cemfi

working paper

2109

Income Risk Inequality: Evidence from Spanish Administrative Records

Manuel Arellano
Stéphane Bonhomme
Micole De Vera
Laura Hospido
Siqi Wei

September 2021

CENTRO DE ESTUDIOS
MONETARIOS Y FINANCIEROS
Casado del Alisal 5, 28014 Madrid, Spain
www.cemfi.es

CEMFI Working Paper No. 2109
September 2021

Income Risk Inequality: Evidence from Spanish Administrative Records

Abstract

In this paper we use administrative data from the social security to study income dynamics and income risk inequality in Spain between 2005 and 2018. We construct individual measures of income risk as functions of past employment history, income, and demographics. Focusing on males, we document that income risk is highly unequal in Spain: more than half of the economy has close to perfect predictability of their income, while some face considerable uncertainty. Income risk is inversely related to income and age, and income risk inequality increases markedly in the recession. These findings are robust to a variety of specifications, including using neural networks for prediction and allowing for individual unobserved heterogeneity.

JEL Codes: D31, E24, E31, J31.

Keywords: Spain, income dynamics, administrative data, income risk, inequality.

Manuel Arellano
CEMFI
arellano@cemfi.es

Stéphane Bonhomme
University of Chicago
sbonhomme@uchicago.edu

Micole De Vera
CEMFI
micole.devera@cemfi.edu.es

Laura Hospido
Banco de España and IZA
laura.hospido@bde.es

Siqi Wei
CEMFI
siqi.wei@cemfi.edu.es

Acknowledgement

This work is part of the Global Income Dynamics (GID) Project, set up by Fatih Guvenen, Luigi Pistaferri, and Gianluca Violante, to produce and maintain a harmonized cross-country database of statistics on income dynamics. The authors thank two anonymous referees, Fatih Guvenen, Mariacristina DeNardi, Clara Martinez-Toledano, Josep Pijoan-Mas, Luigi Pistaferri, Gianluca Violante, as well as seminar participants at Banco de España, CEMFI, IFS/UCL, TSE, University of Minnesota, University of Wisconsin-Madison, and the Global Income Dynamics Conferences for valuable comments and suggestions. We also thank Serdar Ozkan and Sergio Salgado for providing their code and support, and Roberto Ramos for his help with the computation of the effective tax rates. Roque Bescos and Yang Xun provided excellent research assistance. Bonhomme acknowledges support from NSF grant number SES-1658920. De Vera acknowledges funding from Spain's Ministerio de Ciencia, Innovación y Universidades (PRE2018-084485) and Ministerio de Economía, Industria y Competividad (María de Maeztu Programme for Units of Excellence in R&D, MDM-2016-0684). Wei acknowledges funding from Spain's Ministerio de Economía, Industria y Competitividad (BES-2017-082506), and María de Maeztu Programme for Units of Excellence in R&D (MDM-2016-0684). The opinions and analysis are the responsibility of the authors and, therefore, do not necessarily coincide with those of the Banco de España or the Eurosystem.

1 Introduction

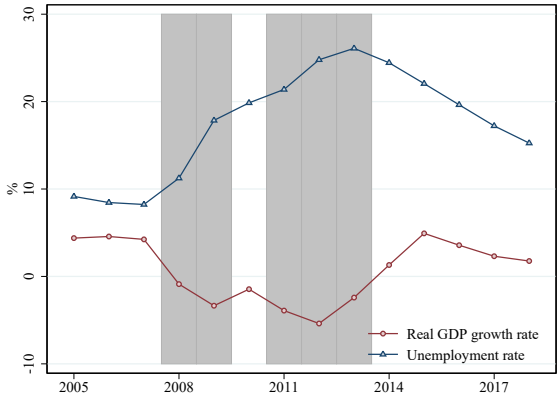
Income inequality is the focus of a large empirical literature, which now spans many countries over decades or centuries (Atkinson, 2003, Alvaredo et al., 2017). However, the measurement of cross-sectional inequality only provides an incomplete understanding of the diversity of individual income trajectories, since it cannot account for upward and downward mobility or the effect of economic shocks on individual careers.

The increased availability of longitudinal records on income and employment has motivated a related literature that concentrates on income dynamics. While a number of contributions are based on survey data (e.g., Gottschalk and Moffitt, 1994, Geweke and Keane, 2000, Meghir and Pistaferri, 2004, Browning et al., 2010, Arellano et al., 2017), there has been a recent surge in the use of administrative income records. Administrative data offers several advantages relative to surveys, such as large representative samples, complete employment spells over long horizons, and high-quality information. The use of administrative data has led to new findings about the dynamics of income, in the US and other countries (e.g., Guvenen et al., 2014, Guvenen et al., forthcoming, Busch et al., forthcoming).

A central motivation of the income dynamics literature is to quantify income risk. In many models and in real life, the ability to forecast one's future income is a key determinant of economic decisions. However, the way researchers measure income risk is usually indirect, based on statistical models of the dynamics of income. The nonparametric approach to income dynamics, which has been put forward in Guvenen et al. (forthcoming) and related work, produces statistics such as conditional moments of log income changes that are related to income risk, yet this approach does not target risk directly. In this paper, we develop a methodology for constructing measures of individual income risk.

We are interested in documenting income risk and uncertainty. Unpredictability of income can have a major impact on consumption and saving decisions (Deaton, 1992). We focus on annual income, although we note that within-year variations may also be relevant sources of income risk (Morduch and Schneider, 2019). Risk, as we define it, differs from income volatility and instability, which have been the focus of a number of studies (Haider, 2001, Gottschalk and Moffitt, 2009, Ziliak et al., 2011), and are at the center of a recent debate

Figure 1: Aggregate conditions in Spain



Notes: Spanish Statistical Office (Instituto Nacional de Estadística). The shaded areas indicate recession years.

in the US (Bloom et al., 2017). Income volatility is typically measured as the dispersion of the changes of log earnings, or of their transitory component. While we will also report such measures, they differ from income risk, which is the part of income changes that cannot be predicted by the agent. To construct individual measures of risk, we will try to capture key determinants of the agent's information set using administrative records.

Our empirical focus is the Spanish economy. The recent Spanish experience is characterized by a high level and large fluctuations of unemployment. In Figure 1 we report the unemployment rate (in triangles), together with real GDP growth (in circles), from 2005 to 2018. Using administrative social security records to study cross-sectional income inequality, Bonhomme and Hospido (2017) found that the double-dipped recession that started in 2008 saw a large increase in inequality (see also Anghel et al., 2018). However, the literature is silent on the nature and evolution of income dynamics in Spain. More broadly, we still lack a description and understanding of the large cross-sectional inequality in individual income risk, at given age and over the life cycle.

Against this background, our first goal is to document a novel set of facts about income dynamics in Spain. To this end, we exploit administrative tax records that were matched to the social security data, and are available since 2005. We are interested in documenting how income inequality and dynamics evolved in recent years. An important goal of this analysis

is to study the level and evolution of moments of the distribution of log income changes, such as dispersion and skewness. In doing so, we follow the model set by the Global Income Dynamics project, and applied to a number of other countries in this volume.

Our second and main goal is to quantify income risk, and to study the inequality of individual income security, taking the Spanish economy as a case study. Our premise is that some people can predict with almost certainty their income one year ahead, while others face considerable uncertainty. In Spain, inequality in income risk is related to the prevalence of high unemployment, but also to the large share of short-term temporary employment that produces high job turnover (Felgueroso et al., 2017). We develop a methodology for constructing measures of income risk as a function of social security employment records, past income, contract type, and demographics. Having obtained an index of individual income risk, we then study its cross-sectional distribution, its persistence, and how it changes with age and the aggregate conditions of the Spanish economy.

In the first part of the paper we focus on income inequality and dynamics. We find that inequality increases strongly in the recession, particularly for males. The increase in inequality characterizes the entire recession period, confirming previous findings in the literature. In addition, the recession is also characterized by an increase in the dispersion of year-to-year log earnings changes, and by a decrease in skewness. While there has been some debate about whether dispersion is countercyclical in the US (e.g., Storesletten et al., 2004, Guvenen et al., 2014), the procyclical skewness of changes in log annual earnings has been documented in several countries (see Busch et al., forthcoming, Hoffmann and Malacrino, 2019, Pora and Wilner, 2020).

In the second part of the paper we study income risk, its determinants, and its evolution. We measure income risk using prediction methods, based on a set of predictors at the individual and aggregate levels. Our main risk measure is a coefficient of variation (CV), computed as the ratio of the mean absolute deviation of income divided by the mean of income, both of them conditional on a set of predictors. For example, a worker with an expected income of 20,000 euros and a CV of 10 percent expects a deviation of her next year's income from its mean of $\pm$2000 euros. The CV is a feature of the predictive distribution of income. Under the assumption that our set of predictors exhausts the agent's information set, this predictive

distribution summarizes the income uncertainty that she faces. Using a calculation in the spirit of Lucas' measurement of the welfare cost of business cycles (Lucas, 1987), we show how, under certain assumptions, the squared CV can be related to how much consumption the agent would have to forgo in order to eliminate income risk. However, the macroeconomic consequences of individual variation in income risk of the magnitude attested by our results are yet to be explored.

The econometrics of measuring income risk is a prediction problem. In our baseline approach, we use as predictors aspects of income and employment history, contract type, and demographics, augmented with a set of indicators of the macroeconomic conditions at the national and provincial level. Our predictive models are based on exponential specifications, and we use Poisson regressions for estimation. Using a large set of predictors is important to compute a reliable risk measure. Indeed, using the final year of our data as a hold-out sample, we show that, relative to a specification solely based on lagged income, including additional predictors improves the prediction of income absolute deviations, the use of employment history being particularly informative.

We find that risk is highly unequal in Spain: more than half of the economy has close to perfect predictability of their income, while some face considerable uncertainty. We also document that the inequality of income risk, as measured by our CV, increases markedly in the recession. Notably, this behavior is only driven by the upper part of the risk distribution. More than half of the Spanish economy faces low levels of risk, which do not vary over the period. Risk affects disproportionately the young, and the individuals in the bottom part of the income distribution. In addition, risk is highly persistent over time: an individual in the bottom half of the risk distribution today is poised to face virtually no risk next year. Overall, these findings suggest that more than half of the Spanish economy is effectively shielded from income risk, whereas the other part of the economy is subject to high levels of risk.

Our risk measure depends on the quality of the predictors and prediction models that we use. We probe the robustness of our baseline approach in various ways. First, we replace the exponential regression models by neural network specifications. Neural networks are universal approximators, and they are increasingly used for flexible modeling (Hornik et al., 1989, Goodfellow et al., 2016, Farrell et al., 2021). Second, we estimate specifications that

allow for unobserved heterogeneity, in addition to observed predictors, following a discrete approach as in Bonhomme et al. (2021). Third, as complements to the CV, we compute quantile-based measures of risk. All these exercises confirm the basic findings obtained using our baseline method. In addition, while the analysis in most of the paper is based on pre-tax income, we show that accounting for the Spanish tax system in the income measure has little impact on our substantive findings. Lastly, we find that, in contrast with the rest of the economy, the CV of Spanish civil servants, who enjoy high levels of job and income security, are all concentrated around low values and do not vary over the period.

In the last part of the paper, we complement our CV measure of income risk, which is based on longitudinal administrative records and a prediction approach, by studying subjective income expectations as reported in survey data. Responses to probabilistic subjective expectations questions can be used to directly quantify the income risk faced by individuals, and thus provide a valuable complement to observational measures of risk (Dominitz and Manski, 1997, Kaufmann and Pistaferri, 2009, Arellano, 2014). By showing a broad agreement between our prediction-based measure and the subjective expectation-based measure, in spite of the many differences in their construction, our confidence in both measures increases. We rely on subjective income expectations questions from the Spanish Survey of Household Finances. Assuming a household-specific log normal random walk predictive income process, we estimate subjective standard deviations of income growth for every household in 2014. We find that, according to this measure, many households face relatively low levels of risk and there is substantial risk dispersion between households. In addition, similarly to our CV measure, subjective standard deviations tend to be higher for the young, and for households with low income.

The paper proceeds as follows. In Section 2 we describe the administrative dataset we use for the analysis. In Section 3 we report a set of facts on income dynamics in Spain. In Section 4 we describe how we measure individual income risk. In Section 5 we document the magnitude and evolution of income risk and income risk inequality in Spain. In Section 6 we compare our risk measure with subjective expectations data. Finally, we conclude in Section 7. An appendix contains additional results.

2 Data

Our main data source comes from the Continuous Work History Sample (Muestra Continua de Vidas Laborales, MCVL, in Spanish), which is a 4% non-stratified random sample from the Spanish population registered with the social security administration in the reference year. Since 2005, individuals who are present in a wave and subsequently remain registered with the social security administration stay as sample members. In addition, the sample is refreshed with new sample members so it remains representative of the population in each wave. To complement our main data source, we match social security employment histories with income tax and census records.

For each employment spell, we observe the start date and end date of the labor contract, the part-time or full-time status of the employee, the type of contract (temporary or permanent), and the sector of employment (public or private). We also observe some information about the establishment, including the province where it is registered and the industry. In addition, by linking the longitudinal data with census records, we have access to individual demographic characteristics such as age, gender, and highest educational attainment.

