convenient to replace ΔInc_{it} by y_{it}, which results in the following equation.

$$y_{it} = \mu_i + \alpha_{s_i} y_{it-1} + x_{it-1}^{'} \beta_{s_i} + v_{it-1}^{'} \lambda + \varepsilon_{it} \tag{5}$$

In order to specify the covariance structure of the error terms, it is useful to stack the observations of each household. Hence, $y_i^{'} = (y_{iT_i^F}, ..., y_{iT_i^L})$, $y_{i,-1}^{'} = (y_{iT_i^F-1}, ..., y_{iT_i^L-1})$ and $X_i^{'} = (x_{iT_i^F-1}, ..., x_{iT_i^L-1})$, also the matrix V_i and the vector ε_i are constructed in a similar fashion. Lastly, μ_i is now a vector filled with the fixed household effect parameter. These stacked observations lead to the following model equation.

$$y_i = \mu_i + \alpha_{s_i} y_{i,-1} + X_i \beta_{s_i} + V_i \lambda + \varepsilon_i \text{ with } \varepsilon_i \sim \mathcal{N}(0, \Sigma_i) \tag{6}$$

In order to estimate the model properly, the covariance matrix Σ_i should be diagonal. Hence, it is assumed that the errors within households are possibly heteroskedastic but serially uncorrelated. However, note that this heteroskedasticity is only modeled implicitly through the down weighting process of observations with large residuals in the outlier robust estimation method. Hence, only one overall volatility parameter σ_s per segment will be estimated. In the remainder of this paper the distribution of the model errors will be written as $\varepsilon_i \sim \mathcal{N}(0, \sigma_{s_i}^2 I_i)$, in order to simplify notation. Next to this, it is assumed that the model errors between households are uncorrelated, so $E(\varepsilon_{it}\varepsilon_{js}) = 0$ for all $i \neq j$ and all t, s.

4.2 Building blocks of disposable income growth

The effects on household disposable income growth that will be estimated by the parameters are not direct effects. These effects are not direct in the sense that the disposable income of a household is build upon different income sources, which in turn are affected by the explanatory variables. Despite the fact that the CNEF data divides household income into six different sources, namely labor, asset, public and private retirement income and public and private transfers. I will further elaborate upon the distinction between labor and non-labor income. If we assume for simplicity that the proportion of paid tax is equal for both income sources, then the following equality holds true.

$$\Delta\text{Disposable income}_{it} = w_{1it}\Delta\text{Labor income}_{it} + w_{2it}\Delta\text{Non-labor income}_{it} \tag{7}$$

Whereby weights w_{1it} and w_{2it} are equal to the proportions of respectively labor and non-labor income at time $t-1$. Further note that the delta's in the equation refer to income growths instead of differences. If we plug the equality given above into the model equation, we get the following model.

$$y_{it} = w_{1it}(\mu_{1i} + \alpha_{1s_i} y_{it-1} + x_{it-1}^{'} \beta_{1s_i} + v_{it-1}^{'} \lambda_1) \tag{8}$$

$$+ \quad w_{2it}(\mu_{2i} + \alpha_{2s_i} y_{it-1} + x_{it-1}^{'} \beta_{2s_i} + v_{it-1}^{'} \lambda_2) + \varepsilon_{it} \tag{9}$$

In order to deduce the proposed model equation given in equation 5 from the model equation with a division in labor and non-labor income, one has to make a few extra assumptions. Firstly, the weighting factors w_{1it} and w_{2it}, since these are time specific, the eventual household specific fixed effect parameter $w_{1it}\mu_{1i} + w_{2it}\mu_{2i}$ will become time specific. Which in turn makes it impossible to and estimate the model via first-differencing. Therefore, household specific labor income weights w_{1i} and w_{2i} will be used, which are obtained by calculating the average weight over time. Next to this, one has to assume that $w_{1i}\alpha_{1s} + w_{2i}\alpha_{2s} = \alpha_s$ and $w_{1i}\beta_{1s} + w_{2i}\beta_{2s} = \beta_s$. Note that without proper segmentation of the households, these two equalities are very unlikely to hold. However, the fact that the eventual disposable income of a certain household is built upon different shares of income sources, is one of the main underlying reasons for choosing a latent class model with segment specific effects. Hence, the idea is that the division of households will lead to homogeneous segments, in the sense that the disposable income of households within a certain segment is mainly built upon similar income sources. In that case the two equalities for the segment specific parameters are more likely to hold. Lastly, one has to make a certain assumption for the non-segment specific parameters λ_1 and λ_2. Recall from the data section that the following household characteristics correspond with these parameters: age, child born dummy, child left dummy, education years, education level up dummy and income growth due to working hours difference. In general these household characteristics are mainly used in the literature to explain the level of labor income growth of a individual or household. On the other hand, one can conclude via logical thinking that their influence on the other five income sources is probably small. An exception to this finding are the two education variables, which might have a positive influence on asset income growth. Moreover, the child dummies might influence the growth in public transfers, for example via governmental child benefits. Note that age only influences public and private retirement income growth via a great peak at the point in time when the retirement income starts. But beyond that, it will not have a major influence on these income sources. Therefore it seems reasonable to assume that the non-segment specific parameter vector for non-labor income (λ_2) is equal to zero. If we use these three assumptions, the divided labor and non-labor income model equation will be simplified in the following manner.

$$y_{it} = w_{1i}\mu_{1i} + w_{2i}\mu_{2i} + \alpha_{s_i} y_{it-1} + x_{it-1}^{'}\beta_{s_i} + w_{1i}v_{it-1}^{'}\lambda_1 \tag{10}$$

Now the only difference between the model equation given in 5 and the equation given above is the labor income weight in front of v_{it-1}. Hence, in the eventual model all household characteristics (v_{it-1}) are multiplied with a household specific weight, that reflects the average proportion of labor income from the total amount of income. The necessity of this multiplication comes from the assumption that these household characteristics only have an influence on disposable income growth via labor income growth.

4.3 Expectation-maximization algorithm

The finite mixture model will be estimated by means of the Expectation-Maximization algorithm (EM algorithm), whereby an outlier robust GMM estimator will be used in the maximization step. Firstly, I will discuss the EM algorithm and show that by performing the expectation step, the resulting model is a dynamic linear panel model that is suitable for the outlier robust GMM estimator. Thereafter, the outlier robust GMM estimation process will be discussed in further detail.

Before explaining the EM algorithm, it is convenient to rewrite the basic model into its first-differences, as the estimator in the maximization step will be based on first differences. Let us define the first difference operator matrix D_i, which is a matrix that contains the values $\{-1, 0, 1\}$ and transforms equation 6 to a first difference equation. Hence, the following two equations are equivalent.

$$y_i - y_{i,-1} = (y_{i,-1} - y_{i,-2})\alpha_{s_i} + (X_i - X_{i,-1})\beta_{s_i} + (V_i - V_{i,-1})\lambda + \varepsilon_i - \varepsilon_{i,-1} \tag{11}$$

$$D_i y_i = D_i y_{i,-1}\alpha_{s_i} + D_i X_i \beta_{s_i} + D_i V_i \lambda + D_i \varepsilon_i \tag{12}$$

The likelihood function of the proposed finite mixture model in first differenced format is then given by.

$$L(\theta) = \prod_{i=1}^{N} \sum_{s=1}^{S} p_s \phi(D_i y_i; D_i \zeta_{is}, \sigma_s^2 I_i) \tag{13}$$

$$\text{with } \zeta_{is} = \alpha_s y_{i,-1} + X_i \beta_s + V_i \lambda \tag{14}$$

Whereby $\phi(D_i y_i; D_i \zeta_{is}, \sigma_s^2 I_i)$ is equal to probability density level of observations $D_i y_i$, given that they are multivariate normally distributed with mean $D_i \zeta_{is}$ and covariance matrix $\sigma_s^2 I_i$. The logarithm of this likelihood function has multiple local maxima and is ill behaved in general. Therefore, maximum likelihood estimation by means of numerical optimization does not work smoothly and should be avoided. Instead, one should ease the maximization step by making use of the EM algorithm. The EM algorithm consists of two steps which are repeated iteratively until convergence. The two steps are the expectation and the maximization step (E-step and M-step), which will both be discussed in further detail.

Let us denote the set of model parameters as $\theta = \{\sigma_s^2, \alpha_s, \beta_s, p_s(s = 1, ..., S), \lambda\}$. In each iteration, all model parameters are estimated in the maximization step. Denote the estimated model parameters of the previous iteration as $\hat{\theta}_{m-1}$. The idea of the E-step is to take the expectation of the log complete data likelihood function with respect to s given y and the current estimate $\hat{\theta}_{m-1}$. The complete data likelihood function and the log complete data likelihood function are given by:

$$L_c(\theta) = \prod_{i=1}^{N} \prod_{s=1}^{S} (p_s \phi(D_i y_i; D_i \zeta_{is}, \sigma_s^2 I_i)^{I(s_i = s)} \tag{15}$$

$$l_c(\theta) = \sum_{i=1}^{N} \sum_{s=1}^{S} I(s_i = s)(\log(p_s) + \log\big(\phi(D_i y_i; D_i \zeta_{is}, \sigma_s^2 I_i)\big)) \tag{16}$$

The E-step requires us to calculate $\mathbb{E}_s(l_c(\hat{\theta}_{m-1})|y)$. Note that the only stochastic components in this expectation are $s_i (i = 1, ..., N)$. Hence, by using the rules of conditional probabilities, these probabilities can be rewritten as follows.

$$p_{is} = \mathbb{E}(I(s_i = s)|y_i) = 1 \times P(s_i = s|y_i) + 0 \times P(s_i \neq s|y_i) = P(s_i = s|y_i) \tag{17}$$

$$P(s_i = s|y_i) = \frac{f(y_i, s_i = s)}{f(y_i)} = \frac{f(y_i|s_i = s)p_s}{\sum_{k=1}^{S} f(y_i|s_i = k)p_k} = \frac{\phi(D_i y_i; D_i \zeta_{is}, \sigma_s^2 I_i)p_s}{\sum_{k=1}^{S} \phi(D_i y_i; D_i \zeta_{ik}, \sigma_s^2 I_i)p_k} \tag{18}$$

Hence, given all model parameters $\hat{\theta}_{m-1}$, the conditional probabilities p_{is} can easily be calculated for all households i and all segments s. Thereafter, the expected value of the log complete data likelihood function can be calculated as follows.

$$\mathbb{E}_s(l_c(\theta)|y) = \sum_{i=1}^{N}\sum_{s=1}^{S} p_{is}\log(p_s) + \sum_{i=1}^{N}\sum_{s=1}^{S} p_{is}\log\big(\phi(D_i y_i; D_i\zeta_{is}, \sigma_s^2 I_i)\big)\big) \tag{19}$$

This expected value should now be maximized by updating the model parameters in an optimal manner. Hence, the M-step can be summarized as follows.

$$\max_{\theta}\Big(\sum_{i=1}^{N}\sum_{s=1}^{S} p_{is}\log(p_s) + \sum_{i=1}^{N}\sum_{s=1}^{S} p_{is}\log\big(\phi(D_i y_i; D_i\zeta_{is}, \sigma_s^2 I_i)\big)\big)\Big) \tag{20}$$

Because there are no overlapping parameters in the first and second part of the optimization function, these can be maximized separately. Maximizing the first part is straightforward and gives the following updated parameters for p_s.

$$\hat{p}_s = \frac{1}{N}\sum_{i=1}^{N} p_{is} \tag{21}$$

The maximization process of the second part has become much easier as well, as it now disintegrates to the optimization of a dynamic linear panel model with N times S 'households'. Or stated differently, maximizing the second part of equation 20 is equivalent with estimating the parameters of the following linear dynamic panel model via a first differenced estimator.

$$y_i = \mu_i + \alpha_s y_{i,-1} + X_i\beta_s + V_i\lambda + \varepsilon_{is} \text{ with } \varepsilon_{is} \sim \mathcal{N}(0, \sigma_s^2 I_i) \tag{22}$$

For $i = 1, ..., N$ and $s = 1, ..., S$. Whereby each observation should be weighted with the square root of the probability of selection. This probability of selection is equal to the multiplication of the probability that household i belongs to segment s (p_{is}) and the household weight of household i, which is given in the dataset. Note that these household weights are not yet introduced into the model equation, since the model needs to be first differenced later on for estimation. Hence, each first differenced time-period will then be weighted with the average of the two corresponding household weights. In the remainder of this section, I will use h_{its} for referring to this square root of the probability of selection of estimation period Δt.

The GMM estimator that has been proposed by Arellano and Bond (1991) is frequently used for estimating the type of model given in equation 22. Which is due to the fact that it gives consistent and efficient estimates, even when the number of observations per household is small. However, this estimator is not robust against anomalous observations, which are almost certainly present in household income data according to the literature. Therefore, I will make use of a GMM estimator that is robust against outliers in the maximization step, which has been proposed by Lucas et al. (1997).

Please note that the EM-algorithm might converge to a local optimum of the likelihood function. Therefore, one should use multiple random starting values for the EM algorithm.

Because finding good starting values for the model parameters can be quite difficult, it is convenient to start the algorithm by randomly drawing values for probability that household i belongs to segment s. For each household, these segment probabilities will be drawn from a Dirichlet distribution with parameters $\alpha_1, ..., \alpha_S$ that are derived from a uniform distribution. To be more precise, each parameter is derived as follows $\alpha_s = \exp(u_s)$ with $u_s \sim U(\log(1/3), \log(3))$. The advantage of using a Dirichlet distribution is that the obtained probabilities add up to one, so they can be used directly. Moreover, by varying the parameters of the Dirichlet distribution over different households, one ensures a high variability level in the obtained starting probabilities.

4.4 Outlier robust generalized method of moments estimation

As has just been mentioned, the outlier robust GMM estimator of Lucas et al. (1997) fits the estimation problem at hand perfectly. In this paper I will discuss the estimation process and some other main findings of the researchers. Hence, the interested reader is referred to Lucas et al. (1997) for more details (e.g. proves of the asymptotic properties of the estimator). The starting point is the dynamic linear panel model given in equation 22. Rearranging some terms in this equations leads to the following model.

$$y_i = \mu_i + Z_i\gamma_s + V_i\lambda + \varepsilon_{is} \text{ with } \varepsilon_{is} \sim \mathcal{N}(0, \sigma_s^2 I_i) \tag{23}$$

$$\text{with } Z_i = (y_{i,-1}, X_i) \text{and } \gamma_s' = (\alpha_s, \beta_s') \tag{24}$$

4.4.1 Instrument specification

The outlier robust GMM estimator will be computed by performing instrumental variable (IV) estimation on the first differences of the model given in equation 23. Which has proven to be the best solution for obtaining unbiased results in a linear dynamic panel setting. For this reason it is convenient to define instrumental variable matrix W_i. First, let us define T_i as the eventual number of model equations of household i, so T_i is equal to $T_i^L - T_i^F + 1$. To illustrate, if a household has data in the first 6 succeeding years, then $T_i^F = 4$, $T_i^L = 6$ and $T_i = 3$. Stated differently, the income growth of the fourth, fifth and sixth year can be modeled, which results in $T_i = 3$. Continuing with the matrix of instrumental variables W_i, which is given in equation 25. One of the explanatory variables in the matrix $D_i Z_i$ was the lagged and first differenced income growth variable $D_i y_{i,-1}$, which is obviously correlated with the lagged error term $\varepsilon_{i,-1}$. Therefore, one needs to resort to instrumental variable (IV) estimation. Under the assumption that the income growth processes behave dynamically, in other words $\mathbb{E}(\alpha_s) \neq 0$ for $s = 1, ..., S$. All previous values of the income growth variable can be used as instruments for the endogenous explanatory variable Δy_{it-1}. Next to this, one can use the other covariates X_i and V_i as instrumental variables. Which time periods to use as instrumental variables depends upon the exogeneity assumptions of the covariates. In general, one can assume three distinct cases for each variable, which are an endogenous, a predetermined or a strictly exogenous covariate (Bond, 2002). For this three assumptions respectively, one can use the covariates as instruments up to time $t-2$, $t-1$ or T_i^L, so in the latter case, one can use the values of all time periods as instruments. The assumption that will be used for each variable will be based on previous literature or simply common sense. For example, several papers have been written on the possible endogeneity of the covariate education level when it is used to explain income. (Block et al., 2010; Blackburn

and Neumark, 1993). On the other hand, common sense can be used to determine that the variable age is strictly exogenous, since it is definitely deterministic. However, for certain other variables, the choice of which assumption to use might be more difficult. Therefore, the validity of the moment conditions will be tested afterwards and misspecified assumptions will be changed accordingly. Now suppose that the number of covariates for which we have assumed endogeneity, predetermined and strict exogeneity are respectively k_1, k_2 and k_3. Then this gives a $T_i \times M$ instrumental variable matrix W_i, whereby the number of moment conditions M is equal to $(\tau+1)\tau/2+k_1(\tau+1)\tau/2+k_2((\tau+2)(\tau+1)/2-1)+k_3(\tau(\tau+3))$. With τ the maximum possible value for T_i, so τ corresponds to the number of model equations of a household that has data for all years that are viewed. The resulting instrumental variable matrix W_i is then defined as follows.