The MCVL records monthly social security contributions, going back to 1980, however these contributions are top and bottom coded. Since 2005, the MCVL is matched to data from the tax authority, which provides us with uncensored individual income from paid employment accumulated in a calendar year, as reported by employers to the tax authority, as well as unemployment benefits and subsidies.[1]

We focus our analysis on annual income. In the first part of the paper in Section 3, we focus on annual labor earnings from paid employment. In the second part starting in Section 4, we use a broader measure of earnings that also includes unemployment benefits and subsidies. All earnings measures are deflated to 2018 euros using the Spanish consumer price index.

The data we rely on have two main limitations. First, the period of observation is relatively short. As mentioned above, for the years prior to 2005, income records are top and

[1]The tax information comes from "model 190", the "Annual summary of retentions and payments for the personal income tax on earnings, economic activities, awards and income imputations." This form is required of all entities that pay wages, pensions or unemployment benefits. It covers all beneficiaries, including those whose wages fall below the legal minimum of exemption for the obligation to declare personal income taxes.

bottom coded, so we focus on the period 2005-2018 where we observe uncensored annual earnings from tax information. Second, the MCVL does not permit to link individuals to households. Hence, our study will necessarily be silent on within-household risk sharing and insurance.

Sample selection. We focus our analysis on workers who are between 25 and 55 years old, are not self-employed, and do not live in the Basque Country or Navarra (for which the tax data does not provide coverage). In the first part of the paper, following the GID project's conventions, we trim annual earnings below a threshold $\underline{y}_t$, which corresponds to working part-time for one quarter at the national minimum wage. This trimming is meant to avoid workers with weak attachment to the labor force. In Appendix Table F1 we report the percentage of observations below the income threshold. It is important to note that the proportion of observations below the threshold is quite large, and that it varies over the period. For this reason, to study income risk we will rely on a broader sample that includes individuals with low or zero annual earnings.

In our analysis of income dynamics in the first part of the paper, we refer to three samples. In the "CS" (cross-sectional) sample, we only impose the restrictions on age and minimum earnings. For the analyses that involve dynamics, we impose additional restrictions on the data and focus on two subsamples. The "LS" (longitudinal) sample only includes observations with non-missing one-year and five-year individual earnings changes. In turn, the "H" (heterogeneity) sample is further restricted to non-missing average earnings over the past three years.

In our analysis of individual income risk in the second part of the paper, we will primarily refer to the "B" (broader) sample, which extends our measure of earnings in two dimensions. First, we use a broader measure of income, which includes both earnings from paid work as well as unemployment benefits. Combining both sources of income allows us to speak towards risk in an earnings measure more relevant to individual consumption and investment decisions. While this income measure does not include other sources of taxes or transfers, which we do not observe in the MCVL, we will also report results based on after-tax income using a simple rule to impute tax amounts to the individuals in our data. Second, we do not

impose a threshold to trim the earnings; that is, we include earnings observations that fall below the threshold, including zeros.[2] A non-negligible share of the Spanish economy has annual earnings below $\underline{y}_t$. This is a salient margin of risk that we want to capture. At the same time, since labor force attachment is lower for females, and we do not have information on the household (e.g., spousal income), inferring income risk for females would raise major challenges. For this reason, we do not include females in the B sample, and we will focus our analysis of income risk on males only.

Descriptive statistics. We provide descriptive statistics about the samples in the appendix.[3] The number of observations and the composition of the sample vary over the period. Indeed, the recession years between 2008 and 2013 are associated with smaller sample sizes, which reflect lower participation to the labor market, and a somewhat older and more educated labor force. The share of females increases slightly, albeit steadily, during the period. Mean income tends to increase in the recession, particularly in the case of males. Moreover, while the percentiles at the bottom of the earnings distribution follow a U-shaped evolution, the earnings percentiles above the median vary little over the period.

3 Income inequality and income dynamics in Spain

In this section we report a set of statistics on the dynamics of income in the Spanish social security data. Here the core quantities are characteristics of the distributions of individual log earnings changes, as in Guvenen et al. (forthcoming) and work inspired by their empirical methodology.

[2] In the MCVL, we only know for sure that an individual is unemployed when she receives unemployment benefits. Years when an individual is not receiving paid work, self-employment income, unemployment benefits, or pension benefits, correspond to zero income. This may overstate the relevant zeros, since the individual may have exited the labor market, found work out of the country where the Social Security agency has no jurisdiction, have returned to further education, or have transitioned to self-employment without official registration. To alleviate this issue, we impose a maximum of two zeros after the end of any observed labor market spell (be it a contract for paid work or a spell of receiving unemployment benefits), and we drop all observations after the imposed maximum of two zeros. We also estimated our baseline specification on samples where we included those observations and treated them as zero income. We found qualitatively similar patterns, with a stronger income risk inequality increase in the recession.

[3] In Appendix Tables F2 and F3 we how summary statistics for the CS sample, and in Appendix Tables F5, F7, and F9 for the LS and H samples (both of them restricted to non-missing 1-year and 5-year changes in log earnings), and for the B sample, respectively. In Appendix Tables F4, F6, F8, and F10 we show the same summary statistics where we convert earnings to US Dollars using the 2018 exchange rate.

3.1 Income inequality

In Figure 2 we start by showing percentiles of log real earnings, by gender, from 2005 to 2018, taking 2005 as the reference year.[4] In the top two graphs, we show the 10th, 25th, median, 75th, and 90th percentiles for males and females, respectively. While the evolution of earnings percentiles over the period shows that earnings inequality increases in the recession, it also highlights a contrast between males and females. For males, earnings percentiles above the median vary little during the period, however the 10th and 25th percentiles drop sharply during the great recession, and only start to recover after 2013. As a result, earnings inequality increases in the recession. This confirms the findings documented in Bonhomme and Hospido (2017). For females, we observe a similar pattern, albeit quantitatively much less pronounced, in line with the findings of Bonhomme and Hospido (2013) on the first part of the period.

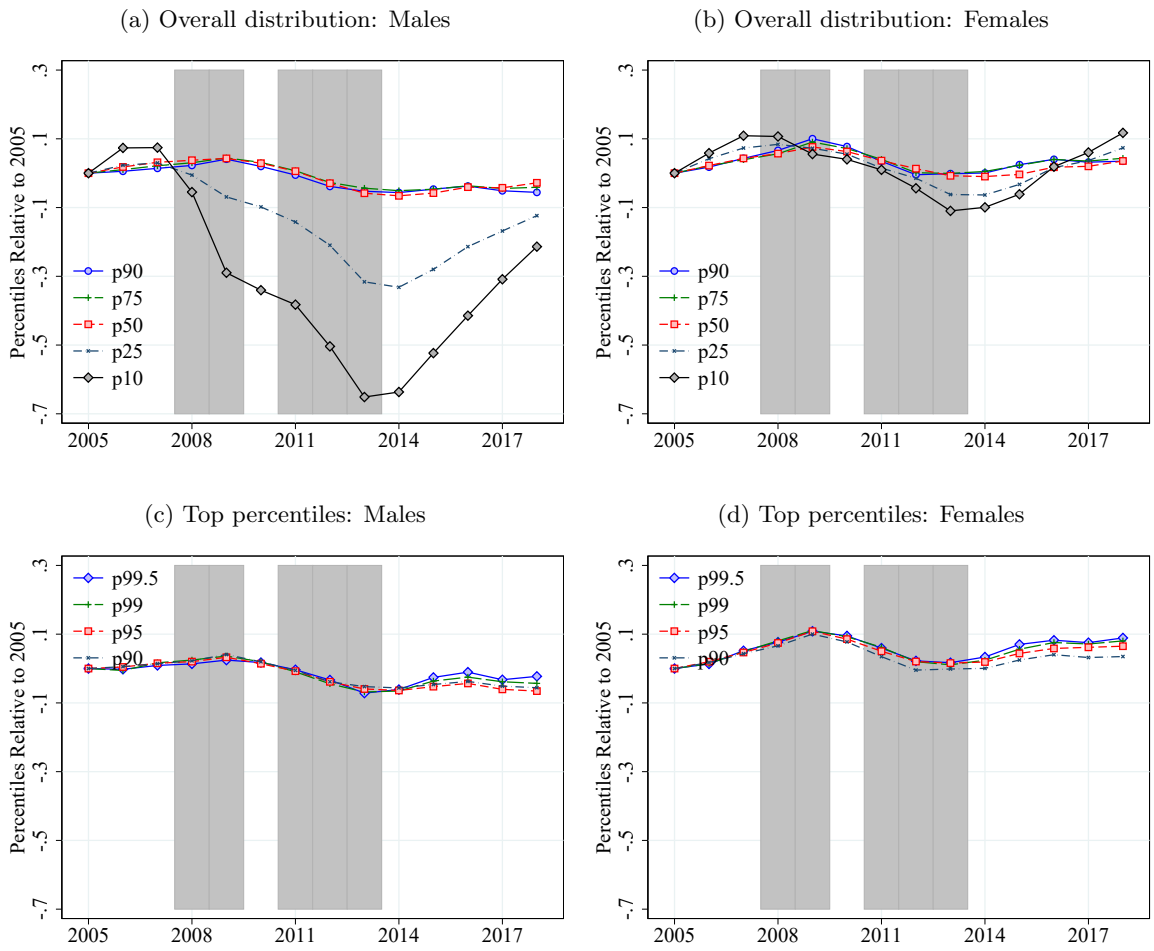
In the bottom two graphs of Figure 2 we show various percentiles at the top of the distribution of log annual earnings, up to the 99.5th percentile. For both genders, top percentiles tend to decrease between 2009 and 2013. However, this decrease is quantitatively small. In addition, the graphs show that all percentiles above the 90th tend to evolve similarly over the period. This suggests that, in Spain, the recession did not affect top labor incomes (i.e., 99th percentile and above) differently from the rest of the top decile. Note that, due to relatively small sample sizes, we are not able to reliably document the evolution of earnings percentiles above the 99.5th in the MCVL. Note also that, given our data, we only include labor earnings, and do not account for capital income in the analysis.

The stability over time of the upper part of the Spanish income distribution, including the right tail, stands in contrast with the experience of other countries, such as the US and the UK (Piketty and Saez, 2013).[5] For Spain, this evidence is consistent with results from survey data in recent years (Anghel et al., 2018). Using top coded administrative records and extrapolation, Bonhomme and Hospido (2017) found that the P90-P50 percentile difference increased substantially between 1988 and 1996, explaining most of the increase in inequality

[4] In Appendix Figure F1 we show the original percentiles, without normalizing them to zero in 2005.

[5] In Appendix Figures F2 and F3 we report Pareto tail coefficients, by gender, estimated on 1% and 5% of the sample, respectively. We find that the tail coefficients are approximately similar in 2005 and 2015, for both genders.

Figure 2: Percentiles of the distribution of log annual earnings



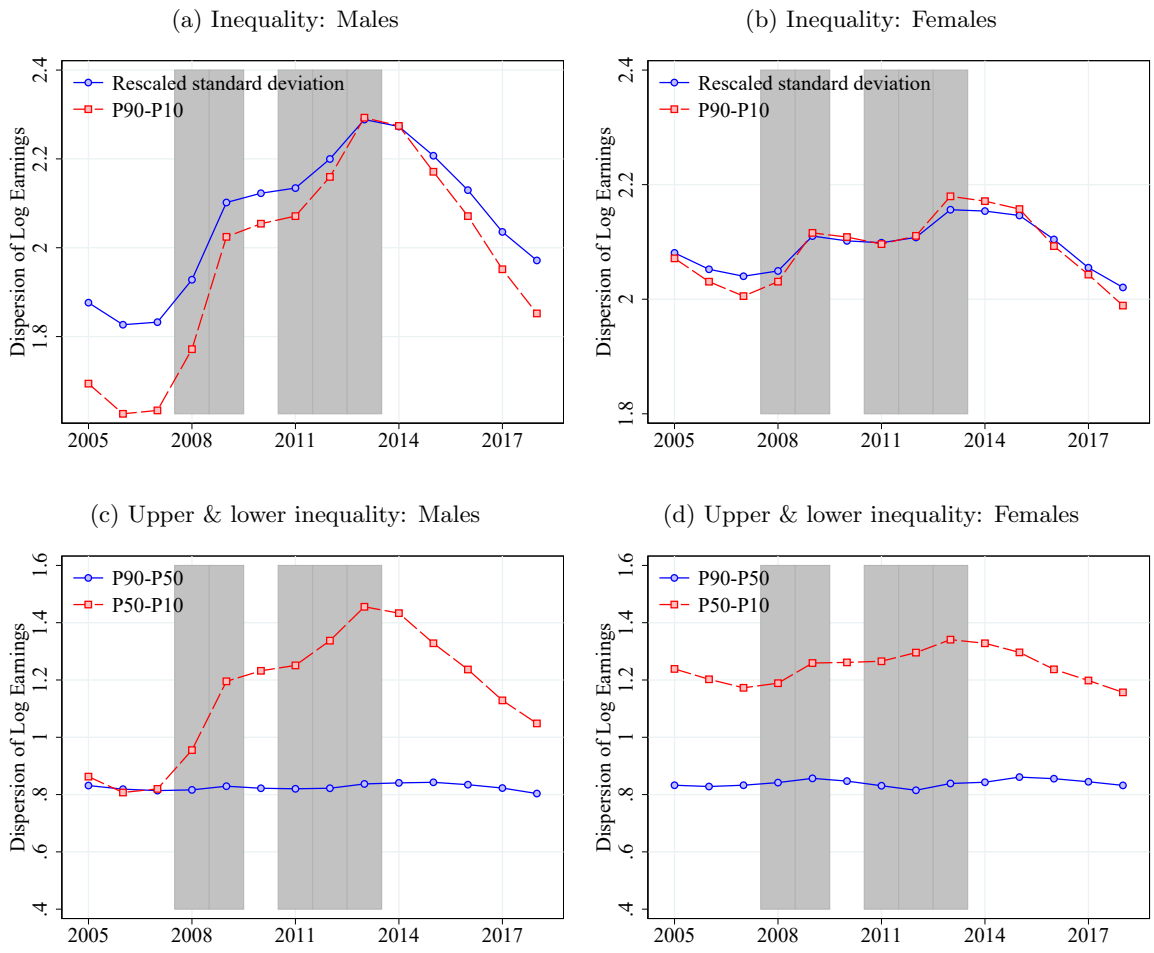
Notes: CS sample, percentiles of log annual earnings, by gender. All percentiles are normalized to 0 in 2005. The shaded areas indicate recession years.

during that period. Despite data differences, this suggests that the recent stability in the upper part of the distribution might not be a long-run phenomenon.

In Figure 3 we show various measures of inequality, by gender and over time.[6] In the top graphs, we focus on overall inequality, as measured by the P90-P10 percentile difference in log annual earnings, as well as by the standard deviation of log annual earnings — suitably scaled in order to facilitate comparability with the P90-P10 measure. The two measures

[6] In Appendix Figure F4 we show the evolution in the overall population, pooling both genders. In Appendix Figures F5 and F6 we show the results controlling for age, and for age and education, respectively.

Figure 3: Income inequality



Notes: CS sample, log annual earnings. In the top graphs, the P90-P10 difference is indicated in squares, and the rescaled standard deviation is indicated in circles (using a scaling factor of 2.56, in order to facilitate comparison between the two measures). In the bottom graphs, the P90-P50 difference is indicated in squares, and the P50-P10 difference is indicated in circles. The shaded areas indicate recession years.

of inequality give a consistent message. For males, inequality increases substantially with the recession, and decreases afterwards. The magnitudes of the fluctuations are substantial. Indeed, the P90-P10 measure increases by 0.7 between 2007 and 2013. For females, the inequality increase associated with the recession is more moderate, with an increase of less than 0.2.

In the bottom graphs of Figure 3 we focus on upper and lower inequality, as measured

by the percentile differences P90-P50 and P50-P10, respectively. For males, inequality in the bottom part of the earnings distribution increases sharply around the recession: indeed, the P50-P10 measure increases by 0.7 between 2007 and 2013. In contrast, upper inequality as measured by the P90-P50 difference is approximately flat over the entire period. This is consistent with the findings of Bonhomme and Hospido (2017), who emphasize the role of sectors, and in particular construction, in the evolution of male inequality in Spain. For females, the P50-P10 also increases in the recession, albeit much less so than for males, and upper inequality is also approximately constant over the period.[7]

When interpreting these features of the Spanish earnings distribution, it is important to take into account the large fluctuations in unemployment over the period. In the second part of the paper we will consider a broader sample, including unemployed individuals with zero labor earnings in a year. As an additional exercise, we have computed measures of inequality based on an income measure that combines labor earnings and unemployment benefits, while keeping the same sample as in the rest of this section. The results show little difference relative to only using labor earnings.[8]

3.2 Income changes

We next turn to the distribution of earnings changes and its evolution. For this purpose, we first focus on the LS sample, and construct residualized log earnings $\varepsilon_{it} = \log y_{it} - x_{it}'\widehat{\beta}$, where x_{it} includes fully-saturated interactions of age dummies, gender and year indicators, and $\widehat{\beta}$ is a regression coefficient, as well as their one-year changes $g_{it} = \Delta\varepsilon_{it} = \varepsilon_{it+1} - \varepsilon_{it}$. We will also refer to multiple-year changes such as $g_{it}^5 = \Delta^5\varepsilon_{it} = \varepsilon_{it+5} - \varepsilon_{it}$.

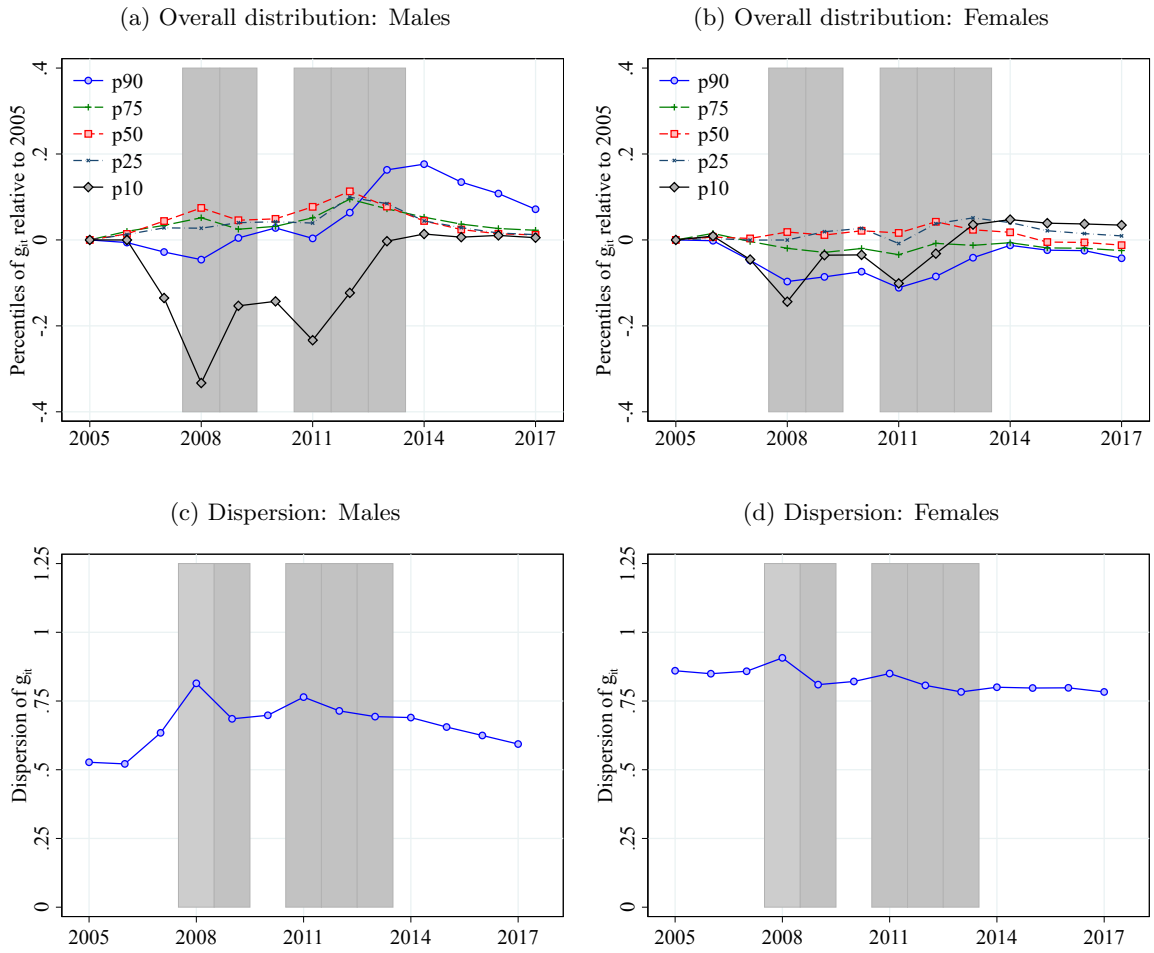
In Figure 4 we start by documenting the evolution over time of percentiles of one-year log earnings changes.[9] All percentiles are relative to the reference year 2005. The top left graph, for males, shows a sharp contrast between the 10th percentile and the other percentiles. Indeed, while most percentiles of log earnings changes increase somewhat over the period,

[7] In Appendix Figure F7 we report the income shares of various percentiles. We find that the share of the bottom 50% decreases substantially around the recession (by 25%), whereas the top 1% remains approximately stable.

[8] See Appendix Figure F8. Another notable aspect of the Spanish economy in this period is the increase in the percentage of immigrants. In Appendix Figure F9 we report earnings percentiles and inequality in a sample without immigrants, and find similar results to the ones based on the sample with immigrants.

[9] In Appendix Figures F10 and F11 we show the densities of one-year and five-year log annual earnings changes, respectively. In Appendix Figures F12 and F13 we show the corresponding log densities.

Figure 4: One-year changes in log earnings, percentiles and dispersion



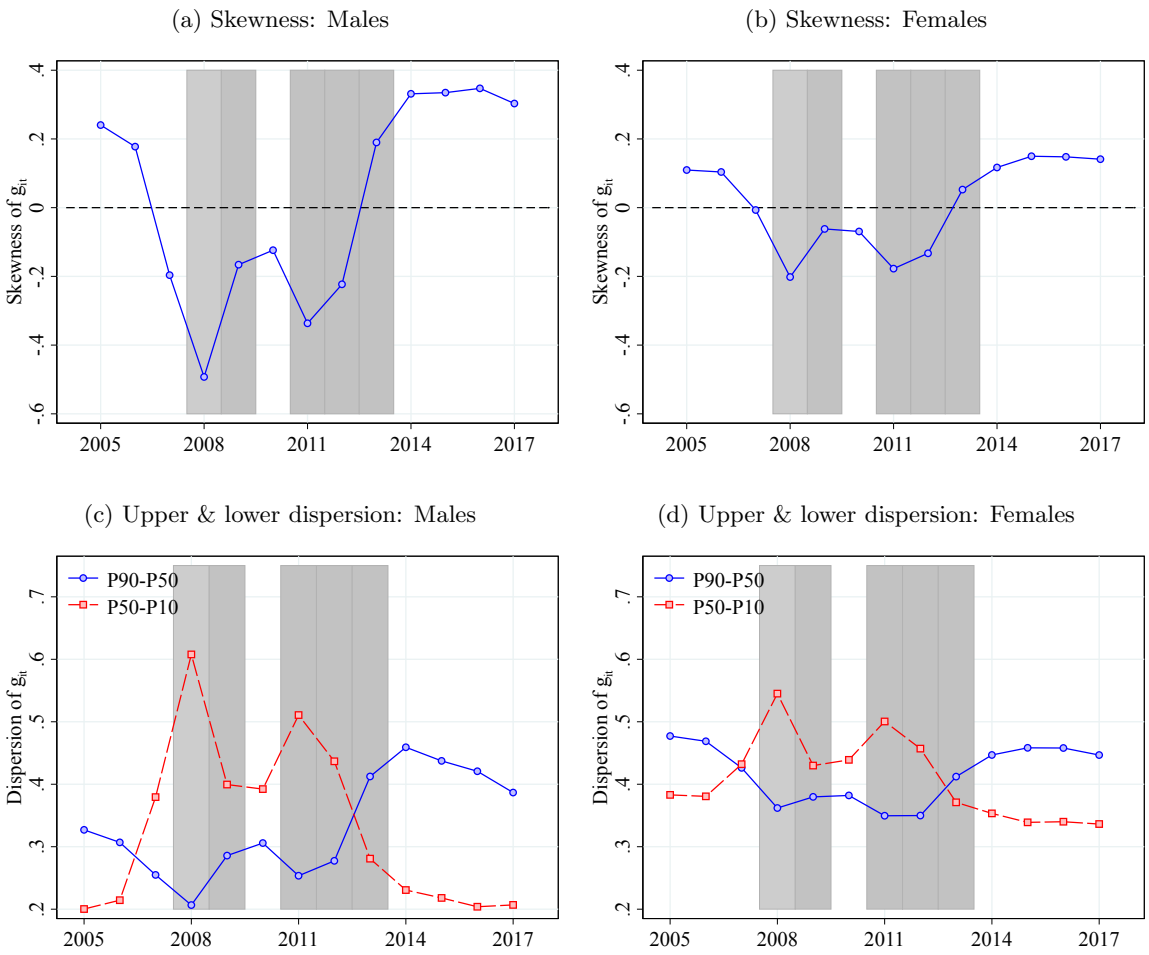
Notes: LS sample, one-year changes in residualized log earnings. In the upper panel, all percentiles are normalized to 0 in 2005. In the lower panel, dispersion measured by $P90 - P10$. The shaded areas indicate recession years.

the 10th percentile decreases sharply around the recession. Moreover, as the comparison to the right graph shows, this evolution is not as pronounced for females.[10] In the lower panel of Figure 4 we show the P90-P10 percentile difference of log earnings changes, alongside the standard deviation, rescaled for comparability. We find that the dispersion of log earnings changes increases at the beginning of the recession, especially for males.[11]

[10] In Appendix Figure F14 we focus on percentiles of log earnings changes above the 90th percentile. We see that the top percentiles tend to move approximately in parallel for both genders.

[11] In Appendix Figure F15 we document the dispersion of five-year log earnings changes.

Figure 5: Skewness and upper & lower dispersion of one-year log earnings changes



Notes: LS sample, one-year changes in residualized log earnings. Kelley skewness is $\frac{P90-2P50+P10}{P90-P10}$. *The shaded areas indicate recession years.*

While in Figure 4 we focus on one-year changes, it is also informative to document changes over long periods. To do so, we compute cumulative earnings changes around the recession, between 2006 and 2014, net of age effects. We find that the distribution of earnings changes over the long period is widely dispersed. While, for males, the 90th percentile of 2006-2014 log-earnings changes is +62%, the 10th percentile is -93%. For females, the corresponding 90th and 10th percentiles are +77% and -77%, respectively.[12]

[12]In Appendix Figure F16 we plot the cumulative earnings changes between 2006 and 2014, against initial earnings percentiles in 2006. The figure shows that the dispersion of log earnings changes in the long period tends to decrease with the level of initial earnings.