$$W_i = \begin{bmatrix} 0^{F_{1i}} & y_i^{T_i^F-2} & 0 & \dots & 0 & 0^{L_{1i}} & 0^{F_{2i}} & U_i^{T_i^F-2} & 0 & \dots & 0 & 0^{L_{2i}} \\ 0^{F_{1i}} & 0 & y_i^{T_i^F-1} & \dots & 0 & 0^{L_{1i}} & 0^{F_{2i}} & 0 & U_i^{T_i^F-1} & \dots & 0 & 0^{L_{2i}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0^{F_{1i}} & 0 & 0 & \dots & y_i^{T_i^L-2} & 0^{L_{1i}} & 0^{F_{2i}} & 0 & 0 & \dots & U_i^{T_i^L-2} & 0^{L_{2i}} \end{bmatrix} \tag{25}$$

Whereby, $0^{F_{1i}}$ and $0^{L_{1i}}$ are zero vectors of length $(T_i^F-3)(T_i^F-4)/2$ and $(\tau+1)\tau/2-(T_i^L-2)(T_i^L-3)/2$, which correspond with the missing years for household i at the beginning and at the end respectively. Note that these vectors have both length equal to zero if household i was part of the survey in all years, or stated differently if $T_i = \tau$. Next to this, y_i^t is defined as $(0, ..., 0, y_{iT_i^F-2}, ..., y_{it})$, with $T_i^F - 4$ number of zeros at the beginning of the vector, which correspond to the number of beginning years for which data is missing for household i. Then, $0^{F_{2i}}$ and $0^{L_{2i}}$ are both zero vectors as well, that serve the same purpose as $0^{F_{1i}}$ and $0^{L_{1i}}$. Lastly U_i^t consists of all endogenous, predetermined and strictly exogenous covariates up to time $t-2$, $t-1$ and T_i^L respectively. Furthermore, it is also filled with zeros at the beginning, just as for y_i^t. On top of this, there are zeros at the end of each exogenous variable, which correspond to missing exogenous covariates in the ending years for which there is no data of household i.

4.4.2 Moment conditions

We are now able to set up the moment conditions that eventually will give us the outlier robust GMM estimator. Let γ_{0s} and λ_0 denote the true parameter vectors for the first differenced version of model 23. Define $e_{its}^0 = e_{its}(\gamma_{0s}) = \Delta y_{it} - \Delta z'_{it-1}\gamma_{0s} - \Delta v_{it-1}\lambda_0$ and define w_{it} as the vector of instrumental variables to be used for time period t of household i. Which is equal to the r-th row of matrix W_i, if r is defined as $r = t - T_i^F + 1$. Then the moment conditions for the outlier robust GMM estimator are defined as.

$$\mathrm{E}\Big(\sum_{t=T_i^F}^{T_i^L} w_{it}\phi_{its}h_{its}e_{its}^0\Big) = 0 \tag{26}$$

$$\text{where } \phi_{its} = \begin{cases} v_t(w_{it})\sigma_s\psi(e_{its}^0/\sigma_s)/e_{its}^0, & \text{if } e_{its}^0 \neq 0 \\ v_t(w_{it}), & \text{otherwise} \end{cases} \tag{27}$$

In the definition given above, $v_t()$ and $\psi()$ are both real valued functions, which will be further specified in the next paragraph. Note that the expected value of the moment conditions given in equation 26 consist of two parts. Firstly, there is the regular GMM orthogonality condition which imply that the instrumental variables w_{it} should be uncorrelated with the error terms e^0_{it}. Next to this, each observation has its own weight ϕ_{its}, which makes the estimator less sensitive to anomalous observations. From this point of view, it is easy to see that the regular GMM estimator proposed by Arellano and Bond (1991) is a special case of this outlier robust GMM estimator, which can be obtained by setting ϕ_{its} equal to one for all observations. The function for the observation weights will now be discussed in further detail. Firstly, it is convenient to rewrite the population function given in equation 27 into an actual data format, which is defined as.

$$\hat{\phi}_{its} = \begin{cases} v_t(w_{it})\hat{\sigma}_s\psi(e_{its}/\hat{\sigma}_s)/e_{its}, & \text{if } e_{its} \neq 0 \\ v_t(w_{it}), & \text{otherwise} \end{cases} \tag{28}$$

The weight factor $\hat{\phi}_{its}$ consists of two parts. The first part ($v_t(w_{it})$) is used to reduce the effect of divergent or persuasive observations in the space of the instrumental variables. While the remainder of the function depends on the error terms and serves to shrink the weight of anomalous observations with large residuals. The weighted median absolution deviation (weighted MAD) will be used to estimate the segment specific volatility parameters, since it is a consistent and outlier robust estimation method. It is defined as follows.

$$\hat{\sigma}_s = 1.4826 \text{ weighted median}_{i,t}|\hat{e}_{its} - \text{weighted median}_{i,t}(\hat{e}_{its})| \tag{29}$$

The volatility parameters will be estimated iteratively with the parameter vector. In the first iteration of the EM algorithm, this procedure will be started by using the instrumental variable weights as starting weights, so $\hat{\phi}_{its} = v_t(w_{it})$. In the subsequent iterations of the EM algorithm, final weights of the previous round will be used as starting weights for this procedure. After estimating the outlier robust GMM estimator, I will perform a sensitivity analysis for these starting weights by reestimating the GMM estimator with all starting weights equal to one and by using starting weights that are drawn from a uniform distribution with minimum zero and maximum one. Next to this, the following starting weights will be tried as well.

$$\hat{\phi}_{its} = \begin{cases} 1, & \text{if } m_k - 3d_k \leq q_{itk} \leq m_k + 3d_k \quad \forall k \\ 0, & \text{otherwise} \end{cases} \tag{30}$$

Whereby, $q_{it} = (\Delta y_{it}, \Delta z_{it-1}, \Delta v_{it-1})$ and m_k and d_k are respectively the weighted median and the weighted MAD of q_k, with $q_k = (q_{11k}, ..., q_{itk}, ..., q_{NT_N-1})$. Hence, an observation gets a non-zero weight value assigned if it does not contain any univariate outlier in the dependent variable or one of the regressors. By performing this sensitivity analysis, one can investigate to which extend the locally robust GMM estimator is globally robust. However, do note that it might be the case that the results of all four methods are affected by outliers that receive a positive starting weight. Hence, from this sensitivity analysis one cannot simply conclude global robustness of the eventual results. Therefore, it might be of interest for further research to investigate more advanced methods for these starting weights. For example, one could use starting weights drawn from a uniform distribution, whereby the bandwidth depends upon q_{it}.

4.4.3 Calculating outlier weights

In order to finalize the estimator, the functional forms of the weight functions $\psi()$ and $v_t()$ need to be specified. Lucas et al. (1997) use a redescending function $\psi()$ which makes use of a fifth degree polynomial $\tilde{\psi}(x)$ such that $\psi()$ is twice continuously differentiable. They specify the weighting function as follows.

$$\psi_{c_1,c_2}(e) = \begin{cases} e, & \text{if } |e| \leq c_1 \\ \text{sign}(e)\tilde{\psi}(|e|), & \text{if } c_1 < |e| \leq c_2 \\ 0, & \text{otherwise} \end{cases} \tag{31}$$

Note that for given values of c_1 and c_2, there is exactly one fifth degree polynomial specification such that $\psi()$ is twice continuously differentiable. Because a fifth degree polynomial has six free parameters, while the condition of twice continuously differentiability gives us 6 restrictions in total. Hence, the parameters of the fifth degree polynomial can be found by solving a linear system of equations. As it is assumed that the error terms are normally distributed, I will use $\sqrt{\chi_1^{-2}(0.975)}$ and $\sqrt{\chi_1^{-2}(0.9975)}$ as values for the breakpoints c_1 and c_2 respectively. Hence, observations with an error value that occurs with probability less than 0.25 percent are discarded from the model. Note that the chosen borders of 0.975 and 0.9975 are a bit more conservative than the borders of 0.99 and 0.999 used by Lucas et al. (1997). This choice is based upon the stylized fact that income data is heavy tailed, which has been investigated by Guvenen et al. (2015). Therefore, I chose to follow many researchers in robust statistics and already start down weighting observations in the 2.5 percent outer part of the distribution. Continuing with the specification of the $v_{tN}()$ function, which should down weight the instrumental variables that are far removed from most of the instrumental variables. For this reason, Lucas et al. (1997) propose to make use of the Mahalanobis distance of the instrumental variables, which is defined as.

$$\delta(w_{it}, m_t, V_t) = \sqrt{(w_{it} - m_t)' V_t^{-1} (w_{it} - m_t)} \tag{32}$$

Whereby, m_t and V_t are the location vector and covariance matrix of the instruments vectors at time t, which obviously need to be estimated in a robust manner. Due to the fact that the panel data is unbalanced, the estimation process of the Mahalanobis distances had to be extended. For a given time period t, columns of the instrumental variable matrix corresponding to the instrument values of the most recent period will be completely filled. However, once we go back in time for k periods, the number of zero values will increase, These zero values correspond to households who participated in the household panel at time t, but did not yet participate at time $t - k$. If one would estimate one covariance matrix for each time period only, these zero values would be included in the estimation process. As a result, households that have participated for a longer period have a higher probability of receiving a zero weight due to outlying instruments. Moreover, the percentage of zero instrument weights will become higher at later periods. Therefore, one has to iterate through all time periods up to the current time period t. Then, for each time period (say time period k), the Mahalanobis distance of the households for which the time period k is their first period will be calculated. These Mahalanobis distances are based upon a mean vector (m_{kt}) and a covariance matrix (V_{kt}) that are estimated by only using the households that already participated at time period k. Note that this estimation process only solves

the issues given above for endogenous or predetermined instrument variables. Which is not an problem in this case, since there are no potentially exogenous instrument variables that contain outlying values.

As was already noted by Lucas et al. (1997), it is important to use a computationally effective estimation method. Because one has to estimate multiple covariance matrices for each time period and the total number of instruments can become quite large. Therefore they propose to use a slight modification of the S-estimator, which makes use of the fact that the instrumental variables for each time period are linear combinations of all observations for a certain household. However, nowadays it is much more attractive from a computational point of view to estimate the outlier robust Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1985). This gain in attractiveness is due to the more recently proposed Deterministic Algorithm (DetMCD) by Hubert et al. (2012), which estimates the MCD estimator in a efficient manner. The only downside of this estimator is that it is not a fully affine equivariant estimator, though it is very close to affine equivariant. However, note that the modification of the S-estimator used by Lucas et al. (1997) also lost its affine equivariance property, due to the usage of the linear combinations.

Continuing with the weight function for the instrumental variables, if we define $\hat{\delta}_{it} = \delta(w_{it}, \hat{m}_{kt}, \hat{V}_{kt})$, with $\hat{m}_{kt}$ and $\hat{V}_{kt}$ robust estimates of m_{kt} and V_{kt}. Then the down weighting function for the instrumental variables is defined as $v_t(w_{it}) = \psi_t(\hat{\delta}_{it}^2)/\hat{\delta}_{it}^2$ if $\hat{\delta}_{itN} \neq 0$ and $v_{tN}(w_{it}) = 1$ otherwise. Whereby $\psi_t()$ make use of the same fifth degree polynomial function as defined in equation 31, but with tuning parameters $c_{1t} = \chi_{p_t}^{-2}(0.99)$ and $c_{2t} = \chi_{p_t}^{-2}(0.999)$, with $p_t = \text{rank}(\hat{V}_{kt})$. Note that these tuning parameters are time specific, since the number of valid instruments increases over time. Just as for the error terms, the down weighting function for the instrumental variables is build upon the assumption that the instrumental variables are multivariate normally distributed. For this reason not all instruments will be used to calculate the instrument weighting factor. To be more precise, the three dummy variables and the variables age and education years are obviously not normally distributed. Because of this, the algorithm will down weight observations which should not be down weighted. For example all households consisting of aged people would be seen as outlying observations if the variable age would be included. Instead, the variables age and education years can easily be checked for irregular patterns beforehand. During this inspection, I could not find any irregular pattern for these variables. Lastly, the instruments corresponding to the macro-economic variables are also not used for calculating the instrumental weighting factors, as these variables are equal for all household within a certain region at each point in time. Hence, the instruments corresponding to previous income growth and previous income growth due to working hours difference values are eventually used as input for calculating the instrumental weights.

4.4.4 Estimator and its asymptotic properties

Now that the moment conditions are fully specified, we can turn our interest to the actual outlier robust GMM estimator. Because there are more moment conditions than parameters to be estimated, we need to minimize the quadratic form of the moment conditions, which

is defined as follows.

$$J_N = \min_{\gamma,\lambda} \Bigg(\Big(S^{-1}N^{-1} \sum_{i=1}^{N} \sum_{s=1}^{S} W_i' \Phi_{is} H_{is} D_i (y_i - Z_i\gamma_s - V_i\lambda) \Big)' A_N \tag{33}$$

$$\Big(S^{-1}N^{-1} \sum_{i=1}^{N} \sum_{s=1}^{S} W_i' \Phi_{is} H_{is} D_i (y_i - Z_i\gamma_s - V_i\lambda) \Big) \Bigg) \tag{34}$$

Whereby Φ_{is} and H_{is} are $T_i \times T_i$ diagonal matrices with $\hat{\phi}_{its}$ and h_{its} as elements on the diagonal respectively. Thus, Φ_{is} is the weighting matrix that down weights outliers, while H_{is} is the weighting matrix of the data, which is obtained by the multiplication of the segment probability p_{is} and the household weights of household i. Next to this, A_N is a $M \times M$ positive definite weighting matrix. Then it is convenient to write down this optimization problem in matrix form. First I will define the vectors and matrices that correspond to one segment and then I will extend these matrices to all segments. Hence, let us define $W' = (W_1', ..., W_N')$ and $\Phi_s = \text{ diag}(\Phi_{1s}, ..., \Phi_{Ns})$. Then define the vector y and the matrices Z and V in a similar fashion as the matrix W. Lastly, the matrices H and D are created in exactly the same way as the matrix Φ_s. Because the matrices W, D and V and the vector y are the same over all segments, they can simply be replicated S times. Hence, $W_{total}' = (W', ..., W')$, $D_{total} = \text{ diag}(D, ..., D)$ and the matrix V_{total} and the vector y_{total} is created in a similar fashion as W_{total}. Next, the segment dependent matrices Φ_{total}, H_{total} and Z_{total} are created in the following manner, $\Phi_{total} = \text{ diag}(\Phi_1, ..., \Phi_S)$ and $Z_{total} = \text{ diag}(Z, ..., Z)$ and the matrix H_{total} can be constructed similarly as Φ_{total}. Note that the matrix Z_{total} has to be a diagonal matrix with elements Z due to fact that the corresponding parameters γ_s are segment specific. Lastly, it is convenient to combine the explanatory variable matrices and parameter vectors. Hence, let us define $C_{total} = (Z_{total}, V_{total})$ and $\kappa' = (\gamma_1', ..., \gamma_S', \lambda')$. Then basic linear algebra can be used to show that the optimal parameter vector is equal to.

$$\hat{\kappa}_N = \frac{\Big(W_{total}' \Phi_{total} H_{total} D_{total} C_{total} \Big)' A_N \Big(W_{total}' \Phi_{total} H_{total} D_{total} y_{total} \Big)}{\Big(W_{total}' \Phi_{total} H_{total} D_{total} C_{total} \Big)' A_N \Big(W_{total}' \Phi_{total} H_{total} D_{total} C_{total} \Big)} \tag{35}$$

Lucas et al. (1997) show in their paper that the outlier robust GMM estimator is consistent and asymptotically normally distributed. More precisely, the estimator is distributed as follows.

$$\sqrt{NS}(\hat{\kappa}_N - \kappa_0) \xrightarrow{\text{d}} \mathcal{N}(0, (M_1' A_0 M_1)^{-1} M_1' A_0 M_2 A_0 M_1 (M_1' A_0 M_1)^{-1}) \qquad \text{with} \tag{36}$$

$$M_1 = \text{E}\Big(\sum_{t=T_i^F}^{T_i^L} w_{it} v_t(w_{it}) \psi'(e_{its}^0/\sigma_s) h_{its} \Delta c_{i,t}' \Big) \tag{37}$$

$$M_2 = \text{E}\Bigg(\Big(\sum_{t=T_i^F}^{T_i^L} \sigma_s w_{it} v_t(w_{it}) \psi(e_{its}^0/\sigma_s) h_{its} \Big) \Big(\sum_{t=T_i^F}^{T_i^L} \sigma_s w_{it} v_t(w_{it}) \psi(e_{its}^0/\sigma_s) h_{its} \Big)' \Bigg) \tag{38}$$

Whereby $\Delta c_{i,t} = (\Delta z_{i,t-1}, \Delta v_{i,t-1})$. Hence, the optimal weighting matrix A_0 from a efficiency perspective is equal to M_2^{-1}. Which in turn leads to the following asymptotic normal

distribution for the estimator

$$\sqrt{NS}(\hat{\kappa}_N - \kappa_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, (M_1' M_2^{-1} M_1)^{-1}) \tag{39}$$

The optimal robust GMM estimator can be estimated via the classical two-step approach. In the first step, a consistent but not efficient estimator $\hat{\kappa}_N^{(1)}$ is estimated by using the identity matrix as weighting matrix. Then, one can use these estimated parameters to calculate the optimal weighting matrix $(M_2^{(1)})^{-1}$ and use this optimal weighting matrix to estimate the most efficient estimator $\hat{\kappa}_N^{(2)}$, which is distributed as follows.

$$\hat{\kappa}_N^{(2)} \sim \mathcal{N}(\gamma_0, 1/(NS)(\hat{M}_1' \hat{M}_2^{-1} \hat{M}_1)^{-1}) \tag{40}$$

Whereby $\hat{M}_1$ and $\hat{M}_2$ are the sample analogues of M_1 and M_2 respectively.

4.5 Model selection

When the model has been estimated for a several number of segments, one should find the optimal number of segments via a model selection criterion. This has to be a robust model selection criterion, since the model is estimated by means of the outlier robust GMM estimator. As has already been noted by Ronchetti and Staudte (1994), robust model selection criteria have not received much attention in the literature. Robust versions of the Akaike information criterion and Mallows C_p statistic have been developed by Ronchetti (1985) and Sommer and Staudte (1995) respectively. However, note that these robust model selection criteria are only valid in case of a standard M, S or MM type regression. Hence, they cannot be used as criterion for the outlier robust GMM results. Agostinelli (2002) derived a robust model selection criterion via weighted likelihood. I will follow his approach in using a weighted likelihood value as base for a Bayesian information criterion (BIC).

Let us denote this final set of model parameters as $\hat{\gamma}_s$ and $\hat{\lambda}$. Furthermore, the outlier robust GMM estimation procedure will produce weights $\hat{\phi}_{its}$ that bound the influence of anomalous observations on the estimated parameters. Next to this define the household weights, which are given in the data as h_{it}^{hh}. Note that the average household weight should be equal to one for calculating a proper BIC value. Using these parameters and weights in the original likelihood function gives us the following likelihood and log-likelihood value.

$$L_S = \prod_{i=1}^{N} \prod_{t=T_i^F}^{T_i^L} \Big(\sum_{s=1}^{S} p_s \big(\phi(\Delta y_{it}; \Delta z_{it-1} \hat{\gamma}_s + \Delta v_{it-1} \hat{\lambda}), \hat{\sigma}_s^2 \big) \Big)^{\hat{\phi}_{its}} \Big)^{h_{it}^{hh}} \tag{41}$$

$$l_S = \sum_{i=1}^{N} \sum_{t=T_i^F}^{T_i^L} h_{it}^{hh} \log \Big(\sum_{s=1}^{S} p_s \big(\phi(\Delta y_{it}; \Delta z_{it-1} \hat{\gamma}_s + \Delta v_{it-1} \hat{\lambda}), \hat{\sigma}_s^2 \big) \Big)^{\hat{\phi}_{its}} \Big) \tag{42}$$

whereby S equals the total number of segments that have been used. Note that direct comparison of log-likelihood values over different segments via BIC is not advisable. Because this would give benefit to models that have a higher proportion of outlying observations with $\hat{\phi}_{its}$ equal to zero. However, one cannot take this issue into account directly. Which is due to the fact that the outlier-weights $\hat{\phi}_{its}$ cannot be extracted from the logarithmic term, as

they are segment specific. To overcome this issue, the outlier-weights over all segments will be merged into one outlying term $\tilde{\phi}_{it}$ in the following manner.

$$\log\Big(\sum_{s=1}^{S} p_s\big(\phi(\Delta y_{it}; \Delta z_{it-1}\hat{\gamma}_s + \Delta v_{it-1}\hat{\lambda}), \hat{\sigma}_s^2\big)\Big)^{\hat{\phi}_{its}}\Big) = \tag{43}$$

$$\tilde{\phi}_{it}\log\Big(\sum_{s=1}^{S} p_s\big(\phi(\Delta y_{it}; \Delta z_{it-1}\hat{\gamma}_s + \Delta v_{it-1}\hat{\lambda}), \hat{\sigma}_s^2\big)\Big) \tag{44}$$

The obtained $\tilde{\phi}_{it}$ values can be used to determine the following adjusted log-likelihood value.

$$l_S^{adj} = \Big(l_S \sum_{i=1}^{N}\sum_{t=T_i^F}^{T_i^L} h_{it}^{hh}\Big)\Big/\Big(\sum_{i=1}^{N}\sum_{t=T_i^F}^{T_i^L} \tilde{\phi}_{it} h_{it}^{hh}\Big) \tag{45}$$

As just has been mentioned, this adjusted log-likelihood value takes differences in the outlying weight vectors into account, such that likelihood values can be compared over different models. The eventual choice of the number of segments will thus be based on the BIC, which is defined as $\text{BIC} = \log\Big(\sum_{i=1}^{N} T_i\Big)k_S - 2l_S^{adj}$. Whereby the total number of free parameters k_S equals $k_1 S + k_2$, with k_1 and k_2 the number of segment specific and non-segment specific explanatory variables respectively.

4.6 Model specification tests

A first check for correct model specification is the standard GMM test of over identifying restrictions, or Sargan test. Which says that the quadratic optimization criterion J_N times the total number of households (N) and the number of segments (S) follows a chi-squared distribution with $M - k$ degrees of freedom. Under the null hypothesis that all moment conditions are valid. Whereby M equals the number of moment conditions and k the number of parameters to be estimated. Note that the test is only valid when the optimal weighting matrix $(M_2)^{-1}$ is used for calculating the quadratic optimization criterion. Hence, if this test is rejected, at least one of the moment conditions is invalid, which means that the model is misspecified. Two frequently occurring model misspecifications that directly lead to invalid moment conditions in the dynamic linear panel setting are original disturbances which are serial correlated (so not the first differenced disturbances) and a too progressive choice for the endogeneity level of one of the regressors.

If the assumed level of exogeneity of some of the covariates were too strong, this leads to a the rejection of the Sargan-test. In order to test this, one should reestimate the model with weaker exogeneity assumptions. Note that the instruments used under these weaker exogeneity assumptions are a subset of the instruments under the stronger exogeneity assumptions. If we denote DS to be difference between the Sargan test statistics under the strong and weak exogeneity assumptions. Then, under the null hypothesis that the additional moment conditions are valid, it holds true that $DS \sim \chi^2(d)$. Whereby d equals the number of extra moment conditions under the stronger exogeneity assumptions. Arellano and Bond (1991) also derived a test statistic for checking the the presence of serial correlation in the original

disturbances. However, this test statistic has only been derived for an unweighted first differenced GMM-estimation and thus cannot be used in this case. In section 7, I will further elaborate on this topic.

Another issue that may arise when estimating a dynamic linear panel model is the problem of weak instruments. Hence, if the correlation between the explanatory variable Δy_{it-1} and its lagged values becomes weak, the first differenced GMM estimator will give biased results. There are two main underlying causes for this issue. Firstly, when the original time series of income growth of a certain household is highly persistent. Hence, if α is getting close to one in the following model $y_{it} = \alpha y_{it-1} + \mu_i(1-\alpha) + v_{it}$, the lagged values of y_{it} will become weak instruments for Δy_{it-1}. Therefore, one should estimate these time-series models on a household level beforehand, to check if the issue of persistent time-series might occur. The second cause of weak instruments is a very large ratio $\text{var}(\mu_i)/\text{var}(v_{it})$. Hence, if the variation in average income growth levels between households is much larger than the variation in income growth levels within households. This can also easily be checked beforehand via estimating the individual time-series given above. When the used instruments in the first differenced GMM estimator are suspected to be weak, one can reduce the forthcoming finite sample bias of the estimates by estimating the extended system GMM estimator, which was proposed by Blundell and Bond (1998). For this extended system GMM estimator one has to assume some extra properties, such as the stationarity assumption of the initial conditions. These extra assumptions provide some extra moment conditions that are not specified in the first differenced setting, but in the original level setting. Therefore they greatly reduce the estimation bias in case of persistent time series.

5 Results

Before discussing the results of the four distinct analyzes, I have to point out a few adjustments that have been carried out in order to make the estimation process feasible. Firstly, the instrument matrix that have been used differs a bit from the instrument matrix introduced in the methodology section. The well known GMM-estimator of Arellano and Bond (1991) has been designed to estimate parameters in a typical micro-panel setting whereby the number of time-periods is small. However, for the analysis with the American data there are for example 41 periods in total. Including all previous values of all variables as instruments would thus lead to a total of 8610 instruments, given that there are nine explanatory variables in total. Besides the fact that the estimation process took very long, the major problem was the estimation of the covariance matrix of the moment conditions. Despite the large number of observations in the dataset, reliable estimation of this large covariance matrix turned out to be impossible. As this covariance matrix is required for the re weighting step of the GMM estimator, as well as for calculating the standard errors of the parameters, its reliability is an essential requirement for proper estimation results. Next to this, there is a risk of biased parameter estimates due to over-fitting if one uses numerous instruments (Roodman, 2009). There are multiple solutions for the issue of too many instruments. In this research the most common solution will be used, which is cutting of at a certain lag instead of using all available lags as instruments. The main question of this solution is at which lag one should cut off the instrumental variable matrix. For sake of simplicity, I will choose one overall cutting point for all variables and all models. This

decision should be based upon the trade-off between the information that is lost by cutting off at a certain lag on one hand and the reduction of the total number of instruments on the other hand. The relative amount of information that a certain lag gives is well represented by the correlation between the first differenced values and the corresponding lagged values of a certain variable. In table 2 the correlations of the first ten lags are given for the German dataset. The patterns within the correlation levels of the other four countries are reasonably well represented by the correlation levels of the German data. Given the size of these tables, they have not been added to the paper. Lastly, note that the correlation values of the micro economic variables are based upon an outlier robust covariance matrix, which has also been estimated via the DetMCD estimator of Hubert et al. (2012). Next to this, the correlations of the macroeconomic variables Inequality level and Redistribution are the averages over the four distinct regions in which Germany has been divided.

	1	2	3	4	5	6	7	8	9	10
Real income growth	-0.81	0.15	0.01	0.00	-0.01	0.01	0.00	-0.01	0.00	0.01
Inequality level	-0.22	-0.05	-0.10	-0.08	-0.18	-0.20	-0.11	0.00	-0.07	0.04
Redistribution	-0.45	-0.48	-0.42	-0.23	-0.03	0.08	0.14	0.03	-0.17	-0.16
GDP growth	-0.72	-0.34	0.28	0.21	-0.09	0.12	-0.14	-0.32	0.09	0.30
Age	0.86	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87
Education years	-0.39	-0.43	-0.44	-0.43	-0.42	-0.42	-0.42	-0.39	-0.37	-0.33
Work hours growth	0.14	0.01	-0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00

Table 2: Correlation between first differenced data and the first ten lagged values for the German data

The correlations of real income growth are close to zero from the third lag onwards, while it is already from the second lag for the variable Work hours growth. In contrast, the other two microeconomic variables Age and Education years seem to have a constant correlation factor over time. Which is logical, given that these variables are close to deterministic. For the macroeconomic variable Inequality level, only small correlation values up to 0.2 can be found until the sixth lag. Redistribution seems to have a significant negative correlation for the first four lags. While GDP growth starts with two negative correlations, followed by two positive ones and then also correlation values close to zero. Combining these findings with the findings of the other four countries led to cutting off at the fifth lag. As this cutting point seems to strike a reasonable balance between the amount of information that is lost and the reduction of the number of instruments. Some researchers might find this cutting point decision a bit subjective. Therefore two other solutions for the issue of too many instruments will be discussed in section 7, which can be of interest for further research.

Next to the reduction of the number instruments, the three dummy variables (education level up, child born and child left) had to be discarded from the model. It turned out that these events were such rare that the outlier robust GMM estimator regularly converged to a state in which the corresponding parameter of a certain dummy variable could not be estimated. Because all observations with a value one for a certain dummy had an outlying weight value of zero in this converged state.

5.1 Persistence properties of household income growth

In table 3 summary results of the individual time-series models for the income growth process per household are given. These individual time-series models should be estimated beforehand, such that the persistence properties of the income growth variable can be checked. In the first three rows of table 3 one can find the relative number of households with an absolute α parameter value larger than a certain threshold. Note that these percentages are based upon households with at least five income growth values. Because estimating an individual model for households with less than five observations would give very volatile parameter estimates, from which one cannot properly derive time-series properties. For all five countries, the share of households with an estimated α value larger than 0.8 is quite small. Next to this, the variance of the model errors is at least twice as large as the variance of the household specific fixed-effect parameters for all countries. Therefore, one can conclude that the income growth time-series are in general not highly persistent. However, the share of α parameter values above 0.5 is substantial. A simulation study performed by Blundell and Bond (1998) revealed that the first differenced GMM estimator already gives a small downward bias when the individual time-series are simulated with an α parameter of 0.5. Therefore, it can be of interest to estimate the extended system GMM estimator of Blundell and Bond (1998) as well, and compare results with the first differenced GMM estimator. However, this comparison falls outside the scope of this research. Because it is expected that the lagged values of the income growth variables will in general not be weak instruments, as the share of highly persistent individual time-series is small.