Recent work has documented the cyclical behavior of the skewness of log earnings changes in the US (Guvenen et al., 2014) and in other countries (e.g., Hoffmann and Malacrino, 2019, Pora and Wilner, 2020, Busch et al., forthcoming). In the top panel of Figure 5 we show the evolution over time of the Kelley measure of skewness of one-year log earnings changes. We see that skewness becomes more negative in the recession, in agreement with the findings of Guvenen et al. (2014) for the US and Busch et al. (forthcoming) for Germany, Sweden and France. This evolution is more pronounced for males than for females. The changes in skewness that we document for males are substantial by international standards.[13]

In the bottom panel of Figure 5 we show the P90-P50 and P50-P10 percentile differences, which measure the upper and lower dispersion of the changes in one-year log earnings, respectively. The dispersion of log earnings changes in the lower part of the distribution increases during the recession, more so for males. The dispersion of log earnings changes in the upper part of the distribution also increases, albeit the increase happens at the end of the recession in this case, and it is most pronounced for males.

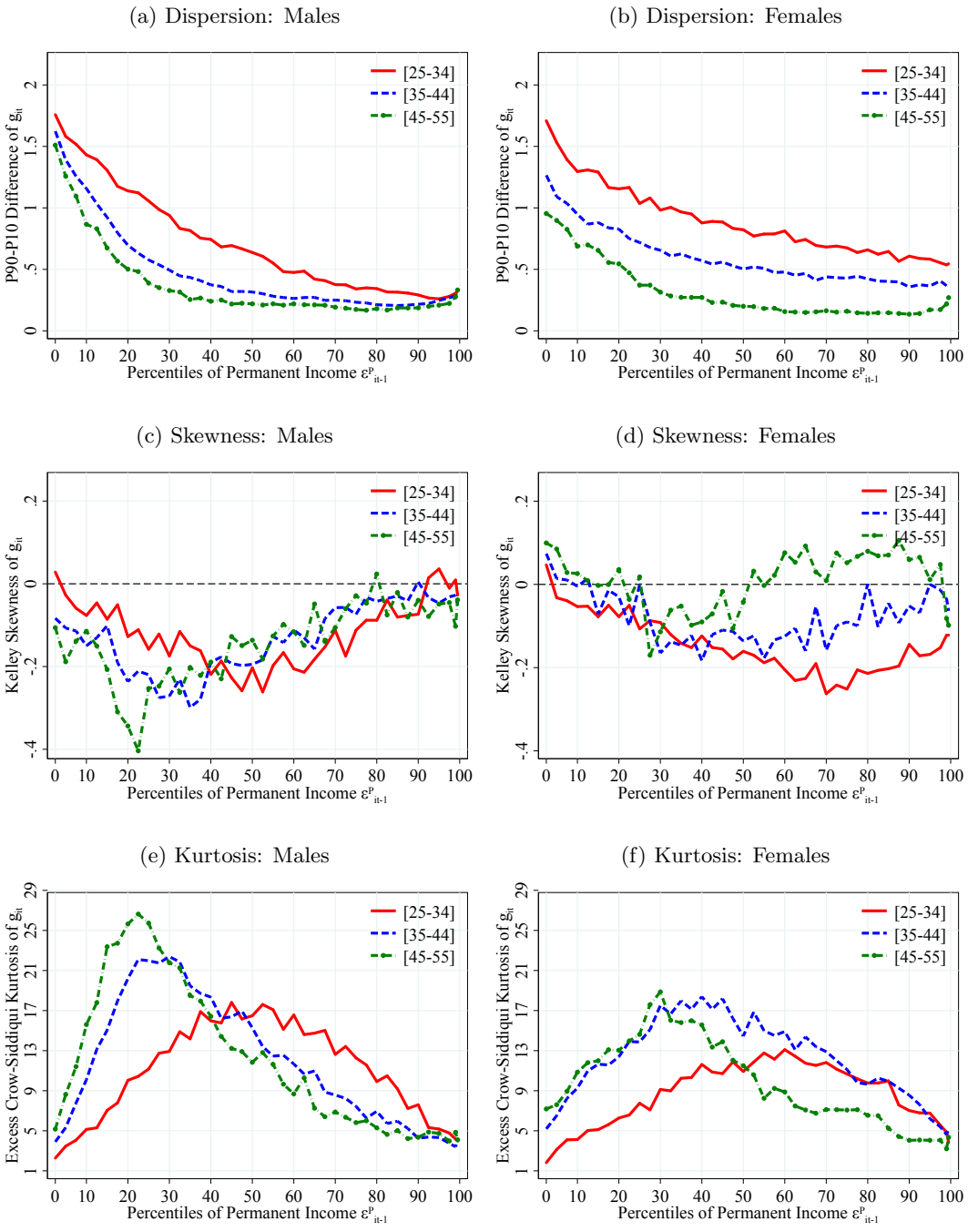
We are interested in relating the dispersion and skewness of log earnings changes to the position of the individual in the earnings distribution. Arellano et al. (2017) and Guvenen et al. (forthcoming) find, using US data, that the dispersion and skewness of income depend on past income. Such measures of conditional dispersion and skewness are particularly relevant to us, given our goal of documenting income risk. Following Guvenen et al. (forthcoming), we construct a measure of "permanent" earnings as $P_{it} = (y_{it-2} + y_{it-1} + y_{it}) / \left(\sum_{\tau=0}^{2} \mathbf{1}\{y_{it-\tau} \geq \underline{y}_{t-\tau}\}\right)$, computed only for those whose earnings are above the threshold $\underline{y}_t$ in at least two of the past three years. We also construct residualized log permanent earnings ε_{it}^{P}.

In Figure 6 we show several measures of dispersion, skewness and kurtosis of one-year log earnings changes, by gender, conditional on lagged residualized log permanent earnings.[14] In the top graphs, we find that the dispersion of log earnings decreases with the level of permanent earnings. Dispersion only increases slightly, for males, at the top levels of perma-

[13] In Appendix Figure F17 we report results based on a moment-based measure of skewness, and find similar results to the ones obtained using the Kelley measure. We also report results for kurtosis, however those are less consistent since quantile-based and moment-based measures disagree to a large extent in this case.

[14] In Appendix Figure F19 we report the corresponding moment-based measures of dispersion, skewness, and kurtosis.

Figure 6: Conditional dispersion, skewness and kurtosis of one-year log earnings changes



Notes: H sample, one-year changes in residualized log earnings. On the x-axis we report percentiles of residualized log permanent earnings ε^P_{it-1}. In the top panel we show the P90-P10 percentile difference, in the middle panel we show Kelley skewness, and in the bottom panel we show excess Crow-Siddiqui kurtosis. The various curves on the graphs corresponds to various age groups: between 25 and 34 years, between 35 and 44, and between 45 and 55 years, respectively.

nent incomes reported on the graph, which correspond to the 99.5 percentile. While sample sizes prevent us from drawing firm conclusions above this level, we checked that dispersion increases somewhat more steeply for the top 0.5%. Moreover, conditional dispersion tends to decrease over the life cycle, for both males and females. The conditional standard deviation of log income given past income may be interpreted as a measure of income risk. In the second part of the paper, we will compare this measure with a prediction-based approach for a broader income measure.

Lastly, in the middle and bottom panels of Figure 6 we show the skewness and kurtosis of one-year log earnings changes, by gender, conditional on permanent earnings P_{it}. The quantile-based measures of higher-order features of the distribution of log earnings suggest that, for both genders, skewness is more negative and excess kurtosis is higher in the middle of the earnings distribution.

3.3 Age profiles and income persistence

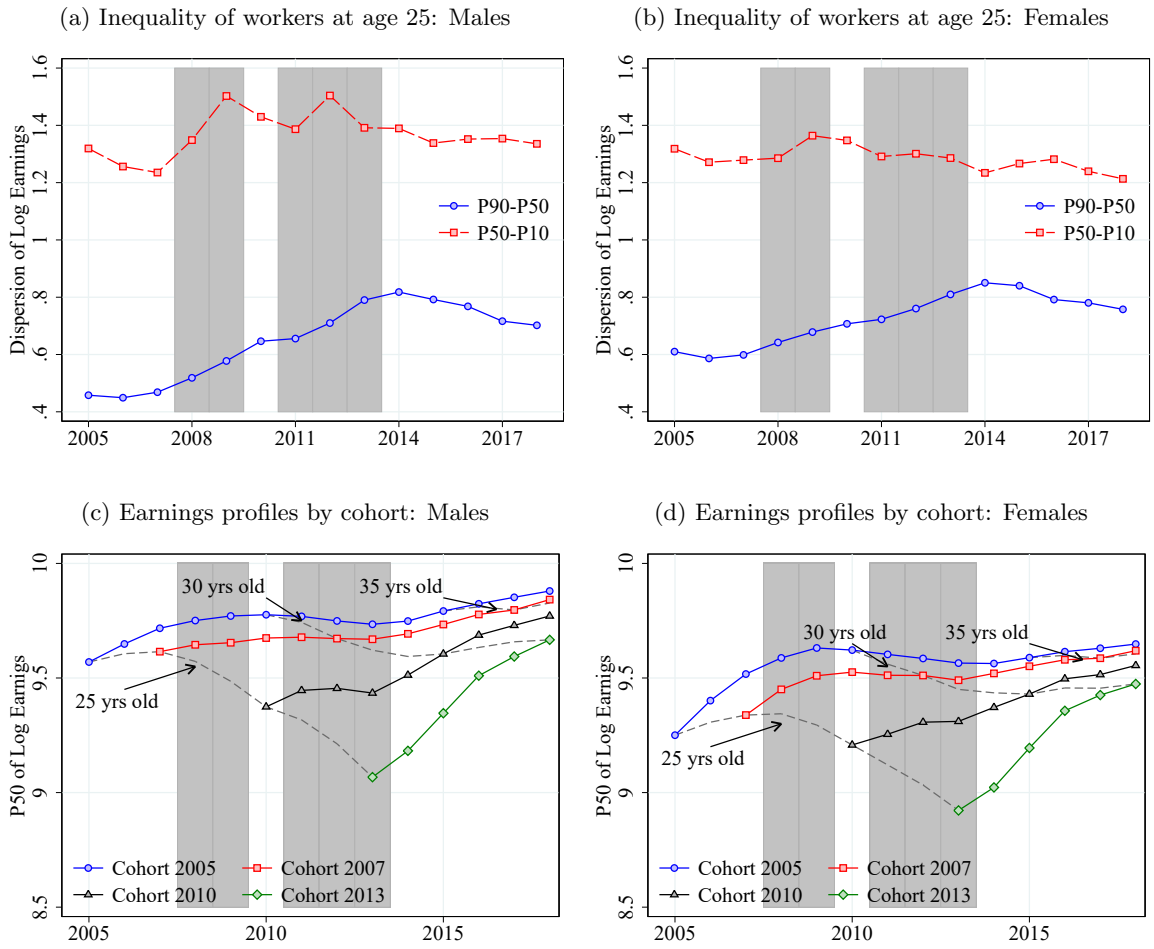
In this last part on income inequality and dynamics, we focus on inequality by cohort and age groups, and on earnings persistence and mobility. In the upper panel of Figure 7 we report the P90-P10, P90-P50, and P50-P10 percentile differences at age 25, by gender, from 2005 to 2018.[15] The results show that, for both genders, inequality in the upper part of the distribution increases during the recession. This pattern for younger workers, which contrasts with the evolution of upper inequality in the whole sample that we documented in Figure 3, reflects in part a fall in median log earnings for young workers during the recession.

Next, in the lower panel of Figure 7 we compare earnings profiles for different cohorts over time. For both males and females, the cohorts of workers who started during the recession have a substantially lower initial level, compared to the cohorts who started in 2005, however their subsequent earnings profile is steeper.[16]

Finally, for the purpose of understanding income dynamics it is also interesting to document to which extent current earnings are associated with future earnings. Pijoan-Mas and Sánchez-Marcos (2010) and Alvarez and Arellano (2021) estimate earnings processes using survey data. Here we report simple measures of earnings mobility based on our administrative

[15] In Appendix Figure F25 we show percentiles of log annual earnings at age 25.
[16] In Appendix Figure F26 we show earnings inequality.