	Germany	Korea	Switzerland	UK	US
α >1	0.021	0.019	0.023	0.031	0.013
α >0.8	0.050	0.041	0.060	0.072	0.034
α >0.5	0.279	0.296	0.354	0.335	0.190
Relative variance	0.184	0.115	0.134	0.438	0.119

Table 3: Percentages of households with an absolute α value larger than a certain threshold given per country. These α values are obtained by estimating the following model for each household $y_{it} = \alpha y_{it-1} + \mu_i(1-\alpha) + v_{it}$. Next to this the relative variance is given of the fixed effect parameters compared to the time-series model errors ($\text{var}(\mu_i)/\text{var}(v_{it})$)

5.2 Choosing the number of segments

In table 4 one can find the Bayesian information criterion values for different choices of the total number of segments. Note that these BIC values are based upon models that are estimated with a random subsample of ten percent of the households, as the estimation process is computationally quite exhaustive. As a first note, the boxes associated with four segments are empty for analysis 1, 2 and 4. This is due to the fact that the EM-algorithm mostly converged to a state in which one of the four segments was empty. For five or more segments, this was always the case. Comparing the BIC values across different segments reveals that the BIC values are the lowest for two segments. Actually, the adjusted likelihood value, from which the BIC is derived, attained the highest value for two segments. If one would use a non-robust estimation process in the maximization step of the EM algorithm,

then adding extra segments would always lead to an higher likelihood value. However, for the likelihood value that is adjusted for the number of outlying observations in the robust estimation procedure, this property does not hold. The underlying reason is that the extra third segment is used to explicitly model the group of households with the most volatile income growth process. Hence, in the two segment setting, the EM-algorithm converged to a state wherein both segments had a relatively small volatility parameter σ_s. Because of this, most of the households with volatile income growth process were threated as outlying observations. Therefore they had no influence on the eventual adjusted likelihood value. In contrast, if one allows for a third segment, then this (smallest) third segment generally had a much larger volatility parameter. Hence, the third segment was used to explicitly model the households with the most volatile income processes. The likelihood values of this high volatility segment are on average lower than the likelihood values of the other two segments. Due to this fact, the eventual adjusted likelihood value will be lower for the three segment case, compared to the two segment case.

	2 segments	3 segments	4 segments
Analysis 1	28020.1	32915.5	
Analysis 2	21970.2	25865.1	25879.0
Analysis 3	15821.3	17887.7	
Analysis 4	14696.8	20299.4	

Table 4: Bayesian information criterion values for each analysis and different number of segments.

For further research, it might be of interest to investigate the addition of a penalty term to the adjusted likelihood value. Whereby the penalty term is based upon the share of observations that are viewed as outlier. For this research, I will simply estimate a model with two segments and a model with three segments for all four distinct analyses. Because the allowance for a smaller third segment, in which the most volatile income processes can be explicitly modeled, might give interesting results. Moreover, it gives the possibility to compare results obtained for two and three segments. Such that eventual conclusions on the effects of macro economic indicators on household income growth are stronger.

5.3 Convergence of the expectation-maximization algorithm

For each of the four distinct analyses, a model with two and a model with three segments have been estimated, using all data available. In the appendix, table 15 up to table 34 summarize the results of the eventual eight models. In subsection 5.5, I will explain how one can interpret these tables. The eventual results are based upon choosing the best local optimum out of five local optima obtained via different random starting points of the EM-algorithm. In order to find the global optimum with a high probability one should probably use more random starting points. However, as noted earlier, the estimation process was computationally quite exhaustive, which made it infeasible to use more than five random starting points. To illustrate, the estimation process of one random starting point generally took around two to five hours. The main reason for this long computation time is the fact that the outlier robust GMM estimator in the M-step of the EM algorithm is also estimated iteratively, just as the EM-algorithm. Within each iteration one has to perform the matrix

multiplication given in equation 35 to find optimal parameter values, given the current outlying weights. This matrix multiplication obviously takes quite some computation time given the sizes of the instrumental variable matrix (W_{total}) and the explanatory variable matrix (C_{total}). For the models that had no 'close to convergence' issues, which will be explained in the next paragraph, a large part of the five local optima were actually equal to the global optimum of these five optima. Therefore, the total number of local optima seems to be small for these mixture models. Hence, if one is able to do further research with more computational power, I would suggest to use around twenty random starts for the EM algorithm.

In general, around twenty to thirty iterations were needed for the outlier robust GMM estimator to converge. However, the EM-algortitm is known for getting close to an optimum within a few iterations, but the eventual convergence rate is generally low. In this case, the EM algorithm was only getting close to convergence for some of the models. Close to convergence in the sense that after 50 iterations of the EM algorithm, the segment probabilities stayed within a certain small range over the subsequent iterations, but never really converged to a stable optimum. Therefore, I decided to use a maximum of 70 iterations. Eventually, if the EM algorithm did not converge within 70 iterations, the iteration with the highest corresponding adjusted likelihood value would be used as output. To give an indication for which of the eight models the issue of non-convergence of the EM algorithm existed, I have added the value 's.d. seg. prob.' to each coefficient table. For all eight models, multiple starting points eventually converged or close to converged to the same local optimum. Hence, standard deviations of the segment probabilities over different starting points that converged to the same local optimum give a good indication for level of convergence of the EM-algorithm. The value 's.d. seg. prob.' is equal to the average of these standard deviations over all segments. Values of this summary statistic are all close to zero for the four models with two segments, so the EM-algorithm converged well for these models. However, the summary statistic indicates that the models with three segments had more converging issues, especially for the latter three analyses. In section 7, I will further elaborate upon a possible cause of the convergence issue of the EM-algorithm and give some suggestions for further research to solve this issue.

5.4 Bootstrapping households for standard errors of parameters

As has already been mentioned in section 3, in order to incorporate the uncertainty associated with the missing values in the CNEF data, one should estimate models for multiple bootstrap samples of all households. To get reliable probability value (p-value) estimates one should use a few hundred bootstrap replicates. Given the very large computation time of the estimator, this is unfortunately infeasible. Moreover, this procedure of bootstrapping should be performed for each model separately. Hence, if one accepts the drop in reliability of the p-values and estimate models for e.g. 10 bootstrap replicates, one still has to estimate 80 models in total. As it can take up to a full day estimate one model, this option is still infeasible. Therefore, a bootstrapping procedure with seven replicates in total has been performed for one of the eight models and the obtained results are also used in the other seven models. To be more precise, bootstrapping the households of the fourth analysis, gave new standard errors for the parameters of the model with two segments. These new standard errors are then divided by the 'normal' model standard errors, for which the uncertainty of

the missing values is not incorporated. This procedure will give an uncertainty inflation factor through the missing values for each parameter. Taking the mean value over these uncertainty inflation factors gives an inflation factor value of 3.16, which best summarizes the extra uncertainty level that is caused by the missing values. Therefore, the standard errors of all eight models will be multiplied with this average uncertainty inflation factor. By doing so, these new standard errors should incorporate the uncertainty caused by the missing values as good as possible, given the in feasibility of the normal bootstrapping procedure.

There are two reasons for choosing the fourth analysis for obtaining standard errors via bootstrapping. Firstly, all five countries are incorporated in this analysis. Next to this, the relative amount of missing values is the highest in this analysis. Hence, it is expected that the uncertainty inflation factor will be lower for the other three analyses. Thus, if one can conclude that a certain parameter value of one of the first three analyses is still significant after using the uncertainty inflation factor of 3.16. It is likely to be significant as well if one would have obtained standard errors via the regular bootstrapping procedure. Then the model with two segments was preferred for bootstrapping over the model with three segments since it is computationally less exhaustive. Moreover, as just has been noted, convergence properties of the EM-algorithm were much better for the models with two segments. Hence, if one would use the model with three segments, it remains unclear if the extra uncertainty obtained via estimating multiple bootstrap replicates comes from the missing values or from the convergence issues of the EM-algorithm. I am aware of the fact that using this general uncertainty inflation factor has not much to do with proper statistical analysis. However, it was the only solution that I could think of which was both computationally feasible and still quite reliable. Quite reliable in the sense that the standard errors are probably overestimated and therewith statistical significance is probably not concluded for insignificant variables. Moreover, in the remainder of this paper I will be careful with concluding significance for borderline cases with a p-value between 0.01 and 0.05.

5.5 Table explanations

Before giving an interpretation of the obtained results, which are summarized in table 15 up to table 34 in the appendix, I will give a short description of these tables. Note that the main findings of these tables are highlighted within this results section in table 5 up to table 8. The first eight result tables in the appendix (table 15 up to 22) give the coefficients, standardized errors and p-values of the eight distinct models. The standardized errors shown in these tables are the regular standard errors obtained via the asymptotic property given in equation 40. Next to this, two p-values are given in each table (p-value and boot p-value). As the names already suggest, the first p-value is obtained via the regular standard errors, while for the latter all standard errors are multiplied with the uncertainty inflation factor caused by missing values of 3.16. The latter one will obviously always be used for determining significance. Then, each number at the end of a segment specific explanatory variable corresponds to its segment number, whereby the segments are ordered according to their size. Lastly, some summary statistics are given on the right side of each table. Two of these summary statistics are '% outlier 0' and '% outlier < 1', which give the percentages of the observations that have received an outlier weight value $\hat{\phi}_{its}$ of respectively zero and smaller than one.

In table 23 up to table 26 one can find the segment probability of each segment together with some segment probabilities of subgroups. To be more precise, subgroups are created by splitting all households according to their value for a certain variable (e.g. age or average disposable income). Eventually, subgroup segment probabilities are calculated by taking the mean over the segment probabilities of all households within that subgroup. Before continuing with a further specification of some of the subgroups, I will first discuss a general transformation of the income variables that is used throughout the remainder of this section. The income variables, which are post-government income, pre-government income and labor income should be comparable over time and the different countries. Therefore, each income variable will be divided by the median value of the corresponding country and year. The subgroup mainly non labor income consists of households for which the share of labor income from their total pre-government income is less than ten percent, given that their pre-government income is higher than fifty percent of the median income. Hence, this subgroup includes all households with a significant amount of asset or private retirement income and hardly any labor income. The subgroups mainly labor income and mainly government income are created in a similar fashion. The subgroups lowest income and highest income consist of ten percent of the households with respectively the lowest and the highest average post-government income. The subgroups volatile and stable are created by taking thirty percent of the households with respectively the highest and the lowest standard deviation of their income growth time-series. Lastly, the group of households with very volatile income process is segmented into three subgroups according to their average disposable income. Note that some other subgroups, such as households with children, have been created as well. Since the segment probabilities of these subgroups did not significantly differ from the overall segment probabilities they are omitted from the tables.

This subsection is concluded by explaining table 27 up to 34, which contain weighted quantile values for household averages of some variables over the different segments. Hence, the second row of this table contains the segment numbers for which the weighted quantiles are determined. These weighted quantile values are calculated by using the segment probability multiplied with the number of observations of a certain household as weight. In other words, the idea of this table is to give a summary of the shape of the distribution of a certain variable for a certain segment. Once again, the transformed values of pre-government, post-government and labor income are used for calculating household averages, such that a cross-time and cross-country comparison is possible. In the remainder of this chapter I will refer to these three groups of tables as coefficient tables, segmentation tables and quantile tables.

5.6 Parameter analysis

Before actually analyzing the segment specific parameters, it is convenient to have a closer look at the segmentation tables, such that some general segmentation patterns can be found. One striking pattern that is present in all eight latent class models, is the segmentation of households with an average or stable income into one segment and households with a below average or volatile income in one or two of the other segments. In table 5 one can find the segment probabilities of the associated subgroups for the third and fourth analyses, but note that similar patterns can be found for the first two analyses. Hence, if we turn our interest to the segmentation tables of the models with two segments, so the left part of table. One can see that there is one segment with segment probabilities for the unemployed, lowest

income and volatile subgroups higher than the overall segment probability. Moreover, the subgroup of stable incomes is underrepresented in these segments, which seems logical, given that the subgroup of volatile incomes are overrepresented. However, do note that these subgroups with relatively more low income households, do not necessarily have relatively less high income households. In general, a similar segmentation can be found for the latent class models with three segments. Though, in this case there is one segment with above average stable incomes and thus below average volatile incomes and less households with a low income or no job. Then, the other two segments have the exact opposite, so relatively less stable incomes for the third analysis. While for the fourth analysis, there is one segment with the exact opposite and one segment with segment probabilities for the stable, volatile, no job and lowest income subgroups that are close to the overall segment probability.

	group 1	group 2	group 1	group 2	group 3
segment probability	0.598	0.402	0.615	0.327	0.058
no job	0.464	0.536	0.433	0.429	0.139
lowest income	0.442	0.558	0.425	0.413	0.162
volatile	0.283	0.717	0.201	0.594	0.205
stable	0.837	0.163	0.913	0.085	0.002
segment probability	0.686	0.314	0.580	0.368	0.052
no job	0.512	0.488	0.717	0.228	0.054
lowest income	0.482	0.518	0.706	0.239	0.056
volatile	0.220	0.780	0.877	0.082	0.041
stable	0.950	0.050	0.209	0.672	0.118

Table 5: Overall segment probabilities and segment probabilities of certain subgroups. Results of the third analysis are given above and results of the fourth analysis are given below.

The fact that lower incomes and more volatile incomes can be found in one subgroup is not surprising, given the positive correlation between these two variables. This positive correlation simply arises from the fact that income growth is calculated relative to your previous income. Therefore, the households with the most volatile incomes have been segmented into three separate groups according to their income level. By doing so, one can see that the volatile middle incomes and volatile high incomes are also overrepresented in the segments with a relatively high amount of low incomes. Hence, the households in these segments are best characterized as households with either a low or a volatile income (or both) and having a higher risk of being unemployed. Lastly, persons that have no job more often, obviously also have a more unstable income growth process, so it is also not surprising that these subgroups follow the same tendency in the segmentation process.

If we turn our interest to the quantile tables, one more interesting pattern can be found. The post-government income quantile values of the segments with a relatively large amount of households with low incomes are obviously lower in general than the corresponding quantile values of the other segments. However, at either the 90 percent or 99 percent quantile, the order of the quantile values is reversed. Hence, the subgroup of the richest 1 percent of all households is also overrepresented in the segment with a relatively large amount of households with low or volatile incomes. A similar reversion in the order of the quantile

values at the 90 or 99 percent quantile can also be found at the pre-government and labor income quantiles. A possible explanation for this finding might be that these very high income households generally have a more volatile income. Because part of there yearly wage is a bonus or they are more dependent on asset income, which is more volatile. Note that this finding only holds true for the two models corresponding to analysis one, three and four.

5.6.1 Lagged income growth parameters

For the analyses with three segments in total, a general pattern is present for the segment specific parameters corresponding to the explanatory variable lagged income growth. This pattern is a significant positive parameter value for the segments with mainly stable incomes and a significant negative value for the other two segments. This finding seems logical, given that an autoregressive process with one lag and a negative parameter is automatically more volatile as it oscillates around the mean value. Exceptions to this finding are the stable segment (segment 1) of analysis 2, which has a weakly significant negative parameter and the more volatile segments 2 of analysis 2 and 3, which have a positive parameter value. There does not seem to be any explanation for these exceptions.

However, the reliability of all parameter values corresponding to the lagged income growth variables can be questioned due to multiple reasons. Firstly, for the analyses with two segments, there is no visible pattern in the signs of the lagged income growth parameter values. Secondly, some of the lagged income growth parameters have an absolute value close to one or even greater than one. This is a bit surprising, given that there were hardly any households who had an highly persistent individual income growth time-series with an absolute lagged income parameter value larger than 0.8. This finding is very well illustrated by the parameter values of analysis 1 with three segments. In table 16 one can find parameter values of respectively 1.13 and -1.08 for the lagged income growth variable of the first and third segment. Whereas the percentage of households with an absolute lagged income parameter value larger than 0.8 is 3.4 percent for the US, which is the lowest of all countries. The last and most important reason pulling reliability of the lagged income growth parameters in doubt are the Sargan-test statistics. These test statistics are equal to one for all eight latent class models. The two main causes for the rejection of the Sargan-test are serially correlated disturbances and a too progressive choice for the endogeneity level of the regressors. As has been explained in the beginning of this section, only the previous five values are used as instruments for each variable. Hence, the latter of the two causes cannot be the problem. Serially correlated models disturbances arise when the proposed model is not able to capture the underlying dynamics of the dependent variable. It sounds reasonable that the chosen dynamic latent class model with one (linear) lag is indeed not capable of capturing the underlying dynamics of household income growth. Because the parameter values for the lagged income growth parameters exhibit indescribable patterns. Moreover, previous literature has demonstrated that the dynamics in income levels of households from other countries exhibit non-linearity (see for example Lokshin and Ravallion (2004)). As has already been pointed out in section 4, the test statistic of no second-order autocorrelation proposed by Arellano and Bond (1991) cannot be used to test if this really is the underlying cause of the rejection of the Sargan-test. Hence, there is no straight forward approach for solving this issue. Therefore, any re-estimation of the models with higher order dynamics included have not been performed, also because of the very long computation time of the estimator.

In section 7, this issue will be discussed in more detail. Note that the parameter values of the other variables in the latent class model are probably also biased, if the proposed latent class model is not capable of capturing the underlying dynamics of the household income growth process. However, it seems reasonable to assume that the biases for the parameters of the other segment specific explanatory variables are much smaller than the biases of the lagged income growth parameters. Because these other macro-economic variables are not directly linked to the internal household income dynamics, Therefore, in the remainder of this section I will still try to answer the research questions by analyzing the results of the other parameters. However, due to these possible biases, I will generally only interpret the signs of the parameters and not the magnitudes.

5.6.2 Inequality level parameters

In table 6 one can find the coefficients and corresponding bootstrapped p-values of the Inequality level parameters. For the latent class models with two segments there is a very clear pattern visible for the parameters in all four analyses. The effect of this variable is namely always positive for the underlined segment with a relatively high amount of stable incomes, while the effect is negative for the other segment with mainly low and volatile incomes. In addition, results of the second analysis also suggests that older households, households with a high average income level and households with non-labor income as main income source generally benefit from a higher level of income inequality, as these households are overrepresented in the first segment.

	Analysis 1		Analysis 2		Analysis 3		Analysis 4	
	coef.	p-vlalue	coef.	p-vlalue	coef.	p-vlalue	coef.	p-vlalue
segment 1	-5.234	0.000	1.369	0.019	1.967	0.000	1.009	0.061
segment 2	4.483	0.002	-15.138	0.000	-4.225	0.000	-3.713	0.003
segment 1	0.836	0.535	-1.669	0.045	2.345	0.258	-2.416	0.000
segment 2	-0.287	0.883	3.715	0.080	-6.867	0.483	2.043	0.014
segment 3	-4.793	0.188	-0.925	0.879	1.045	1.486	0.532	0.665

Table 6: Coefficients of the macroeconomic variable inequality level with associated bootstrapped p-values. The segments with a relative large amount of stable and middle and high incomes are underlined.

Results of the latent class models with three segments are less clear-cut as the results for the two segment models were. However, a similar pattern for the parameters of inequality level can be found in the third and fourth analysis. In the first two analyses, all parameters associated with the level of inequality are insignificant. If one would convert the found pattern to these two analyses, then the third segments with a relative high amount of low and volatile incomes should experience a negative effect. A possible reason for the insignificance of the parameter in the third segment of the first analysis might be the increase in the parameter standard error due to a higher segment standard deviation (σ_3). The higher segment standard deviation indicates that the households with volatile incomes are actually modeled in this case. Whereas they are seen as outliers in the first segment of the two class model. Hence, from this finding one may conclude for the first analysis that the effect

of inequality is significantly negative for households with a low income, while it might be insignificant for households with a volatile income.

From the values in the segmentation and the quantile tables one can derive that the first segments of the first two analyses are similarly structured as the second and first segment of respectively the first and the second analysis of two class models. Therefore, one would also expect a significant positive effect for these segments. For the first analysis, it might be the case that effects of inequality level and redistribution have switched due to non-stationarity of the corresponding time-series. In table 15 one can see that parameter values are equal to 4.48 and 0.42 for the second segment of the two class model, while they are equal to 0.84 and 6.52 for the first segment of the three class model. As can be seen in figure 1 and figure 2, both the income inequality level as well as the level of redistribution increased over time for the United States. Due to this similarity, it might be the case that the effects of inequality level and redistribution on income growth have interchanged for the first segment of the three class analysis. Note that the stationarity properties of the inequality level, redistribution and GDP-growths have not been studied for this paper, as these three variables are in the long term all theoretically stationary. However, for a shorter time-span, they might become empirically non-stationary. Therefore, it can be of interest for further research to deeper investigate this issue.

Another aspect of the first two analyses with three segments are the segment probabilities of the subgroups volatile low income, volatile middle income and volatile high income. As one can see in the segmentation tables, these segment probabilities are inclining for the second segment, while they are strongly declining for the third segment. On the other hand, the segment probabilities of the subgroups no job, lowest income, volatile and stable are relative to their overall segment probabilities roughly equal for the second and third segment, especially in the second analysis. Due to this setting, one is able to derive if there are any differences in the effect of inequality level on the income growth levels of households from these three subgroups. In both analyses, the parameter value of the third segment is lower than the parameter value of the second segment. Which suggests that the subgroup volatile low income experiences a greater negative effect from the level of inequality than the subgroups volatile middle income and volatile high income. Lastly, the income inequality coefficients of the third and fourth analyses with three segments. These coefficients are almost completely in line with the earlier described pattern. The only exception to this finding is the insignificant coefficient for the third segment of the third analysis. Because low and volatile incomes are over represented in this segment, one would expect a significant negative parameter value.

A last interesting feature of the inequality level parameters can be found in the segment probabilities of the participating countries. The general pattern is that the segment probabilities of Germany and Switzerland are the highest of all participating countries for the segments with a significant positive parameter for the inequality level. In contrast, the segments with a significant negative parameter have higher segment probabilities for the other three countries. The only exception to this rule is in the fourth analysis with two segments. As the segment probability of Switzerland is slightly higher than the segment probability of the US for the segment with a negative parameter. In figure 1 one can see that the level of inequality is significantly lower for Germany and Switzerland, compared to the South-

Korea and the US. Moreover, the weighted average of the income inequality coefficients with the overall segment probabilities as weights is generally negative. From these findings one may conclude that the overall impact of the inequality level on household income growth is negative. Which is in line with the conclusions of Ostry et al. (2014). Furthermore, if one assumes that the higher inequality level within South-Korea and the US is partly due to a larger income gap between low income households and the other households. Which seems reasonable, as both South-Korean and US households as well as the ten percent relative lowest income households are always over represented in the segments with a negative income growth parameter. Then, results of these latent class models are in line with the conclusion of Cingano (2014) that the gap between low income households and the rest of the population matters most for subsequent growth. Regarding the conclusions of this paragraph, two remarks have to be made. Firstly, the UK does not follow the general pattern as described above. Because UK households are over represented in the same segments as South-Korean and US households, while the level of inequality within the UK is close to those of Germany and Switzerland. The lower level of income mobility in the UK might be the reason for the fact that UK households are also over represented in the segments with a negative impact. For example, the intergenerational earnings elasticity is equal to 0.5 in the UK, while it is 0.32 in Germany (Corak, 2006). Secondly, the conclusions drawn in this paragraph are partly based upon the magnitude of the income inequality parameters. Therefore, one has to bear in mind that these conclusions might be partly incorrect due to the possible bias in the parameters.

To summarize the findings given above, households with lower incomes generally experience a negative impact of a higher inequality level on their income growth. Next to this, households with the most volatile income processes seem to experience a negative effect of a higher inequality level. Though, it has to be noted that the first two analyses with three segments put the significance of this relation in doubt, especially for the middle and high income groups with volatile incomes. In contrast, households with a stable middle or high income generally benefit from a higher inequality level. In practice these results suggest that the level of inequality within a country is a self-generating system that lead towards very high inequality levels if the government does not intervene via redistribution. Self-generating in the sense that a high level of inequality has a negative impact on the income growth level of poor households, while it has a positive effect on stable middle and high incomes. Then through these effects, the level of inequality will grow in the future, which in turn has its effects on the income growth levels of the households, and so on.

5.6.3 Level of redistribution parameters

For the latent class models with two segments all parameters corresponding to the variable redistribution are insignificant, which can be seen in table 7. This finding is in line with the findings of Ostry et al. (2014), who concluded that redistribution generally had no direct effect on GDP-growth. However, only half of the coefficients associated with redistribution in the three segment models are insignificant. Because insignificance of the redistribution parameter seems to be the most logical outcome, given the results of the two segment analyses and the results of Ostry et al. (2014). I will try to give possible explanations for all of the significant redistribution coefficients in the remainder of this section.

	Analysis 1		Analysis 2		Analysis 3		Analysis 4	
	coef.	p-vlalue	coef.	p-vlalue	coef.	p-vlalue	coef.	p-vlalue
segment 1	-0.079	0.962	1.281	0.164	0.238	0.680	1.640	0.079
segment 2	0.420	0.820	-4.580	0.420	-0.134	0.888	-1.780	0.364
segment 1	6.523	0.000	-0.513	0.726	-0.065	0.933	0.833	0.322
segment 2	-10.540	0.003	2.021	0.588	4.516	0.017	3.663	0.010
segment 3	12.402	0.006	10.679	0.361	-11.650	0.219	-11.177	0.016

Table 7: Coefficients of the macroeconomic variable redistribution with associated bootstrapped p-values.

In the first analysis the coefficient of redistribution is positive for the first and third segment. On a first sight, this positive coefficient seems to be more logical for the third segment, as this segment has a relative large number of poor and unemployed households. The intuition behind this statement is that poor households may benefit from a higher level of redistribution through more progressive taxes, while unemployed people can benefit through a higher level of government transfers. The positive coefficient of the first segment might be explained by the interchange with the inequality level parameter of this segment, as has been explained in the previous section. Lastly, there is a significant negative effect of redistribution on the households of the second segment. From the segmentation table it become clear that this segment is overrepresented with households of the subgroups with volatile middle incomes and volatile high incomes. For these two subgroups, a higher level of tax might indeed lead towards less intent to work. As these households probably do not have a permanent job with fixed working hours, given their high income volatility. Moreover, they do not necessarily always need to work, given their higher income.

In the third analysis, the parameter of redistribution is weakly positive significant for the second segment. As this segment mainly consists of low income and unemployed households, this finding is not very surprising. The coefficient of the second segment of the fourth analysis is also weakly positive significant. A possible explanation for this finding can be found in the quantile table. The 10 percent quantile of pre-government income is the lowest for the second segment, while this segment relatively has the least number of unemployed households. From these two findings one may conclude that this segment has a very large proportion of retired households with hardly any private retirement income. These households can have large benefits from a higher redistributions level through higher public retirement incomes. Lastly, the coefficient of the third segment in the fourth analysis is negative. From the quantile table it becomes clear that the households with high incomes in this segment mainly have high labor incomes. Moreover, these high incomes are taxed more heavily than the incomes of the high income households of the other two segments. This finding is derived from the fact that the 75, 90 and 99 percent labor income quantiles are the highest for third segment, while the 75, 90 and 99 percent post-government income quantiles are the lowest for the third segment. Hence, through this relatively large tax on labor income, it might be that the effect of redistribution is negative for this segment.

5.6.4 Gross domestic product growth parameters

In table 8 one can see that the GPD growth parameters are generally insignificant for all segments in all eight models. Hence, the general conclusion for this variable is that the income growth of households does generally not depend significantly on the economical well-being of their country. This result is a bit counterintuitive, especially for the segments with more volatile incomes. A possible reason might be the fact that GDP-growths on a country level are used as proxy for the state of the economy. Whereas regional GDP-growths can differ significantly within large countries, such as the US. Hence, for further research, it might be of interest to study the effect of regional GDP-growths on household income growth. In the remainder of this section, the parameters corresponding to GDP-growth that actually were significant will be discussed.

	Analysis 1		Analysis 2		Analysis 3		Analysis 4	
	coef.	p-vlalue	coef.	p-vlalue	coef.	p-vlalue	coef.	p-vlalue
segment 1	-0.178	0.791	-0.094	0.699	-0.627	0.001	0.226	0.235
segment 2	1.380	0.163	1.653	0.329	1.855	0.000	-1.148	0.120
segment 1	-2.333	0.099	-0.463	0.351	-0.264	0.354	-0.877	0.047
segment 2	5.149	0.000	4.280	0.003	0.900	0.359	1.117	0.001
segment 3	-3.875	0.059	-3.644	0.375	3.974	0.484	-0.347	0.796

Table 8: Coefficients of the macroeconomic variable GDP-growth with associated bootstrapped p-values.

For the models with two segments, only the parameters of the third analysis are significant. The positive parameter value for the second segment seems quite intuitive, as this segment has a relatively large degree of households with volatile incomes and unemployed households. The negative effect of the first segment is a bit counter-intuitive though. As one can see in the segmentation table, this segment has a relative large degree of German households that have labor income as their main income source. Moreover, this segment has a relative large fraction of very stable incomes, given that the volatility parameter of this segment is the lowest of all segments over all eight models. Hence, a possible explanation of the negative coefficient might be that the wages in Germany usually rise after a period of high economical growth, when the economical growth itself is already in a downward cycle.

The second segment of the first, second and fourth analyses has a significant positive parameter value in the three class model. As noted earlier, the second segment of the first and second analyses has a relative large amount of households from the subgroups volatile middle income and volatile high income. A significant positive dependency of these subgroups upon the general well-being of the economy seems reasonable, given that a large part of these subgroups are probably freelancers. However, the significant positive effect for the second segment of the fourth analysis is less logical, as this group mainly consists of households with stable incomes.

5.6.5 Other parameters

Next to the segment specific macroeconomic variables, three household level variables were used as explanatory variables. These variables were age, education years and income growth due to working hours difference, whereby the latter two could only be estimated in the first two analyses due to data availability. The parameters of the variable age show the same pattern across the two types of models (2 and 3 segments). They are significantly negative for the first three analyses and insignificant for the fourth analyses. The fact that age generally has a negative impact on income growth is off course very intuitive. The number of education years is generally insignificant. Which might be explained by the fact that the starting salary of higher educated people is already higher, causing the income growth to be not significantly higher. However, the significant negative value for this variable at the first analysis with two segments is a bit counter-intuitive. Especially since the US is well-known for the large income-gap between lower and higher educated people. Lastly, the variable income growth due to working hours difference for which one would expect a positive effect. This has also been estimated for the second analysis with three segments. However, the parameters are significantly negative for the first two analyses with two segments. Any sensible explanation for this finding could not be given.

5.6.6 Starting weight sensitivity

The outlier robust GMM-estimator is estimated iteratively, whereby the process is started by choosing some starting outlier weights. The parameters that just have been analyzed are obtained by using the instrumental outliers as starting weights. In order to investigate the sensitivity of the eventual results to this starting criterion, three other starting weights have been tried, which are explained in section 4. The sensitivity analysis has been performed for one of the eight latent class models, as the estimation process is quite time-consuming. Just as for the bootstrapping procedure, the two segment latent class model of analysis 4 has been used for the sensitivity analysis. This model was chosen since data of all five countries is included and the EM-algorithm converged well for this model. The eventual results were equal to the original parameter estimates for all three starting criteria. The insensitivity of the eventual estimator for these distinct starting criteria can be explained by the fact that the outlier robust GMM-estimator is used within the EM-algorithm. Recall that the starting outlying weights are only used in the first iteration of the EM-algorithm, as the outlying weights of the previous iteration are employed in the other iterations. Moreover, multiple starting segment probabilities are used to find a global optimum out of several local optima. Therefore, the probability of finding the same global optimum for different starting outlying weights will rise with the chosen number of starts of the EM-algorithm.

6 Conclusion

The aim of this research was to investigate which macroeconomic factors drive household income growth, to what extend these driving forces were heterogeneous across different subpopulations and how these subpopulations were best characterized. To answer these questions, a latent class model has been estimated with the harmonized household panel data of the Cross-National Equivalent File. The EM-algorithm has been used to estimate the latent class model, whereby new parameter values in the maximization step where determined

via the outlier robust GMM-estimator of Lucas et al. (1997). In order to fit the data at hand and to draw stronger conclusions, four datasets have been modeled with a varying number of countries and time-frames. It turned out that latent class models with two and three segments were the best options in all cases. Hence, eight distinct latent class models were estimated in total.

The sargan test was strongly rejected for all eight models. The most likely underlying reason for these rejections was the incapability to capture the underlying dynamics of the income growth processes of the households through the chosen model. However, this could not be formally tested, as the corresponding test of no serial-autocorrelation from Arellano and Bond (1991) did not fit the latent class framework. The latent class models were not re estimated as the estimation process was computationally very exhaustive. Moreover, the expected bias in the model parameters associated with the macroeconomic variables was relatively small. Nevertheless, conclusions on the effects are generally based on significance of the parameters, not on their magnitude, as a result of this possible bias.