Figure 7: Inequality and age profiles for young workers

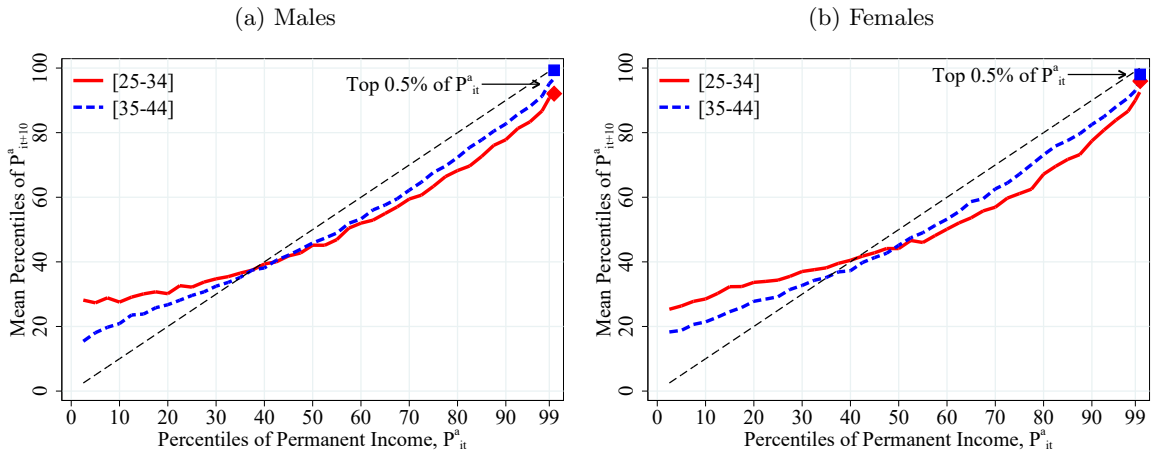


Notes: CS sample, log annual earnings. In the top panel the sample is restricted to age-25 workers only. In the bottom panel the different curves correspond to different cohorts of workers. The shaded areas indicate recession years.

sample. In Figure 8 we report 10-year average rank-rank mobility for two age groups: 25-34 and 35-44. The figure shows reversion towards the mean, and relatively small changes with age. There is upward mobility for those at the bottom of the permanent income distribution, and downward mobility for those at the top of the distribution. These patterns are similar for males and females, and more pronounced for the young.[17]

[17] In Appendix Figures F22, F23, and F24, we show mobility over the life cycle, over time, and at a 10-year horizon, respectively.

Figure 8: Evolution of 10-year earnings mobility over the life cycle



Notes: Sample of individuals for which the alternative permanent income measures, P^a_{it} and P^a_{it+10}, exist, where $P^a_{it} = (y_{it} + y_{it-1} + y_{it-2})/3$, for individuals with non-missing earnings $y_{it}, y_{it-1}, y_{it-2}$ for whom at least one of them is above the threshold. This figure shows average rank-rank 10-year earnings mobility. The various curves on the graph correspond to different age groups measured at time t: solid corresponds to 25-34 and dashed corresponds to 35-44. The squares and diamonds correspond to the top 0.5 percentile of the distribution of permanent income at t.

4 Measuring income risk

In the second part of the paper we now study income risk and income risk inequality in Spain. We first describe how we measure inequality in income risk. In the empirical analysis, we extend the notion of earnings to include observations below the income threshold $\underline{y}_t$ that we used in the first part of the paper, including zeros, as well as unemployment benefits. Including both sources of income provides a closer approximation to the income risk faced by individuals when making consumption and investment plans. We envisage an individual that factors in employment transitions within the year and takes both sources of income into account when forming expectations of her income over the next year. In this section and the next we restrict the analysis to males.

4.1 A CV measure of income risk

Our goal is to produce summary measures of the uncertainty of an individual agent's one-year-ahead predictive income distribution. We propose to mimic the agent's prediction problem as

closely as we can, using the administrative records at our disposal. We target the distribution of income levels Y_{it} given predictors X_{it}, which we conceive are predictors also considered by the agent.

We use both micro and macro predictors in X_{it}. The *micro* predictors include a cubic polynomial in past log labor income, $\log Y_{it-1}$, interacted with an indicator that Y_{it-1} is positive; the log of income from unemployment benefits at $t-1$; an indicator that income from unemployment benefits is positive; the number of days worked in $t-1$; dummy variables that indicate working full year at $t-1$, $t-2$, and $t-3$; an indicator for full-time employment status in the main job (defined as the job spell that contributes the largest fraction to total annual earnings); an indicator for permanent contract of the main job; and indicators of educational attainment. We have also tried a larger set of predictors including firm- and family-related variables (firm size, industry, and family size), finding qualitatively and quantitatively similar results to the ones we report below. In some specifications, we will augment this list to include unobserved heterogeneity (discussed in Appendix B.3). We interact all micro predictors with a quadratic in age.

In turn, the *macro* predictors include GDP growth and unemployment rate at $t-1$, $t-2$, and $t-3$, at the national and provincial level, as well as their interactions with age. We use aggregate covariates such as GDP growth and unemployment in an attempt to mimic the agent's information set in the presence of aggregate uncertainty. Alternatively, one could assume perfect foresight about next year's macroeconomic conditions, and estimate risk models with time-varying parameters. Given that aggregate conditions end up playing a small quantitative role in our results, following such an approach does not materially affect any of the conclusions below.

We propose to measure income risk using the following coefficient of variation (CV hereafter):

$$\mathrm{CV}(\mathrm{X_{it}}) = \frac{\overbrace{\mathbb{E}\left(|Y_{it} - \mathbb{E}(Y_{it} \mid X_{it})| \mid X_{it}\right)}^{\text{mean absolute deviation}}}{\underbrace{\mathbb{E}(Y_{it} \mid X_{it})}_{\text{mean}}}. \tag{1}$$

The CV is a ratio between two measures: the mean absolute deviation, which is a measure of dispersion of the predictive distribution of income, and the mean, which is a measure of

location. In words, an individual with an expected income of 20,000 euros and a CV of 0.1 expects a deviation of her next year's income from its mean of ±2000 euros.

We use the mean absolute deviation instead of the standard deviation in the numerator to minimize sensitivity to extreme observations. A rescaled version of $\mathrm{CV}(\mathrm{X_{it}})$ is directly comparable to the usual coefficient of variation that has the standard deviation in the numerator, the scaling factor being $\sqrt{\frac{\pi}{2}} \approx 1.25$. When the CV is small, it is approximately equal to the rescaled standard deviation of log income, conditional on the predictors; that is, $\mathrm{CV}(\mathrm{X_{it}}) \approx \sqrt{\frac{2}{\pi}}\,\mathrm{Std}(\log(\mathrm{Y_{it}})|\mathrm{X_{it}})$. However, unlike the standard deviation of log income, the CV remains well-defined when $Y_{it} = 0$. We will also report results based on other robust counterparts to CV, using the conditional median instead of the mean.

Discussion. To assess the magnitude of the risk measures that we report, we find it informative to provide a simple welfare interpretation in the spirit of Lucas (1987). To do so, we approximate the welfare gain to an individual associated with fully eliminating the income risk that she faces. To proceed, consider an individual with utility $U_i(C_{it}) = \frac{C_{it}^{1-\theta_i}-1}{1-\theta_i}$, with consumption $C_{it} = \lambda(X_{it})Y_{it}$ for some proportionality factor $\lambda(X_{it})$. Suppose also that Y_{it} given X_{it} is log-normally distributed. The welfare gain of eliminating income risk faced by i at t can then be approximated in percentage of consumption as

$$\text{Welfare gain} \approx \frac{1}{2} \times \theta_i \times \mathrm{Var}(\log(\mathrm{Y_{it}})|\mathrm{X_{it}}).$$

That is, alternatively,

$$\text{Welfare gain} \approx \frac{\pi}{4} \times \theta_i \times \mathrm{CV}(\mathrm{X_{it}})^2, \tag{2}$$

where $\mathrm{CV}(\mathrm{X_{it}})$ is given by (1). Based on this calculation, we would interpret a CV value lower than 0.1 as reflecting relatively low individual income risk (e.g., corresponding to less than 2% of consumption when $\theta_i = 2$), whereas values of 0.3 or higher correspond to substantial amounts of risk that can potentially impact individual welfare in major ways (e.g., corresponding to more than 14% of consumption when $\theta_i = 2$).

An important limitation of this derivation is that it relies on income being conditionally log-normal. As we documented in the first part of the paper, log-normality may not be a good approximation in our setting. In this case, conditional higher-order moments of income

such as skewness and kurtosis will also matter in order to assess the welfare gains associated with eliminating income risk. As a result, the CV will not necessarily accurately measure the income risk faced by individuals, possibly underestimating it. Extending our approach to estimate the full conditional distribution of income, as we mention in Subsection 5.3 and detail in Appendix B, it is in principle possible to compute the welfare gains of eliminating risk given individual preferences. Although we do not pursue this possibility here, we will also report quantile-based risk measures as a complement to the CV.

Another limitation of the above welfare calculation is that it relies on a specific, possibly restrictive form for individual preferences. To illustrate, suppose the individual's utility function takes a Stone-Geary form, $U_i(C_{it}) = \frac{(C_{it}-C_m)^{1-\theta_i}-1}{1-\theta_i}$, where C_m is a subsistence consumption level. In Appendix A we show that, if $C_{it} - C_m$ is log-normal, and using the same approximation as in (2), the welfare gain of eliminating income risk can be approximated as

$$\text{Welfare gain} \approx \frac{\pi}{4} \times \theta_i \times \frac{\mathbb{E}(C_{it} \,|\, X_{it})}{\mathbb{E}(C_{it} \,|\, X_{it}) - C_m} \times \text{CV}(\text{X}_{\text{it}})^2. \tag{3}$$

Hence, for non-negligible values of $C_m/\mathbb{E}(C_{it} \,|\, X_{it})$ — e.g., for individuals whose average consumption is close to the subsistence level — the squared CV underestimates the welfare cost of income risk. Moreover, given our empirical finding that risk and income are negatively correlated, a CV-based measure will then tend to underestimate the degree of income risk inequality.

Lastly, it is important to note that, since the CV is based on a predictive income distribution, its interpretation hinges on the chosen predictors. While we attempt to mimic the agent's information set using the administrative data, it is of course possible that the agent's information does not coincide with the set of predictors that we rely on. This fundamental challenge in risk measurement will motivate us to consider specifications with different sets of observed and unobserved predictors. In addition, we will use Spanish civil servants as a convenient test sample that we expect to face very low income risk, and we will compare our prediction-based risk measures with estimates based on subjective expectations.

4.2 Income risk: econometric approach

Estimating the numerator and denominator of the coefficient of variation in (1) requires performing two prediction tasks. Here we describe a simple and parsimonious approach to predict income and quantify income risk. In Section 5.3 we will describe several extensions of this approach, and report results based on them.

Since income is non-negative, a parametric estimator can be based on the two following exponential specifications:

$$\mathbb{E}(Y_{it}|X_{it}) = \exp(X_{it}'\beta),$$

and

$$\mathbb{E}\left(|Y_{it} - \mathbb{E}(Y_{it} \mid X_{it})| \mid X_{it}\right) = \exp(X_{it}'\gamma),$$

where X_{it} includes all the micro and macro predictors that we listed in the previous subsection. We estimate β and γ using two Poisson regressions.[18] First, we regress Y_{it} on X_{it}, which gives us $\widehat{\beta}$. To alleviate issues related to outliers in the prediction of the conditional mean, we censor the upper tail of predicted values at the maximum value of total income in the data, which only affects a handful of observations. Then, we regress $|Y_{it} - \exp(X_{it}'\widehat{\beta})|$ on X_{it}, which gives us $\widehat{\gamma}$. Finally, given estimates $\widehat{\beta}$ and $\widehat{\gamma}$, we compute our estimate of the risk faced by individual i in year t as

$$\widehat{\mathrm{CV}}_{it} = \exp\left(X_{it}'(\widehat{\gamma} - \widehat{\beta})\right). \tag{4}$$

In the next section, we will document several key features of the distribution of income risk and income risk inequality, based on our risk measure $\widehat{\mathrm{CV}}_{it}$. Before doing so, we perform two exercises in order to better understand what $\widehat{\mathrm{CV}}_{it}$ measures. The first exercise is a quantification of prediction performance associated with the two tasks of predicting income absolute deviations (to estimate the numerator of CV), and predicting income levels (to estimate the denominator of CV). We document in-sample performance using data for the years 2006-2017. In addition, we document out-of-sample performance using data for the years 2006-2017 as our estimation sample and data for 2018 as our hold-out sample. Note that this exercise measures prediction performance for a given set of predictors, so accurate prediction need not imply that we correctly capture the income risk that agents face.

[18] We found Poisson estimates to be more numerically stable than estimates from exponential regressions.

Table 1: Prediction performance

(a) Mean (CV denominator)

	In sample				Out of sample			
	Income	+Days	+Days+Age	All	Income	+Days	+Days+Age	All
MSE	23027126	22857682	22708526	21391777	19464323	19626672	19392590	20010842
MAE	3007	3003	3002	3058	2770	2790	2771	2938
Log-lik	221363	221369	221383	221483	224057	224044	224056	224189

(b) Absolute deviation (CV numerator)

	In sample				Out of sample			
	Income	+Days	+Days+Age	All	Income	+Days	+Days+Age	All
MSE	12956397	12313304	12014627	10219270	12723757	11272098	11028598	9714735
MAE	2734	2484	2427	2200	2768	2410	2337	2211
Log-lik	25953	26380	26447	26777	25121	25646	25709	26043

Notes: B sample. MSE is mean squared error, with 99th percentile trimmed. MAE is mean absolute error, with 99th percentile trimmed. Log-lik is the log likelihood value divided by the number of observations. Exponential regression models, using lagged log income and an indicator of past income being zero ("Income"), adding days worked in the year ("+Days"), adding days worked and age ("+Days+Age"), and using all micro and macro predictors ("All"). In sample is for 2006-2017. Out of sample is for 2018. The bottom panel corresponds to performance in the prediction of the absolute deviation, using the "All" specification as the estimate for the mean to maintain comparability between columns.