One striking pattern that is present in all eight latent class models, is the segmentation of households with an average or stable income into one segment and households with a below average or volatile income in one or two of the other segments. Next to this, the relative number of unemployed is always higher in the segment with below average and volatile incomes. Generally the segments with average and stable incomes experience a positive effect of the income inequality level within a country. Whereas this effect is negative for the segments with a below average or volatile income. Though, this negative relation is not very strong for the middle and high income groups with volatile incomes. Moreover, the findings of Ostry et al. (2014) and Cingano (2014) are confirmed by the estimated models. Hence, income inequality has a negative impact on the overall income growth and what matters most is the income gap between low income households and the rest of the population. However, do note that the conclusion for the overall impact on income growth might be partly incorrect due to the possible bias in the parameters. The parameters corresponding to the level of redistribution and GDP-growth were generally insignificant. Redistribution occasionally had a positive effect on segments with a relative high degree of poor households, unemployed households or retired households with no private retirement income. In contrast, households of the subgroups volatile middle and volatile high income seem to experience a negative effect of redistribution, while the effect of GDP-growth is generally positive for these subgroups.

The obtained results suggest that the level of inequality within a country is a self-generating system in the sense that high income inequality levels lead to even higher inequality levels in the future. Interestingly, a similar result has been found by Thomas Piketty for capital inequality. The underlying reason for the self-generating system of capital inequality is the general law that returns on equity are higher than the economic growth. Hence, for further research it can be interesting to inspect if this vicious cycle is indeed also present for income inequality and what might be the underlying cause of this vicious cycle. Luckily governments have a very powerful tool to cope with this self-generating system, as redistribution generally has no significant direct impact on household income growth. Hence, a higher level of redistribution will generally have a positive overall effect on household income growth, which is in line with the findings of Ostry et al. (2014). Whereby the overall effect is the

insignificant direct effect combined with the indirect positive effect via lowering the level of income inequality.

7 Discussion

This section is built up as follows, first the convergence issues of the EM-algorithm will be discussed in greater detail. Then the probable underlying cause of the Sargan-test rejections will be reviewed. Lastly, some other solutions for the issue of too many instruments and small extensions to improve the model will be given.

As was already noted in section 5.3, the EM-algorithm did not always converge perfectly, especially for the latent class models with three segments. No perfect convergence in the sense that segment probabilities converged to a certain range of ultimately five percent, but then never really converged to a steady state afterwards. Note that I have checked the validity of the algorithm and the self-programmed R-code via estimating a latent class model on a simulated dataset. When there was no noise added to the data, convergence of the EM-algorithm was generally very fast. However, after adding an extra random error term for ten percent of the simulated households, the same issue of slow or sometimes even only close to convergence of the EM-algorithm arose. Luckily, the estimated parameter estimates were still within range of the simulated parameters. A possible underlying reason for the fact that the EM-algorithm does not always perfectly converge might be the observations that have a model error outlying weight between zero and one. Recall that absolute standardized model errors between $\sqrt{\chi_1^{-2}(0.975)}$ and $\sqrt{\chi_1^{-2}(0.9975)}$ were down weighted via a fifth degree polynomial. Hence, this fifth degree polynomial is quite steep, or stated differently, a small change in the absolute standardized model error can result in quite a large change of the eventual model error outlying weight. Moreover, the observations within this range of standardized error values have a large impact on the eventual parameter estimates due to this large model error. Therefore, it might be the case that very small adjustments of the standardized model errors of these close to outlying observations over subsequent iterations of the EM-algorithm, lead to the fact that the EM-algorithm does not perfectly converge. If this really is the underlying cause of the issue, one could try to use a less steep down weighting function to solve the problem. Another solution that might be of interest for further research is to use one of the several acceleration methods for faster convergence of the EM-algorithm, such as the acceleration method by means of Quasi-Newton methods (Jamshidian and Jennrich, 1997). However, if the issue of no perfect convergence still arises with these acceleration methods, one could also opt for another estimation technique of the latent class model, such as a Bayesian analysis (Titterington et al., 1985).

The Sargan tests were strongly rejected for all eight models. A plausible cause for these rejections is the fact that the chosen model specification is not able to capture the underlying dynamics of the household income growth processes. As a result, the original disturbances (so not the first-differenced) will be serially correlated, which in turn will lead to a rejection of the Sargan test. In order to check if the original disturbances are serially correlated, one can test for no second-order serial correlation in the first-differenced residuals. Under the null hypothesis of no second-order serial correlation, Arellano and Bond (1991) define the

associated test statistic as follows.

$$m_2 = \frac{\hat{e}'_{-2}\hat{e}}{\sqrt{\hat{e}_*}} \xrightarrow{\text{d}} \mathcal{N}(0,1) \tag{46}$$

Whereby $\hat{e}_{-2}$ are the twice lagged disturbances per household, $\hat{e}$ all disturbances per household except for the first two and$\sqrt{\hat{e}_*}$ a scaling term. Further details of this scaling term can be found in Arellano and Bond (1991). The problem of using this test statistic in the current setting is that all observations are equally weighted. However, the model errors corresponding to a segment for which the segment probability is very small for a certain household, will generally be large and serially correlated. Therefore, it is essential to down weight these model errors with the household segment probability. Unfortunately, plugging in these segment probability weights into the test statistic is not straightforward (as the scaling term is quite comprehensive), neither is the derivation of the eventual asymptotic distribution. Therefore, the derivation and the usage of this test statistic falls outside the scope of this research. Another option to solve the issue is by means of trial and error. Hence, one could for example first extend the model by allowing for non-linearity in the dynamics of the income growth process, as was also done by Lokshin and Ravallion (2004). When the Sargan test is still rejected, one could also add higher order lags to the model. However, one cannot be sure that the incapability of capturing the underlying dynamics of the income growth process actually is the cause of rejection of the Sargan test. Hence, given the computational intensity of the estimator, I decided not to follow this trial and error procedure.

In this paper the total number of instruments has been reduced via cutting off at a certain lag. However, the chosen cutting off point might be seen as a bit subjective by other researchers. Therefore two other interesting methods will be given, which may be used in further research on this topic. Firstly, Roodman (2009) gives a pretty straightforward method to reduce the number of instruments. In place of the standard first-differenced GMM moment conditions ($\text{E}(y_{i,t-l}\Delta\epsilon_{it}) = 0$ for each $t \geq 3$ and $l \geq 2$), one imposes $\text{E}(y_{i,t-l}\Delta\epsilon_{it}) = 0$ for each $l \geq 2$. Hence, these moment conditions are based upon the same orthogonality conditions, but the empirical moments are only minimized over l, instead of l and t. By doing so, one may reduce the number of columns of the instrument variable matrix via placing all lagged values of the same order into one column. Hence, the first column will consists of y_{i1}, y_{i2}, y_{i3}..., which are the first order lags associated with respectively Δy_{i2}, Δy_{i3}, Δy_{i4}... The advantage of this method is that more information is potentially retained, as no lags are actually dropped. Another more advanced method, has been proposed by Mehrhoff (2009). His idea is to reduce the number of instruments via principal component analysis. This method seems very suitable for this issue, as the objective of principal component analysis is to reduce the dimensionality with a minimal amount of information loss. However, do note that the stronger assumption of independence of $y_{i,t-l}$ and $\Delta\epsilon_{it}$ is required for using the principal components within the moment conditions of the GMM-estimator.

Next to the variables that are used in these latent class models, one could think of other variables that might have a significant influence on household income growth. A good example are tax-reforms, which obviously directly influence the post-government income of households. Therefore, it could be of interest for further research to look at major tax reforms and include these as dummies to the model. However, do note that these tax reform event dummies are also rare events. Therefore they might experience the same issue of being seen

as outlying observations. As this was also the case for the three dummy variables that were initially added these models. For this reason, investigation of major tax reforms was left out of this research.

References

Aaberge, R., Bjorklund, A., Jantti, M., Palme, M., Peder, J., Smith, N., and Wennemo, T. (2002). Income inequality and income mobility in the scandinavian countries compared to the us. *Review of Income and Wealth*, 4(48):443–469.

Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology. *Statistics and Probability Letters*, 56(3):289–300.

Alderson, A. S. and Nielsen, F. (2002). Globalization and the great u-turn: Income inequality trends in 16 oecd countries. *American Journal of Sociology*, 107(5):1244–1299.

Aquaro, M. and Cizek, P. (2013). Robust estimation of dynamic fixed-effects panel data models. *Statistical Papers*, 55(1):169–186.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *Review of Economic Studies*, 58:277–297.

Aristei, D. and Perugin, C. (2015). The drivers of income mobility in europe. *Economic Systems*, 39(2):197–224.

Behr, A., Bellgardt, E., and Rendtel, U. (2005). Extent and determinants of panel attrition in the european community household panel. *European Sociological Review*, 21(5):489–512.

Blackburn, M. and Neumark, D. (1993). Omitted-ability bias and the increase in the return to schooling. *Journal of Labor Economics*, 11(3):521–544.

Block, J. H., Hoogerheide, L., and Thurik, R. (2010). Are education and entrepreneurial income endogenous and do family background variables make sense as instruments? a bayesian analysis. *Tinbergen Institute Discussion Paper*, 10-024/4.

Blundell, R. W. and Bond, S. R. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87:115143.

Blundell, R. W., Bond, S. R., and Windmeijer, F. (2000). Estimation in dynamic panel data models: Improving on performance of the standard gmm estimators. *The Institute for Fiscal Studies*, Working Paper 00-12.

Bond, S. (2002). Dynamic panel data models: a guide to micro data methods and practice. *Portuguese Economic Journal*, 1:141.

Burgess, S. and Propper, C. (1998). An economic model of household income dynamics, with an application to poverty dynamics among american women. *Discussion Paper No. 1830, Centre for Economic Policy Research, London.*

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

Cingano, F. (2014). Trends in income inequality and its impact on economic growth. *OECD Social, Employment and Migration Working Papers*, No. 163.

Clark, A., Etil, F., Postel-Vinay, F., Senik, C., and der Straeten K, V. (2005). Heterogeneity in reported well-being: Evidence from twelve european countries. *The Economic Journal*, 115(502):C118–C132.

Corak, M. (2006). Do poor children become poor adults? lessons from a cross country comparison of generational earnings mobility. *IZA Discussion Paper No. 1993*.

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

Dhaene, G. and Zhu, Y. (2009). Median-based estimation of dynamic panel models with fixed effects. *Computational Statistics and Data Analysis*, 113:398–423.

DiPrete, T. A. and McManus, P. A. (1999). The sensitivity of family income to changes in family structures and job change in the united states and germany. *Vierteljahrshefte zur Wirtschaftsforschung*, 68(2):171–176.

DiPrete, T. A. and McManus, P. A. (2000). Family change, employment transitions, and the welfare state: Household income dynamics in the united states and germany. *American Sociological Review*, 65(3):343–370.

Dynan, K., Elmendorf, D., and Sichel, D. (2012). The evolution of household income volatility. *The B.E. Journal of Economic Analysis and Policy*, 12(2):Article 3.

Fields, G., Cichello, P., Freije, S., Menendez, M., and Newhouse, D. (2003). Household income dynamics: a four country study. *The Journal of Development Studies*, 40(2):30–54.

Fitzgerald, J. (2011). Attrition in models of intergenerational links in health and economic status in the psid. *Berkeley Electronic Journal of Economic Analysis and Policy*, 11(3):Article 2.

Frick, J. R., Jenkins, S. P., Lillard, D. R., Lipps, O., and Wooden, M. (2007). The cross-national equivalent file (cnef) and its member country household panel studies. *Schmollers Jahrbuch*, 127:627–654.

Glewwe, P. and Hall, G. (1998). Are some groups more vulnerable to macroeconomic shocks than others? hypothesis tests based on panel data from peru. *Journal of Development Economics*, 56(1):181–206.

Guvenen, F., Karahan, F., Ozkan, S., and Song, J. (2015). What do data on millions of u.s. workers reveal about life-cycle earnings risk? *NBER Working Paper*, 20913.

Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21(3):618–637.

Jalan, J. and Ravallion, M. (2002). Household income dynamics in rural china. *WIDER Discussion Papers // World Institute for Development Economics (UNUWIDER)*, No. 2002/10.

Jamshidian, M. and Jennrich, R. I. (1997). Acceleration of the em algorithm by using quasi-newton methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3):569–587.

Jenkins, S. (2000). Modelling household income dynamics. *Journal of Population economics*, 13(529).

Kentor, J. (2001). The long term effects of globalization on income inequality, population growth, and economic development. *Social Problems*, 48(4):435–455.

Kowarik, A. and Templ, M. (2016). Imputation with the r package vim. *Journal of Statistical Software*, 74(7).

Lillard, L. A. and Panis, C. W. A. (1998). Panel attrition from the panel study of income dynamics: Household income, marital status, and mortality. *The Journal of Human Resources*, 33(2):437–457.

Lipps, O. (2007). Attrition in the swiss household panel. *Methoden, Daten, Analysen*, 1:45–68.

Lokshin, M. and Ravallion, M. (2004). Household income dynamics in two transition economies. *Studies in Nonlinear Dynamics and Econometrics*, 8(3):Article 4.

Lucas, A., van Dijk, R., and Kloek, T. (1997). Outlier robust gmm estimation of leverage determinants in linear dynamic panel data models.

Mehrhoff, J. (2009). A solution to the problem of too many instruments in dynamic panel data gmm. *Discussion Paper Series 1: Economic Studies No 31/2009*.

Okun, A. M. (1975). *Equality and Efficiency: The Big Tradeoff*. The Brookings Institution.

Ostry, J. D., Berg, A., and Tsangarides, C. G. (2014). Redistribution, inequality, and growth. *IMF staff discussion note*.

Roed, K. and Strom, S. (1999). Progressive taxes and the labour market: Is the trade-off between equality and efficiency inevitable? *Memorandum, Department of Economics, University of Oslo*, No. 1999,19.

Ronchetti, E. (1985). Robust model selection in regression. *Statistics and Probability Letters*, 3:21–23.

Ronchetti, E. and Staudte, R. (1994). A robust version of mallowss cp. *Journal of the American Statistical Association*, 89:550–559.

Roodman, D. (2009). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics*, 71(1).

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, pages 283–297.

Sommer, S. and Staudte, R. (1995). Robust variable selection in regression in the presence of outliers and leverage points. *Australian Journal of Statistics*, 37:323–336.

Titterington, D. M., Smith, A. F. M., and E, M. U. (1985). *Statistical analysis of finite mixture distributions*. Wiley, New York.

van Ewijk, C., Jacobs, B., de Mooij, R., and P, T. (2003). Tien doelmatigheidsargumenten voor progressieve belastingen. *Tijdschrift voor Openbare Financin*, 35(1):30–35.

Woolard, I. and Klasen, S. (2005). Determinants of income mobility and household poverty dynamics in south africa. *The Journal of Development Studies*, 41(5):865–897.