We compare four specifications: (1) only using as predictors income lagged one year and the indicator that income is positive, (2) adding the number of days worked, (3) adding age to income and days worked, and (4) including all the micro and macro predictors that we listed in the previous subsection. In Table 1 we report the mean squared error (MSE) and the mean absolute error (MAE), both trimmed at the 99th percentile in order to reduce sensitivity to extreme observations, as well as the average log likelihood value (Log-lik) of the Poisson model. In the top panel we focus on the prediction of income levels, and in the bottom panel we show the results for the prediction of income absolute deviations — in which case we assess the prediction for $|Y_{it} - \exp(X_{it}'\widehat{\beta})|$, where $\widehat{\beta}$ is estimated based on the 2005-2017 sample using the most comprehensive specification. We see that, while the various specifications perform similarly in sample and out of sample to predict income levels (top panel), adding other predictors beyond lagged income tends to improve the prediction of income absolute deviations (bottom panel).

In the second exercise, we attempt to document the main sources of variation in the CV

Table 2: Explaining the variation in CV

	Age categories		
	26-30	36-40	46-50
Business cycle	0.0205	0.0077	0.0076
Permanent (t-1)	0.0019	0.0008	0.0001
Full time (t-1)	0.0168	0.0117	0.0082
Days worked (t-1)	0.6937	0.4512	0.3519
Income (t-1)	0.0014	0.0013	0.0000

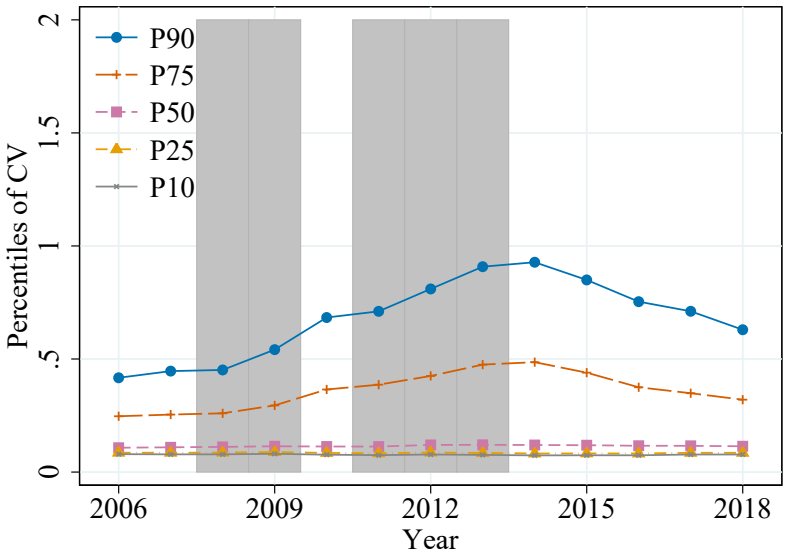
Notes: B sample. Partial R^2 in linear regressions of $\widehat{\text{CV}}_{it}$ on various determinants. Exponential specification that includes all macro and micro predictors. "Business cycle" includes the macro predictors, i.e., GDP growth and unemployment rate at $t-1, t-2$ and $t-3$ at the national and provincial level.

risk measure, using regressions. Specifically, we regress $\widehat{\text{CV}}_{it}$ in (4) on five sets of covariates, and we report the partial R^2 coefficients associated with each of them. In Table 2 we show the results, split by age categories. The sets of covariates are: an indicator of permanent labor contract, an indicator of full-time labor contract, the number of days worked, and the income level (all of them lagged), and our macro indicators. We see that the number of days worked in the past year explains the largest part of the variation in the CV. Net of the impact of days worked, the macro indicators, the features of the labor contract, and the income level, all have low explanatory power. The Spanish economy experiences high levels of unemployment and employment turnover related to the large share of short-term temporary employment. The partial R^2 coefficients in Table 2 suggest that these features contribute substantially to the empirical variation in income risk.

5 Income risk inequality in Spain

A large part of the literature that studies cross-sectional inequality concentrates on the inequality in the levels of income. In this section, we document the magnitude and evolution of inequality in income risk in Spain, where we measure individual risk using our proposed CV.

Figure 9: Income risk over the period, percentiles of CV



Notes: B sample. Exponential specification, using all macro and micro predictors. The shaded areas indicate recession years.

5.1 Income risk inequality over the period

In Figure 9 we show the evolution of different percentiles of CV over time. In Table 3 we report selected quantiles of the income risk distribution over time, as well as various measures of income risk inequality.[19] We see that both the level and evolution of income risk vary very differently along the distribution. The lower part of the income risk distribution corresponds to CV values of at most 0.12. This suggests that at least half of the Spanish economy faces little uncertainty in their future income. In addition, for this part of the sample, risk levels stay remarkably constant over the period. In contrast, the 75th and 90th percentiles of income risk have large CV values, and those vary widely over the period: the 75th (respectively, the 90th) percentile ranges between 0.3 (resp., 0.7) at the beginning of the period and 0.5 (resp., 1.2) at the end of the recession

As a result, inequality in income risk tends to increase in the recession. As shown by the left panel of Figure 10, while median risk remains constant during the entire period, income

[19] In Appendix Figure G1 we report the coefficient estimates that we use to construct our CV.

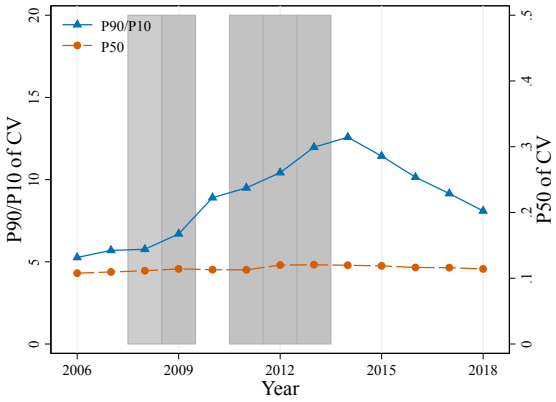
Table 3: Income risk over the period, in numbers

	All	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
P90/P10	8.94	5.26	5.69	5.76	6.70	8.90	9.50	10.42	11.97	12.58	11.43	10.15	9.15	8.09
P90/P50	6.03	3.87	4.07	4.06	4.74	6.05	6.30	6.75	7.53	7.75	7.15	6.48	6.13	5.51
P50/P10	1.48	1.36	1.40	1.42	1.41	1.47	1.51	1.55	1.59	1.62	1.60	1.57	1.49	1.47
p10	0.08	0.08	0.08	0.08	0.08	0.08	0.07	0.08	0.08	0.07	0.07	0.07	0.08	0.08
p25	0.08	0.09	0.09	0.09	0.09	0.08	0.08	0.09	0.08	0.08	0.08	0.08	0.09	0.09
p50	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.12	0.12	0.12	0.12	0.12	0.12	0.11
p75	0.34	0.25	0.25	0.26	0.29	0.37	0.39	0.43	0.48	0.49	0.44	0.38	0.35	0.32
p90	0.69	0.42	0.45	0.45	0.54	0.68	0.71	0.81	0.91	0.93	0.85	0.75	0.71	0.63

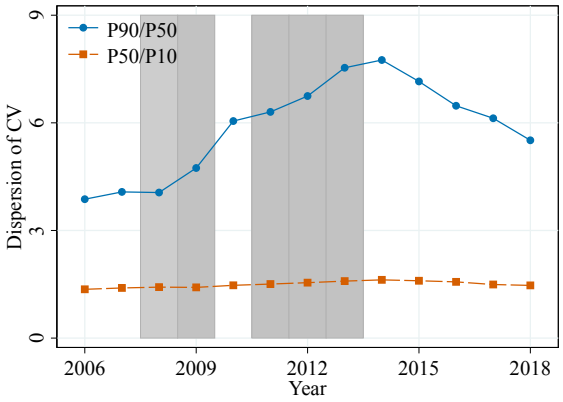
Notes: B sample. Exponential specification, using all macro and micro predictors.

Figure 10: Income risk inequality over the period

(a) Median risk and inequality



(b) Upper & lower inequality



Notes: B sample. Exponential specification, using all macro and micro predictors. The shaded areas indicate recession years.

risk inequality — as measured by the P90/P10 ratio — increases substantially during the recession, with a more than threefold increase between 2006 and 2013. This evolution is qualitatively in line with the one of earnings inequality, see Figure 3. However, as shown by the right panel of Figure 10, in the case of income risk inequality, the changes happen at the top of the income risk distribution. Indeed, the P90/P50 percentile ratio of CV increases by more than 2 with the recession, whereas the P50/P10 ratio remains approximately constant.

5.2 Correlates of income risk

We next turn to documenting several features of income risk and income risk inequality. We start by studying variation over the life cycle. In the upper left graph of Figure 11 we show the percentiles of CV by age. We find that younger individuals (less than 30 years old) tend to face higher levels of income risk. In addition, younger individuals face larger risk dispersion than older individuals.

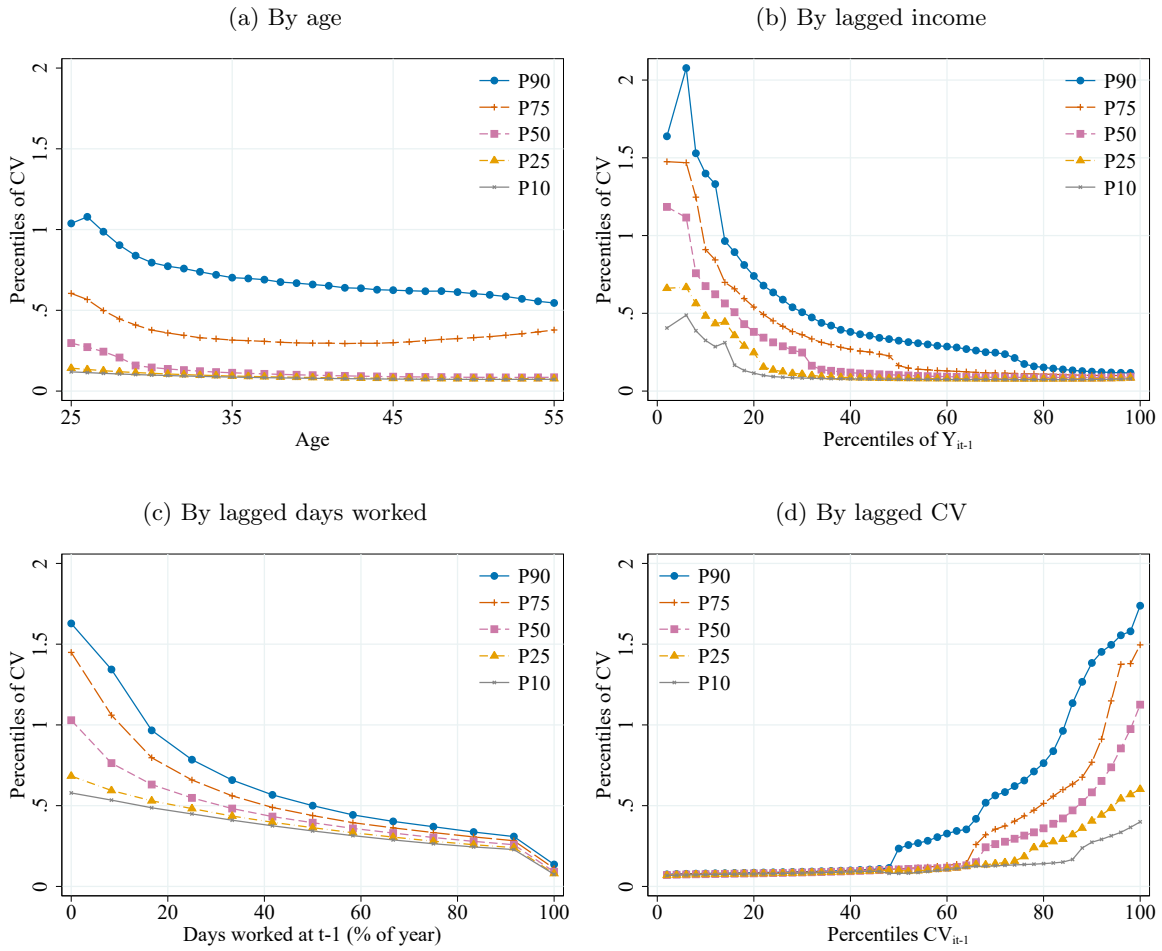
In order to illustrate the magnitude of the life-cycle variation in income risk, in Table 4 we report the ratio of age-specific percentiles of CV to the unconditional percentiles, by age. For example, the third row shows that, at the median, 25-year-olds experience almost three times as much risk — as measured by CV — compared to 35-year-olds. These patterns show remarkable variation in income risk and income risk inequality over the life cycle.

We next study how income risk and income risk inequality vary along the income distribution. For this purpose, in the upper right graph of Figure 11 we plot percentiles of CV as a function of lagged income percentiles. To produce the graph, we bin income into 50 categories, where the first category corresponds to zero income. We see a clear negative relationship between income and income risk. In addition, while high-income individuals face low levels and a small dispersion of income risk, individuals at the bottom of the income distribution face not only higher average income risk, but also a higher dispersion of CV.

It is interesting to compare our income risk measure with the income-based measures that we reported in the first part of this paper. Indeed, in Section 3 we documented several features of the dispersion of earnings changes conditional on lagged income. In order to compare such a measure to our CV, in Figure 12 we compare the distribution of income risk as measured by the CV, to the distribution of the conditional standard deviation of log income given lagged income. For the purpose of this comparison, we restrict the sample to positive income, and we rescale the standard deviation so as to make it comparable to the CV.[20] Compared to the conditional standard deviation of log income, the CV implies a larger proportion of low risk in the data, while also showing a long right tail, pointing to a substantial part of the economy facing high uncertainty in future income. Indeed, compared to the density of the conditional

[20] In Appendix Tables F11 and F12, we report summary statistics of the B sample conditional on positive income in 2018 euros and 2018 US dollars, respectively.

Figure 11: Correlates of income risk



Notes: B sample. Exponential specification, using all macro and micro predictors.

standard deviation, the density of CV is more skewed to the right and has a larger mass of observations near zero.[21]

Another key determinant of income risk is past employment. In the lower left graph of Figure 11 we show how the CV depends on the days worked in the past year. The graph shows a clear decreasing relationship between days worked and income risk. Individuals

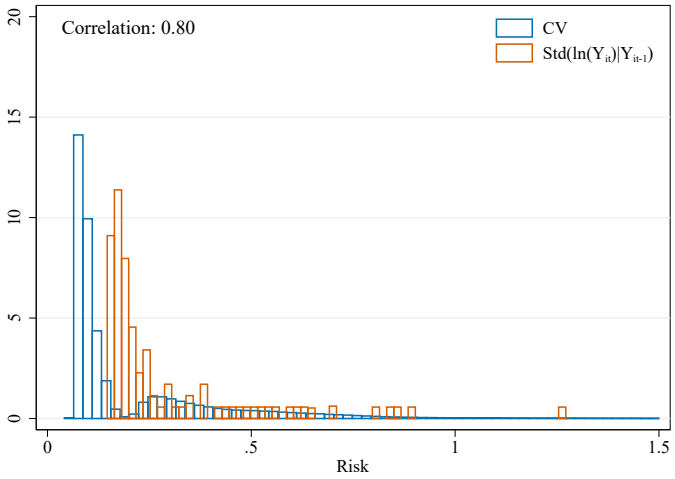
[21] In Appendix Figure G1 we compare various conditional percentiles of CV with the conditional standard deviation $\text{Std}(\log(Y_{it})|Y_{it-1})$, as a function of lagged income percentiles. The conditional standard deviation of log income suggests a level of risk that is close to the 90th percentile of risk implied by our CV. In addition, for any income level, our CV implies additional risk heterogeneity compared to the standard deviation of log income.

Table 4: Relative percentiles of income risk by age

	25	30	35	40	45	50	55
P1010	1.55	1.29	1.13	1.03	0.97	0.94	0.95
P2525	1.67	1.30	1.10	0.99	0.93	0.90	0.91
P5050	2.61	1.28	1.00	0.86	0.78	0.75	0.75
P7575	1.76	1.10	0.92	0.86	0.87	0.96	1.10
P9090	1.51	1.16	1.02	0.96	0.91	0.88	0.79

Notes: B sample. Exponential specification, using all macro and micro predictors. We report the relative percentiles $P\tau\tau = Q_\tau(\mathrm{CV_{it}}|age)/\mathrm{Q}_\tau(\mathrm{CV_{it}})$*, where* $Q_\tau(\mathrm{CV_{it}}|age)$ *is the* τ*th conditional percentile of CV given age, and* $Q_\tau(\mathrm{CV_{it}})$ *is the* τ*th unconditional percentile of CV.*

Figure 12: Comparing CV and standard deviation



Notes: B sample, with positive income. Exponential specification, using all macro and micro predictors. We compare the CV with a rescaled conditional standard deviation of log income. The correlation coefficient is computed after trimming the 99th percentiles of both measures.

working less than half the year face substantially higher risk, and a higher dispersion of risk. Individuals working full year face low risk and little risk dispersion.

As a fourth dimension of income risk, we next study its persistence at the individual level. In the lower right graph of Figure 11 we show how, for a given individual i, $\widehat{\mathrm{CV}}_{it}$ and $\widehat{\mathrm{CV}}_{it-1}$ relate to each other. We see that, when current risk CV is below the median, it is highly likely that the CV in the following year will be low.[22] This suggests that more than half of

[22] In addition, risk persistence tends to increase with age, as we show in Appendix Figure G2.

the Spanish economy is effectively shielded from income risk, at least in the short run. In contrast, current CV values exceeding the 60th percentile are associated with high CV values in the following period. Both the level and dispersion of future CV increase with current CV.

5.3 Robustness checks and extensions

Here we summarize results on income risk and income risk inequality, based on several alternative income measures and estimation techniques (details can be found in Appendix B). We produce risk estimates that accounts for income taxes. Moreover, we probe the robustness of of our results by extending our baseline specification in two ways. We first estimate the CV using neural networks, instead of the low-dimensional exponential specifications that we rely on for our main results. Second, we augment the set of predictors by including unobserved heterogeneity types in the specification, following the two-step grouped fixed-effects approach of Bonhomme et al. (2021). Lastly, we report results based on a median-based counterpart of the CV, in an attempt to minimize the impact of outliers. In all these specifications we obtain results that are qualitatively and quantitatively similar to our main results. As a last extension, we use quantile regressions estimate the entire conditional distribution of income given the predictors.[23]

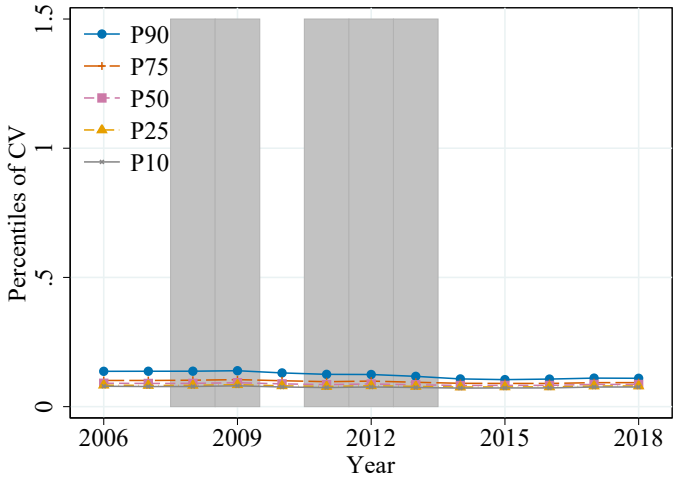
5.4 The income risk for civil servants

In this subsection we consider a particular category of workers, civil servants (*funcionarios*), as a convenient test case of the ability of our administrative records-based CV measure to correctly represent the risk individuals face. In Spain, as in other countries, civil servants are known to enjoy high levels of job and income security (see Antón and Muñoz de Bustillo, 2015). Thus, we expect them to face low income risk. In Figure 13 we plot the distribution of the CV for civil servants under permanent contracts.[24] The income risk levels we find are low compared to the rest of the economy: indeed, the 90th percentile of CV among civil servants is comparable to the median of the overall CV distribution. Moreover, the distribution is virtually unaffected by the recession. We interpret this exercise as suggesting that, for the

[23] We use this approach to document the level and evolution of the skewness of the predictive income distribution in Appendix Figure G17.

[24] In Appendix Table G8 we report the corresponding numbers.

Figure 13: CV over the period, civil servants



Notes: B sample, restricted to civil servants under permanent contracts. Exponential specification, using all macro and micro predictors. The shaded areas indicate recession years.

subsample of workers in civil service jobs, the CV accurately captures the low level and low variability of income risk that we would expect for contractual reasons.

6 Income risk: what do subjective expectations data say?

Our income risk measure is based on income and employment histories. However, the administrative data has no direct information on the agent's information set and beliefs. As a complement, in this subsection we compare our CV with an income risk measure calculated from data on subjective income expectations. For this purpose, we use the subjective probabilistic expectation question included in the Spanish Survey of Household Finances (Encuesta Financiera de las Familias, EFF). The EFF is a longitudinal survey undertaken by the Banco de España, which has been conducted since 2002 to obtain information about the wealth and financial conditions of Spanish households. Based on this information, we directly measure the uncertainty that households face about their future income growth by obtaining a subjective standard deviation for each respondent. If there is a broad agreement between the prediction-based measure and the subjective expectation-based measure, despite the many differences in the way they are constructed, this will strengthen our confidence in

both measures.

Starting in 2014, the EFF introduced a question to elicit household income probabilistic expectations (Bover et al., 2018). Households were asked to distribute ten points among five different scenarios concerning the change of their income over the next 12 months. In this way, respondents provide information not only about point expectations, but also about the probabilities they assign to different future outcomes. The exact wording is the following:

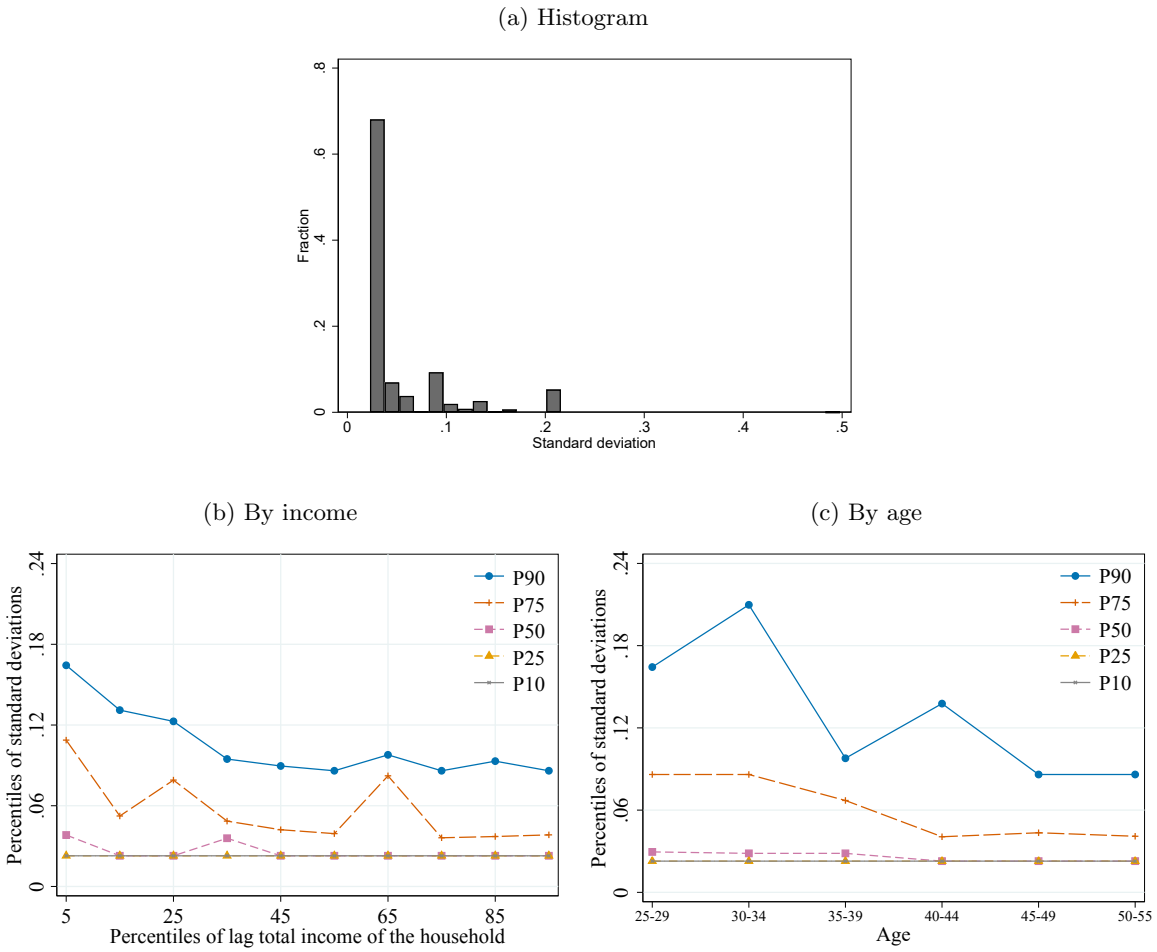
We are interested in knowing how you think the total annual income of your household will change in the next 12 months. Divide 10 points among the five options given below, assigning more points to the options you think are more likely (assign 0 point to options you think are impossible):

- *Drop of more than 10%*
- *Drop between 2% and 10%*
- *Approximately steady (falls or rises of no more than 2%)*
- *Increase between 2% and 10%*
- *Increase of more than 10%*

Thus, for every person who answered this question, we observe the fraction of points $\widehat{p}_j$ allocated to each event $j = 1, \ldots, J$ (adding up to 1), where $J = 5$. From that information, we calculate summary measures of dispersion under the assumption that the underlying probabilities are normally distributed. We provide the details of the method in Appendix E. Let us define $\widetilde{c}_j = \frac{\sum_{k=1}^{j} \widehat{p}_k + \frac{j}{2m}}{1+\frac{J}{2m}}$ to be regularized estimates of cumulative frequencies, and let $\widetilde{q}_j = \Phi^{-1}(1 - \widetilde{c}_j)$ be the standard normal quantiles of the complementary frequencies. The regularization parameter m can be thought of as a measure of the accuracy of the elicitation process. For the results that we now present, we take $m = 10$ — and verified that using m between 5 and 100 had small effects on the results. We then compute the following standard deviation estimates

$$\widetilde{\sigma} \quad = \quad \frac{2}{5\left(\widetilde{q}_1 - \widetilde{q}_4\right) + 25\left(\widetilde{q}_2 - \widetilde{q}_3\right)}.$$

Figure 14: Estimated subjective standard deviations



Notes: Estimated subjective standard deviations from the EFF 2014.