A Tables

A.1 Summary statistics

	Germany	Korea	Switzerland	UK	US
mean	0.009	0.087	0.010	0.041	0.013
median	0.004	0.024	0.005	0.023	0.012
standard deviation	0.358	1.220	0.945	0.853	0.588
MAD	0.143	0.353	0.196	0.239	0.232
skewness	0.789	1.426	1.051	0.626	-0.077
kurtosis	208.557	23.656	94.469	83.873	73.301
outer 2.5%	0.152	0.148	0.175	0.153	0.125

Table 9: Summary statistics of household real income growth

	mean	median	stan. dev.	MAD	% missing
number of children	0.53	0.00	0.89	0.00	1.02
post-gov. income	27649.76	23526.17	22362.01	14350.33	1.02
labor income (house)	26728.55	20400.00	32658.37	30245.04	1.02
pre-gov. income	29204.71	22231.00	38945.73	30839.56	1.02
age	49.33	48.00	17.70	19.27	1.33
labor income (person)	18205.64	12570.10	23320.45	18636.44	1.01
education level	1.98	2.00	0.63	0.00	3.51
education years	11.62	11.00	2.53	1.48	5.02
working hours	1197.16	1182.00	1092.40	1713.89	1.31

Table 10: Summary statistics of German data, income values are in euros

	mean	median	stan. dev.	MAD	% missing
number of children	1.30	1.00	1.03	1.48	2.29
post-gov. income	3197.84	2480.00	3253.70	1971.86	2.29
labor income (house)	2754.33	2340.00	2692.67	2134.94	2.29
pre-gov. income	3526.84	2829.00	3591.75	2400.33	17.63
age	46.65	45.00	14.64	14.83	2.13
labor income (person)	1878.14	1440.00	2478.08	1423.30	5.65

Table 11: Summary statistics of Korean data, income values are in 10.000 Korean won

	mean	median	stan. dev.	MAD	% missing
number of children	0.74	0.00	1.05	0.00	3.28
post-gov. income	80921.56	72151.53	63670.22	41414.15	3.28
labor income (house)	90287.38	83200.00	99925.96	88659.48	3.28
pre-gov. income	96321.17	86800.00	112129.73	85528.29	3.28
age	47.18	46.00	16.35	16.31	3.06
labor income (person)	54354.20	46456.00	71770.73	58337.34	3.06

Table 12: Summary statistics of Swiss data, income values are in Swiss francs

	mean	median	stan. dev.	MAD	% missing
number of children	0.72	0.00	1.05	0.00	4.21
post-gov. income	24462.42	18941.75	85946.96	13109.46	18.11
labor income (house)	18685.09	12009.24	23957.32	17804.90	5.80
pre-gov. income	25300.18	19815.24	85351.75	20419.56	14.17
age	48.46	47.00	18.29	19.27	4.01
labor income (person)	11939.52	7890.07	15985.18	11697.82	7.38

Table 13: Summary statistics of UK data, income values are in pounds

	mean	median	stan. dev.	MAD	% missing
number of children	0.88	0.00	1.24	0.00	18.42
post-gov. income	30569.98	20313.30	44474.66	17586.07	18.61
labor income (house)	31560.64	17500.00	61144.46	25631.19	18.61
pre-gov. income	36705.47	21400.00	66679.43	23532.98	18.62
age	43.60	42.00	18.01	19.27	17.58
labor income (person)	19349.68	9500.00	44839.57	14084.70	17.70
education level	2.16	2.00	0.79	1.48	18.36
education years	12.59	12.00	2.78	2.97	21.39
working hours	1432.28	1715.00	1131.77	1065.99	17.68

Table 14: Summary statistics of US data, income values are in dollars

A.2 Coefficient tables

The coefficient tables give the coefficients, standardized errors and p-values of the eight distinct models. The standardized errors shown in these tables are the regular standard errors obtained via the asymptotic property given in equation 40. Next to this, two p-values are given in each table (p-value and boot p-value). As the names already suggest, the first p-value is obtained via the regular standard errors, while for the latter all standard errors are multiplied with the uncertainty inflation factor caused by missing values of 3.16. Then, each number at the end of a segment specific explanatory variable corresponds to its segment number, whereby the segments are ordered according to their size. Lastly, some summary statistics are given on the right side of each table.

	coef.	std. error	p-value	boot p-val.		
Δ Income$_{t-1}$ 1	-0.604	0.007	0.000	0.000	Adj. likelihood	-84370.9
GDP growth 1	-0.178	0.214	0.404	0.791	s.d. seg. prob.	0.000
Inequality level 1	-5.234	0.349	0.000	0.000	Sargan test	1.000
Redistribution 1	-0.079	0.526	0.881	0.962	% Outlier 0	0.189
Δ Income$_{t-1}$ 2	0.690	0.019	0.000	0.000	% Outlier < 1	0.294
GDP growth 2	1.380	0.313	0.000	0.163		
Inequality level 2	4.483	0.458	0.000	0.002	σ_1	0.427
Redistribution 2	0.420	0.587	0.474	0.820	σ_2	0.412
Age	-0.015	0.001	0.000	0.000		
Education years	-0.842	0.089	0.000	0.003		
Work hours growth	-0.379	0.013	0.000	0.000		

Table 15: Coefficient table of analysis 1 with 2 segments

	coef.	std. error	p-value	boot p-val.		
Δ Income$_{t-1}$ 1	1.131	0.014	0.000	0.000	Adj. likelihood	-105082.7
GDP growth 1	-2.333	0.448	0.000	0.099	s.d. seg. prob.	0.007
Inequality level 1	0.836	0.427	0.050	0.535	Sargan test	1.000
Redistribution 1	6.523	0.235	0.000	0.000	% Outlier 0	0.171
Δ Income$_{t-1}$ 2	-0.505	0.016	0.000	0.000	% Outlier < 1	0.266
GDP growth 2	5.149	0.365	0.000	0.000		
Inequality level 2	-0.287	0.621	0.643	0.883	σ_1	0.504
Redistribution 2	-10.540	1.116	0.000	0.003	σ_2	0.385
Δ Income$_{t-1}$ 3	-1.076	0.029	0.000	0.000	σ_3	0.704
GDP growth 3	-3.875	0.650	0.000	0.059		
Inequality level 3	-4.793	1.154	0.000	0.188		
Redistribution 3	12.402	1.431	0.000	0.006		
Age	-0.005	0.001	0.000	0.006		
Education years	0.023	0.108	0.830	0.946		
Work hours growth	0.000	0.006	0.969	0.990		

Table 16: Coefficient table of analysis 1 with 3 segments

	coef.	std. error	p-value	boot p-val.		
Δ Income$_{t-1}$ 1	-0.385	0.007	0.000	0.000	Adj. likelihood	-57439.3
GDP growth 1	-0.094	0.077	0.223	0.699	s.d. seg. prob.	0.001
Inequality level 1	1.369	0.186	0.000	0.019	Sargan test	1.000
Redistribution 1	1.281	0.291	0.000	0.164	% Outlier 0	0.148
Δ Income$_{t-1}$ 2	0.798	0.029	0.000	0.000	% Outlier < 1	0.235
GDP growth 2	1.653	0.537	0.002	0.329		
Inequality level 2	-15.138	1.136	0.000	0.000	σ_1	0.277
Redistribution 2	-4.580	1.800	0.011	0.420	σ_2	0.851
Age	-0.010	0.001	0.000	0.000		
Education years	-0.213	0.116	0.066	0.560		
Work hours growth	-0.213	0.012	0.000	0.000		

Table 17: Coefficient table of analysis 2 with 2 segments

	coef.	std. error	p-value	boot p-val.		
Δ Income$_{t-1}$ 1	-0.087	0.012	0.000	0.026	Adj. likelihood	-63427.1
GDP growth 1	-0.463	0.157	0.003	0.351	s.d. seg. prob.	0.029
Inequality level 1	-1.669	0.264	0.000	0.045	Sargan test	1.000
Redistribution 1	-0.513	0.463	0.269	0.726	% Outlier 0	0.146
Δ Income$_{t-1}$ 2	0.264	0.014	0.000	0.000	% Outlier < 1	0.220
GDP growth 2	4.280	0.462	0.000	0.003		
Inequality level 2	3.715	0.673	0.000	0.080	σ_1	0.265
Redistribution 2	2.021	1.182	0.087	0.588	σ_2	0.535
Δ Income$_{t-1}$ 3	-0.369	0.033	0.000	0.000	σ_3	0.548
GDP growth 3	-3.644	1.301	0.005	0.375		
Inequality level 3	-0.925	1.931	0.632	0.879		
Redistribution 3	10.679	3.706	0.004	0.361		
Age	-0.008	0.001	0.000	0.001		
Education years	0.176	0.110	0.110	0.612		
Work hours growth	0.133	0.011	0.000	0.000		

Table 18: Coefficient table of analysis 2 with 3 segments

	coef.	std. error	p-value	boot p-val.		
Δ Income$_{t-1}$ 1	-0.311	0.013	0.000	0.000	Adj. likelihood	-53992.7
GDP growth 1	-0.627	0.060	0.000	0.001	s.d. seg. prob.	0.000
Inequality level 1	1.967	0.142	0.000	0.000	Sargan test	1.000
Redistribution 1	0.238	0.183	0.194	0.680	% Outlier 0	0.163
Δ Income$_{t-1}$ 2	-0.102	0.013	0.000	0.014	% Outlier < 1	0.261
GDP growth 2	1.855	0.151	0.000	0.000		
Inequality level 2	-4.225	0.236	0.000	0.000	σ_1	0.239
Redistribution 2	-0.134	0.302	0.657	0.888	σ_2	0.443
Age	-0.008	0.001	0.000	0.000		

Table 19: Coefficient table of analysis 3 with 2 segments

	coef.	std. error	p-value	boot p-val.		
Δ Income$_{t-1}$ 1	0.025	0.024	0.308	0.747	Adj. likelihood	-118080.3
GDP growth 1	-0.264	0.090	0.003	0.354	s.d. seg. prob.	0.036
Inequality level 1	2.345	0.258	0.000	0.004	Sargan test	1.000
Redistribution 1	-0.065	0.245	0.791	0.933	% Outlier 0	0.148
Δ Income$_{t-1}$ 2	0.191	0.021	0.000	0.004	% Outlier < 1	0.249
GDP growth 2	0.900	0.310	0.004	0.359		
Inequality level 2	-6.867	0.483	0.000	0.000	σ_1	0.279
Redistribution 2	4.516	0.600	0.000	0.017	σ_2	0.527
Δ Income$_{t-1}$ 3	-1.582	0.035	0.000	0.000	σ_3	1.599
GDP growth 3	3.974	1.801	0.027	0.484		
Inequality level 3	1.045	1.486	0.482	0.824		
Redistribution 3	-11.650	3.002	0.000	0.219		
Age	-0.012	0.001	0.000	0.000		

Table 20: Coefficient table of analysis 3 with 3 segments

	coef.	std. error	p-value	boot p-val.		
Δ Income$_{t-1}$ 1	0.302	0.029	0.000	0.001	Adj. likelihood	-170154.8
GDP growth 1	0.226	0.060	0.000	0.235	s.d. seg. prob.	0.001
Inequality level 1	1.009	0.171	0.000	0.061	Sargan test	1.000
Redistribution 1	1.640	0.296	0.000	0.079	% Outlier 0	0.174
Δ Income$_{t-1}$ 2	-1.150	0.027	0.000	0.000	% Outlier < 1	0.270
GDP growth 2	-1.148	0.234	0.000	0.120		
Inequality level 2	-3.713	0.400	0.000	0.003	σ_1	0.365
Redistribution 2	-1.780	0.621	0.004	0.364	σ_2	0.973
Age	-0.005	0.001	0.000	0.130		

Table 21: Coefficient table of analysis 4 with 2 segments

	coef.	std. error	p-value	boot p-val.		
Δ Income$_{t-1}$ 1	-0.401	0.016	0.000	0.000	Adj. likelihood	-105995.8
GDP growth 1	-0.877	0.140	0.000	0.047	s.d. seg. prob.	0.049
Inequality level 1	-2.416	0.214	0.000	0.000	Sargan test	1.000
Redistribution 1	0.833	0.267	0.002	0.322	% Outlier 0	0.199
Δ Income$_{t-1}$ 2	0.452	0.054	0.000	0.009	% Outlier < 1	0.321
GDP growth 2	1.117	0.110	0.000	0.001		
Inequality level 2	2.043	0.264	0.000	0.014	σ_1	0.436
Redistribution 2	3.663	0.450	0.000	0.010	σ_2	0.342
Δ Income$_{t-1}$ 3	-0.224	0.154	0.144	0.644	σ_3	0.298
GDP growth 3	-0.347	0.425	0.414	0.796		
Inequality level 3	0.532	0.390	0.172	0.665		
Redistribution 3	-11.177	1.474	0.000	0.016		
Age	-0.004	0.001	0.000	0.071		

Table 22: Coefficient table of analysis 4 with 3 segments

A.3 Segmentation tables

The segmentation tables give the segment probability of each segment together with some segment probabilities of subgroups. To be more precise, subgroups are created by splitting all households according to their value for a certain variable (e.g. age or average disposable income). Eventually, subgroup segment probabilities are calculated by taking the mean over the segment probabilities of all households within that subgroup.

	group 1	group 2	group 1	group 2	group 3
segment probability	0.573	0.427	0.473	0.359	0.168
mainly labor income	0.552	0.448	0.519	0.358	0.123
mainly non labor income	0.567	0.433	0.508	0.370	0.121
mainly government income	0.629	0.371	0.412	0.365	0.223
no job	0.688	0.312	0.343	0.380	0.277
young (<35)	0.604	0.396	0.443	0.370	0.187
middle (35-65)	0.592	0.408	0.473	0.371	0.156
old (>65)	0.556	0.444	0.505	0.370	0.125
lowest income	0.680	0.320	0.331	0.324	0.344
highest income	0.507	0.493	0.565	0.351	0.084
volatile	0.848	0.152	0.122	0.446	0.432
volatile low income	0.848	0.152	0.114	0.398	0.488
volatile middle income	0.850	0.150	0.130	0.493	0.378
volatile high income	0.849	0.151	0.139	0.539	0.321
stable	0.309	0.691	0.779	0.192	0.029
USA	0.573	0.427	0.473	0.359	0.168

Table 23: Segmentation table of analysis 1, results of the model with two segments are in the two left-hand columns and results of the model with three segments are in the three right hand columns

	group 1	group 2	group 1	group 2	group 3
segment probability	0.831	0.169	0.645	0.285	0.070
mainly labor income	0.843	0.157	0.666	0.273	0.061
mainly non labor income	0.900	0.100	0.652	0.312	0.036
mainly government income	0.885	0.115	0.677	0.263	0.060
no job	0.800	0.200	0.523	0.366	0.111
young (<35)	0.775	0.225	0.591	0.314	0.096
middle (35-65)	0.830	0.170	0.632	0.295	0.073
old (>65)	0.927	0.073	0.724	0.246	0.030
lowest income	0.714	0.286	0.497	0.340	0.163
highest income	0.884	0.116	0.703	0.257	0.040
volatile	0.676	0.324	0.309	0.530	0.161
volatile low income	0.615	0.385	0.284	0.502	0.214
volatile middle income	0.719	0.281	0.326	0.553	0.120
volatile high income	0.761	0.239	0.345	0.558	0.097
stable	0.932	0.068	0.872	0.106	0.022
USA	0.797	0.203	0.608	0.302	0.090
Germany	0.865	0.135	0.666	0.281	0.054

Table 24: Segmentation table of analysis 2, results of the model with two segments are in the two left-hand columns and results of the model with three segments are in the three right hand columns

	group 1	group 2	group 1	group 2	group 3
segment probability	0.598	0.402	0.615	0.327	0.058
mainly labor income	0.667	0.333	0.688	0.273	0.040
mainly non labor income	0.558	0.442	0.520	0.402	0.078
mainly government income	0.603	0.397	0.588	0.326	0.085
no job	0.464	0.536	0.433	0.429	0.139
young (<35)	0.593	0.407	0.601	0.347	0.053
middle (35-65)	0.622	0.378	0.627	0.315	0.058
old (>65)	0.656	0.344	0.643	0.299	0.058
lowest income	0.442	0.558	0.425	0.413	0.162
highest income	0.612	0.388	0.623	0.311	0.066
volatile	0.283	0.717	0.201	0.594	0.205
volatile low income	0.229	0.771	0.154	0.578	0.269
volatile middle income	0.315	0.685	0.231	0.607	0.162
volatile high income	0.305	0.695	0.215	0.591	0.194
stable	0.837	0.163	0.913	0.085	0.002
USA	0.532	0.468	0.538	0.402	0.060
Germany	0.767	0.233	0.758	0.219	0.023
UK	0.543	0.457	0.555	0.357	0.088

Table 25: Segmentation table of analysis 3, results of the model with two segments are in the two left-hand columns and results of the model with three segments are in the three right hand columns

	group 1	group 2	group 1	group 2	group 3
segment probability	0.686	0.314	0.580	0.368	0.052
mainly labor income	0.755	0.245	0.527	0.420	0.054
mainly non labor income	0.613	0.387	0.671	0.267	0.062
mainly government income	0.669	0.331	0.577	0.367	0.056
no job	0.512	0.488	0.717	0.228	0.054
young (<35)	0.684	0.316	0.605	0.338	0.057
middle (35-65)	0.699	0.301	0.581	0.375	0.044
old (>65)	0.698	0.302	0.575	0.375	0.050
lowest income	0.482	0.518	0.706	0.239	0.056
highest income	0.693	0.307	0.553	0.384	0.063
volatile	0.220	0.780	0.877	0.082	0.041
volatile low income	0.171	0.829	0.901	0.061	0.038
volatile middle income	0.253	0.747	0.862	0.096	0.043
volatile high income	0.227	0.773	0.872	0.087	0.042
stable	0.950	0.050	0.209	0.672	0.118
USA	0.756	0.244	0.600	0.373	0.027
Germany	0.859	0.141	0.379	0.579	0.042
UK	0.632	0.368	0.654	0.258	0.088
Korea	0.432	0.568	0.811	0.166	0.023
Switzerland	0.703	0.297	0.570	0.397	0.033

Table 26: Segmentation table of analysis 4, results of the model with two segments are in the two left-hand columns and results of the model with three segments are in the three right hand columns

A.4 Quantile tables

The quantile tables contain weighted quantile values for household averages of some variables over the different segments. Hence, the second row of this table contains the segment numbers for which the weighted quantiles are determined. These weighted quantile values are calculated by using the segment probability multiplied with the number of observations of a certain household as weight. In other words, the idea of this table is to give a summary of the shape of the distribution of a certain variable for a certain segment.

		1%	10%	25%	50%	75%	90%	99%
post-government income	1	0.146	0.331	0.562	0.978	1.502	2.108	4.284
	2	0.208	0.420	0.727	1.210	1.729	2.256	3.894
pre-government income	1	0.007	0.149	0.443	0.966	1.670	2.521	5.824
	2	0.003	0.182	0.642	1.288	1.982	2.722	5.044
labor income	1	0.000	0.079	0.381	0.960	1.693	2.551	5.584
	2	0.000	0.035	0.579	1.323	2.066	2.854	5.098
income growth	1	-0.397	-0.069	-0.026	0.005	0.041	0.101	0.385
	2	-0.199	-0.050	-0.017	0.011	0.043	0.092	0.312

Table 27: Quantile table of analysis 1 with 2 segments

		1%	10%	25%	50%	75%	90%	99%
post-government income	1	0.209	0.439	0.751	1.231	1.738	2.287	3.932
	2	0.184	0.373	0.624	1.032	1.556	2.157	4.298
	3	0.089	0.240	0.393	0.705	1.198	1.717	4.190
pre-government income	1	0.003	0.194	0.681	1.323	2.011	2.764	5.254
	2	0.009	0.177	0.499	1.028	1.723	2.602	5.824
	3	0.003	0.112	0.285	0.653	1.257	2.001	5.704
labor income	1	0.000	0.039	0.624	1.357	2.088	2.873	5.178
	2	0.000	0.085	0.432	1.025	1.743	2.602	5.724
	3	0.000	0.070	0.262	0.622	1.258	2.004	4.943
income growth	1	-0.178	-0.048	-0.016	0.011	0.042	0.087	0.267
	2	-0.282	-0.063	-0.024	0.006	0.042	0.103	0.359
	3	-0.838	-0.117	-0.037	0.003	0.044	0.121	0.773

Table 28: Quantile table of analysis 1 with 3 segments

		1%	10%	25%	50%	75%	90%	99%
post-government income	1	0.204	0.427	0.694	1.086	1.542	2.114	4.074
	2	0.117	0.308	0.508	0.866	1.304	1.886	3.725
pre-government income	1	0.001	0.091	0.495	1.108	1.766	2.592	5.488
	2	0.009	0.173	0.414	0.858	1.457	2.222	4.917
labor income	1	0.000	0.001	0.398	1.112	1.810	2.660	5.305
	2	0.000	0.149	0.395	0.852	1.468	2.245	4.902
income growth	1	-0.190	-0.050	-0.018	0.009	0.038	0.082	0.283
	2	-0.670	-0.078	-0.023	0.010	0.052	0.132	0.691

Table 29: Quantile table of analysis 2 with 2 segments

		1%	10%	25%	50%	75%	90%	99%
post-government income	1	0.218	0.457	0.726	1.116	1.570	2.138	4.055
	2	0.161	0.361	0.598	0.958	1.424	2.029	4.082
	3	0.084	0.248	0.410	0.769	1.237	1.786	3.765
pre-government income	1	0.001	0.084	0.526	1.157	1.808	2.639	5.458
	2	0.003	0.144	0.451	0.944	1.581	2.408	5.546
	3	0.005	0.136	0.327	0.755	1.338	2.080	4.700
labor income	1	0.000	0.000	0.433	1.172	1.865	2.714	5.275
	2	0.000	0.056	0.384	0.922	1.583	2.397	5.393
	3	0.000	0.117	0.312	0.755	1.361	2.121	4.185
income growth	1	-0.175	-0.044	-0.015	0.010	0.038	0.081	0.280
	2	-0.361	-0.070	-0.025	0.007	0.042	0.100	0.392
	3	-0.973	-0.078	-0.024	0.010	0.050	0.125	0.705

Table 30: Quantile table of analysis 2 with 3 segments

		1%	10%	25%	50%	75%	90%	99%
post-government income	1	0.264	0.520	0.754	1.102	1.505	2.006	3.522
	2	0.159	0.387	0.600	0.953	1.365	1.969	4.530
pre-government income	1	0.000	0.071	0.449	1.169	1.902	2.726	5.173
	2	0.000	0.088	0.343	0.916	1.699	2.660	6.395
labor income	1	0.000	0.000	0.273	1.201	2.054	3.025	5.702
	2	0.000	0.000	0.201	0.887	1.842	2.927	6.620
income growth	1	-0.190	-0.045	-0.014	0.013	0.044	0.090	0.276
	2	-0.608	-0.096	-0.030	0.015	0.068	0.158	0.806

Table 31: Quantile table of analysis 3 with 2 segments

		1%	10%	25%	50%	75%	90%	99%
post-government income	1	0.270	0.527	0.760	1.105	1.498	1.992	3.477
	2	0.187	0.407	0.625	0.970	1.395	1.993	4.272
	3	0.076	0.258	0.435	0.756	1.246	2.012	5.668
pre-government income	1	0.000	0.074	0.478	1.188	1.906	2.719	5.152
	2	0.000	0.093	0.345	0.908	1.700	2.658	6.119
	3	0.000	0.043	0.204	0.675	1.553	2.788	8.180
labor income	1	0.000	0.000	0.314	1.234	2.075	3.028	5.672
	2	0.000	0.000	0.206	0.859	1.811	2.901	6.603
	3	0.000	0.000	0.066	0.567	1.638	3.013	7.421
income growth	1	-0.192	-0.046	-0.014	0.012	0.043	0.084	0.239
	2	-0.399	-0.085	-0.027	0.017	0.072	0.161	0.541
	3	-2.510	-0.224	-0.042	0.022	0.095	0.316	2.978

Table 32: Quantile table of analysis 3 with 3 segments

		1%	10%	25%	50%	75%	90%	99%
post-government income	1	0.244	0.509	0.746	1.096	1.505	2.022	3.644
	2	0.121	0.337	0.555	0.910	1.333	1.979	4.587
pre-government income	1	0.000	0.066	0.458	1.164	1.883	2.720	5.287
	2	0.000	0.088	0.327	0.851	1.575	2.541	6.254
labor income	1	0.000	0.000	0.292	1.195	2.026	3.016	5.841
	2	0.000	0.000	0.177	0.787	1.610	2.715	6.474
income growth	1	-0.205	-0.058	-0.019	0.011	0.048	0.095	0.262
	2	-0.861	-0.131	-0.043	0.020	0.092	0.252	2.039

Table 33: Quantile table of analysis 4 with 2 segments

		1%	10%	25%	50%	75%	90%	99%
post-government income	1	0.147	0.401	0.637	0.993	1.416	1.987	4.179
	2	0.257	0.522	0.765	1.118	1.533	2.059	3.619
	3	0.228	0.492	0.700	1.022	1.392	1.870	3.527
pre-government income	1	0.000	0.093	0.376	0.967	1.699	2.587	5.736
	2	0.000	0.046	0.476	1.221	1.941	2.773	5.269
	3	0.000	0.053	0.375	1.072	1.884	2.846	5.642
labor income	1	0.000	0.000	0.219	0.937	1.787	2.826	6.106
	2	0.000	0.000	0.313	1.262	2.071	3.045	5.828
	3	0.000	0.000	0.144	1.110	2.150	3.402	6.652
income growth	1	-0.488	-0.093	-0.030	0.018	0.072	0.160	1.169
	2	-0.191	-0.050	-0.018	0.008	0.039	0.081	0.245
	3	-0.645	-0.104	-0.030	0.017	0.067	0.134	0.390

Table 34: Quantile table of analysis 4 with 3 segments

B Figures

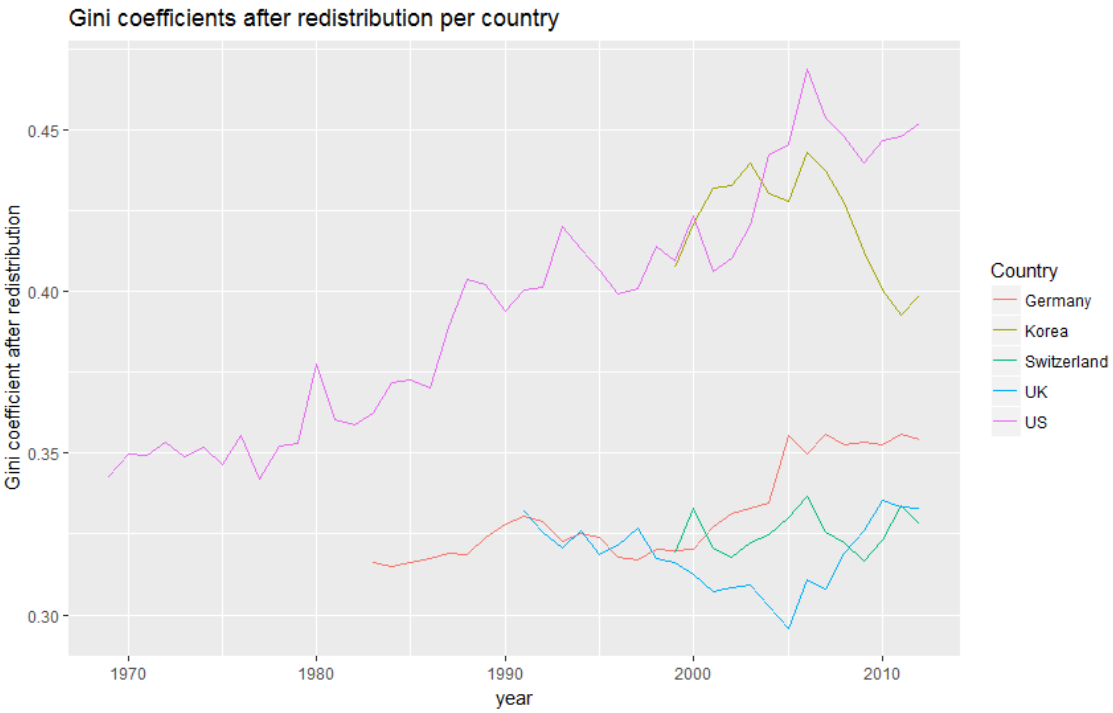

Figure 1: Gini coefficient after redistribution per year

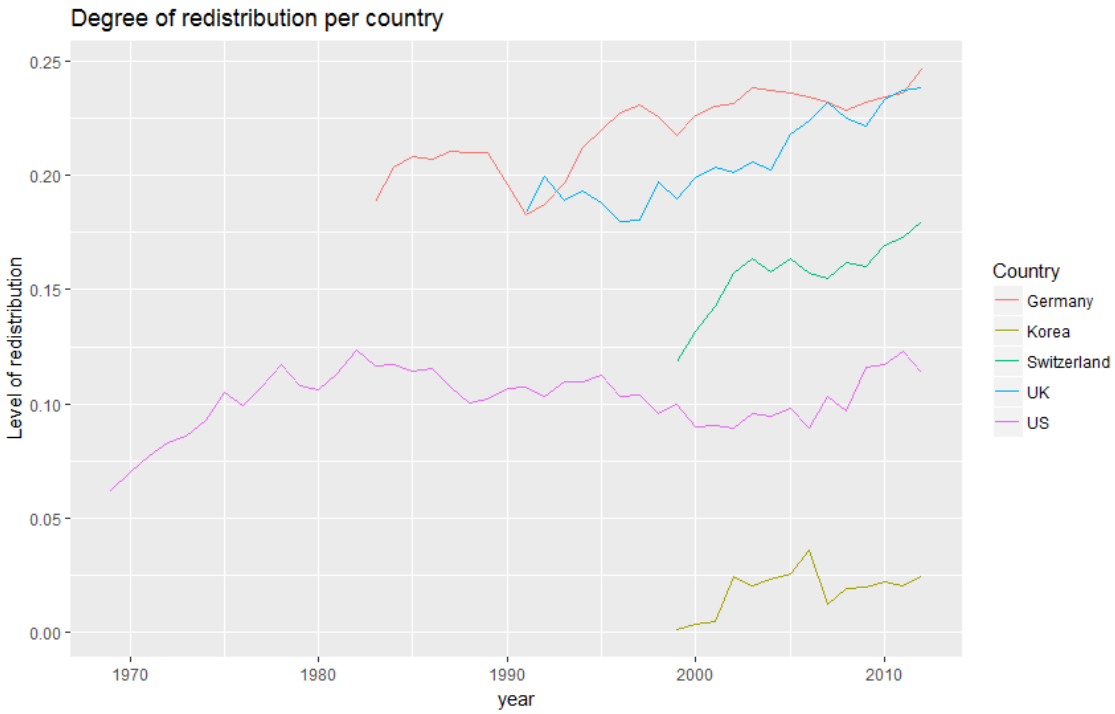

Figure 2: Redistribution per year

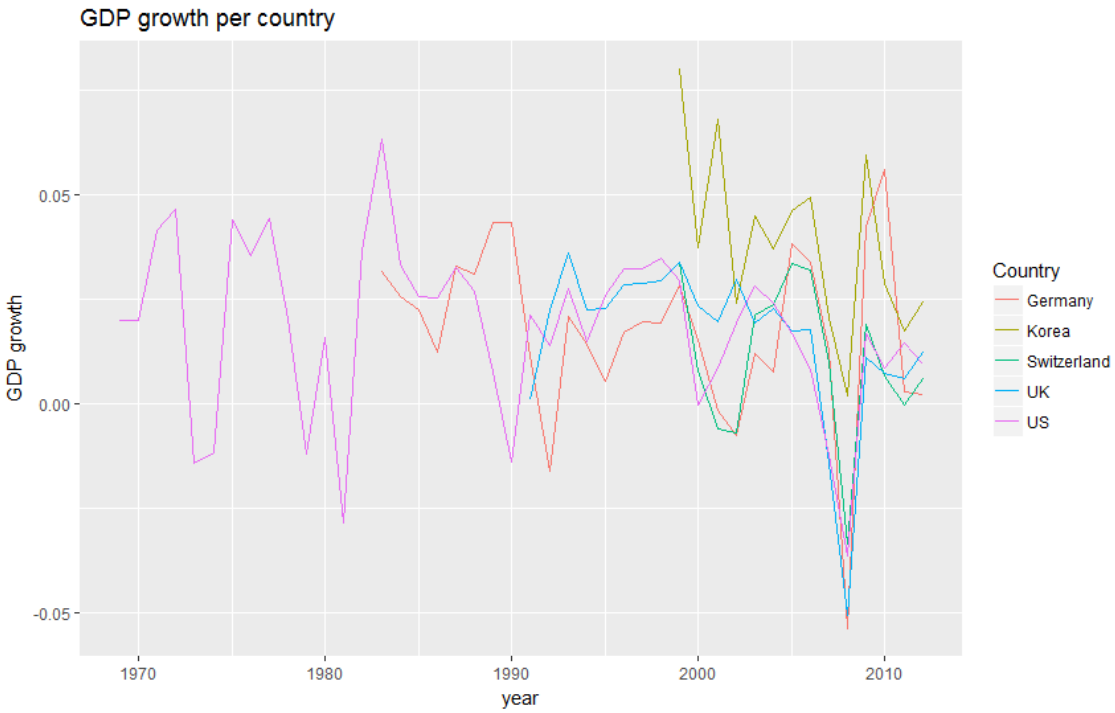

Figure 3: GDP growth per year