For this exercise, we use data from the 2014 wave of the EFF. We select all male household heads, aged between 25 and 55 years, who responded to the question about subjective expectations. A histogram of the standard deviation estimates $\widetilde{\sigma}$ in the top graph of Figure 14 shows a large proportion of low subjective risk together with a long right tail. In the bottom graphs of Figure 14 we show those standard deviation estimates by total income of the household in the previous year (on the left) and by age (on the right). Even though there are major differences in the way we capture income risk compared to our main analysis based on the MCVL, the calculations based on the EFF are qualitatively consistent with several of

the main lessons of the previous sections. Importantly, the subjective expectations question in the EFF refers to household income, as opposed to individual income as in the MCVL data.

The subjective standard deviations that we compute are close to 0.05 on average, which is consistent with a large share of the sample facing relatively low levels of income risk. Moreover, Figure 14 shows that there is substantial dispersion across households in terms of subjective standard deviations, which is qualitatively consistent with the evidence from the administrative data. In addition, similarly to what we obtained with our prediction-based CV measure of income risk, the figure shows that subjective standard deviations are higher in the bottom part of the income distribution, and also for younger household heads. While there is a good overall qualitative agreement, subjective risk is somewhat muted by comparison with MCVL risk. The nature of information extraction, the income concept, and the operation of household insurance are some of the factors that may play a role in explaining these differences.

7 Conclusion

We have developed a methodology for constructing measures of individual income risk and for quantifying the inequality of income risk. We have documented a number of new empirical facts regarding the dispersion, evolution, and dynamics of both income and income risk.

We have found evidence of high inequality of income security in the Spanish economy. A large mass of workers with negligible risk in their incomes coexists with many who anticipate fluctuations in their next year's income larger than 10 or 20 percent of their expected incomes. Additional key findings are that: (i) income risk is more unequal and higher on average among the young; (ii) inequality of income risk increases during the recessions; (iii) risk decreases with income, and (iv) lower levels of risk are more persistent than higher levels of risk. Beyond income inequality, inequality of income risk is thus a key feature of the Spanish economy. It would be of great interest to document it in other settings.

Some of the underlying causes of the inequality of income risk that we have documented are familiar to the labor economists that studied Spanish unemployment and the consequences of temporary/permanent dual labor markets. However, we have taken a different perspective

that abstracts from shorter-term labor market transitions and puts the focus on the unequal income risks that individuals face on a relevant time horizon.

The analysis could be extended in a number of directions. First, an open question is the extent to which individual risks are mitigated at the household level, and how demographic risks interact with income risks in the short and long run. Second, since different components of income may have different degrees of persistence, it would be valuable to map our approach into models with multiple latent components, which are key features of the permanent income hypothesis and the literature on consumption insurance (Friedman, 1957, Hall and Mishkin, 1982, Blundell et al., 2008). Third, although we have not distinguished the sources of risk that are exogenous to the agent from those that are the result of choice, this distinction is important to account for example for labor market attachment and labor force participation. Fourth, while we have only studied annual income risk, the MCVL administrative records may also be useful to document within-year income fluctuations and their risk consequences (Morduch and Schneider, 2019). Finally, an interesting direction will be to structurally estimate the welfare costs of income risk, and the inequality of those economic costs, along the lines of the discussion in Section 4.

References

Alvaredo, Facundo, Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman (2017) "Global Inequality Dynamics: New Findings from WID.world," *American Economic Review*, 107 (5), 404–09, 10.1257/aer.p20171095.

Alvarez, Javier and Manuel Arellano (2021) "Robust likelihood estimation of dynamic panel data models," *Journal of Econometrics*, https://doi.org/10.1016/j.jeconom.2021.03.005.

Anghel, Brindusa, Henrique Basso, Olympia Bover, José María Casado, Laura Hospido, Mario Izquierdo, Ivan A. Kataryniuk, Aitor Lacuesta, José Manuel Montero, and Elena Vozmediano (2018) "Income, consumption and wealth inequality in Spain," *SERIEs*, 9 (4), 351–387, 10.1007/s13209-018-0185-1.

Antón, José-Ignacio and Rafael Muñoz de Bustillo (2015) "Public-private sector wage differentials in Spain. An updated picture in the midst of the Great Recession," *Investigación Económica*, 74 (292), 115–157, https://doi.org/10.1016/j.inveco.2015.08.005.

Arellano, Manuel (2014) "Uncertainty, Persistence, and Heterogeneity: A Panel Data Perspective," *Journal of the European Economic Association*, 12 (5), 1127–1153, 10.1111/jeea.12105.

Arellano, Manuel, Richard Blundell, and Stéphane Bonhomme (2017) "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework," *Econometrica*, 85 (3), 693–734, https://doi.org/10.3982/ECTA13795.

Atkinson, A. B. (2003) "Income Inequality in OECD Countries: Data and Explanations," *CESifo Economic Studies*, 49 (4), 479–513, 10.1093/cesifo/49.4.479.

Bloom, Nicholas, Fatih Guvenen, Luigi Pistaferri, John Sabelhaus, Sergio Salgado, and Jae Song (2017) "The great micro moderation," Working Paper.

Blundell, Richard, Luigi Pistaferri, and Ian Preston (2008) "Consumption Inequality and Partial Insurance," *American Economic Review*, 98 (5), 1887–1921, 10.1257/aer.98.5.1887.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2021) "Discretizing unobserved heterogeneity," Working Paper.

Bonhomme, Stéphane and Laura Hospido (2013) "Earnings inequality in Spain: new evidence using tax data," *Applied Economics*, 45 (30), 4212–4225, 10.1080/00036846.2013.781261.

——— (2017) "The Cycle of Earnings Inequality: Evidence from Spanish Social Security Data," *The Economic Journal*, 127 (603), 1244–1278, 10.1111/ecoj.12368.

Bover, Olympia, Laura Crespo, Carlos Gento, and Ismael Moreno (2018) "The Spanish survey of household finances (EFF): description and methods of the 2014 wave," Technical report, Banco de España, https://repositorio.bde.es/bitstream/123456789/6404/1/do1804e.pdf.

Browning, Martin, Mette Ejrnæs, and Javier Alvarez (2010) "Modelling Income Processes with Lots of Heterogeneity," *The Review of Economic Studies*, 77 (4), 1353–1381, 10.1111/j.1467-937X.2010.00612.x.

Busch, Christopher, David Domeij, Fatih Guvenen, and Rocio Madera (forthcoming) "Skewed idiosyncratic income risk over the business cycle: Sources and insurance," *American Economic Journal: Macroeconomics*.

Deaton, Angus (1992) *Understanding consumption*, Oxford England New York: Clarendon Press Oxford University Press.

Dominitz, Jeff and Charles F. Manski (1997) "Using Expectations Data to Study Subjective Income Expectations," *Journal of the American Statistical Association*, 92 (439), 855–867, 10.1080/01621459.1997.10474041.

Farrell, Max H., Tengyuan Liang, and Sanjog Misra (2021) "Deep Neural Networks for Estimation and Inference," *Econometrica*, 89 (1), 181–213, https://doi.org/10.3982/ECTA16901.

Felgueroso, Florentino, José Ignacio García-Pérez, Marcel Jansen, and David Troncoso Ponce (2017) "Recent trends in the use of temporary contracts in Spain," Working Paper.

Friedman, Milton (1957) *A theory of the consumption function*, Princeton: Princeton University Press.

Geweke, John and Michael Keane (2000) "An empirical analysis of earnings dynamics among men in the PSID: 1968–1989," *Journal of Econometrics*, 96 (2), 293–356, https://doi.org/10.1016/S0304-4076(99)00063-9.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016) *Deep Learning*: MIT Press, http://www.deeplearningbook.org.

Gottschalk, Peter and Robert Moffitt (1994) "The Growth of Earnings Instability in the U.S. Labor Market," *Brookings Papers on Economic Activity*, 25 (2), 217–272, https://www.brookings.edu/bpea-articles/the-growth-of-earnings-instability-in-the-u-s-labor-market/.

——— (2009) "The Rising Instability of U.S. Earnings," *Journal of Economic Perspectives*, 23 (4), 3–24, 10.1257/jep.23.4.3.

Guvenen, Fatih, Fatih Karahan, Serdar Ozkan, and Jae Song (forthcoming) "What do data on millions of US workers reveal about life-cycle earnings," *Econometrica*.

Guvenen, Fatih, Serdar Ozkan, and Jae Song (2014) "The Nature of Countercyclical Income Risk," *Journal of Political Economy*, 122 (3), 621–660, 10.1086/675535.

Haider, Steven J. (2001) "Earnings Instability and Earnings Inequality of Males in the United States: 1967–1991," *Journal of Labor Economics*, 19 (4), 799–836, 10.1086/322821.

Hall, Robert E. and Frederic S. Mishkin (1982) "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households," *Econometrica*, 50 (2), 461–481, http://www.jstor.org/stable/1912638.

Hoffmann, Eran B. and Davide Malacrino (2019) "Employment time and the cyclicality of earnings growth," *Journal of Public Economics*, 169, 160–171, https://doi.org/10.1016/j.jpubeco.2018.09.009.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989) "Multilayer feedforward networks are universal approximators," *Neural Networks*, 2 (5), 359–366, https://doi.org/10.1016/0893-6080(89)90020-8.

Kaufmann, Katja and Luigi Pistaferri (2009) "Disentangling Insurance and Information in Intertemporal Consumption Choices," *American Economic Review*, 99 (2), 387–92, 10.1257/aer.99.2.387.

Lucas, Robert E. (1987) *Models of Business Cycles*, Yrjö Jahnsson lectures: Basil Blackwell, https://books.google.es/books?id=JiB9swEACAAJ.

Meghir, Costas and Luigi Pistaferri (2004) "Income Variance Dynamics and Heterogeneity," *Econometrica*, 72 (1), 1–32, https://doi.org/10.1111/j.1468-0262.2004.00476.x.

Morduch, J. and R. Schneider (2019) *The Financial Diaries: How American Families Cope in*

a World of Uncertainty: Princeton University Press, https://books.google.es/books?id=M3SYDwAAQBAJ.

Pijoan-Mas, Josep and Virginia Sánchez-Marcos (2010) "Spain is Different: Falling Trends of Inequality," *Review of Economic Dynamics*, 13 (1), 154–178, 10.1016/j.red.2009.10.002.

Piketty, Thomas and Emmanuel Saez (2013) "Top Incomes and the Great Recession: Recent Evolutions and Policy Implications," *IMF Economic Review*, 61 (3), 456–478, 10.1057/imfer.2013.14.

Pora, Pierre and Lionel Wilner (2020) "A decomposition of labor earnings growth: Recovering Gaussianity?" *Labour Economics*, 63, 101807, https://doi.org/10.1016/j.labeco.2020.101807.

Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron (2004) "Cyclical Dynamics in Idiosyncratic Labor Market Risk," *Journal of Political Economy*, 112 (3), 695–717, http://www.jstor.org/stable/10.1086/383105.

Ziliak, James P., Bradley Hardy, and Christopher Bollinger (2011) "Earnings volatility in America: Evidence from matched CPS," *Labour Economics*, 18 (6), 742–754, https://doi.org/10.1016/j.labeco.2011.06.015, European Association of Labour Economists, 3rd World Conference EALE/SOLE, London UK, 17-19 June2010.

APPENDIX to "Income Risk Inequality: Evidence from Spanish Administrative Records"

A The welfare cost of income risk

Consider an individual with utility function $U_i(C_{it}) = \frac{(C_{it}-C_m)^{1-\theta_i}-1}{1-\theta_i}$, with consumption $C_{it} = \lambda(X_{it})Y_{it}$ for some proportionality factor $\lambda(X_{it})$. Suppose that

$$\ln(C_{it} - C_m) \,|\, X_{it} = x \sim \mathcal{N}(\mu(x), \sigma(x)^2).$$

The individual is willing to give up $a\%$ of consumption each period in order to eliminate income risk, where a solves

$$U_i(\mathbb{E}(C_{it} \,|\, X_{it})(1-a)) = \mathbb{E}\left[U_i(C_{it}) \,|\, X_{it}\right].$$

Equivalently, omitting the dependence on X_{it} for simplicity, a solves

$$\begin{aligned}\left(\left(C_m + \exp\left(\mu + \frac{1}{2}\sigma^2\right)\right)(1-a) - C_m\right)^{1-\theta_i} &= \mathbb{E}\left[(C_{it} - C_m)^{1-\theta_i}\right] \\ &= \exp\left[(1-\theta_i)\mu + \frac{1}{2}(1-\theta_i)^2\sigma^2\right].\end{aligned}$$

It follows that

$$a = 1 - \frac{C_m + \exp\left[\mu + \frac{1}{2}(1-\theta_i)\sigma^2\right]}{C_m + \exp\left(\mu + \frac{1}{2}\sigma^2\right)}.$$

Hence, for small σ,

$$a = \frac{1}{2}\theta_i \frac{\exp(\mu)}{C_m + \exp(\mu)}\sigma^2 + o(\sigma^2).$$

Now, we have

$$\mathrm{CV} = \sqrt{\frac{2}{\pi}}\frac{\exp(\mu)}{\mathrm{C_m} + \exp(\mu)}\sigma + \mathrm{o}(\sigma).$$

Hence

$$a = \frac{\pi}{4}\theta_i \frac{C_m + \exp(\mu)}{\exp(\mu)}\mathrm{CV}^2 + \mathrm{o}(\sigma^2).$$

This implies (3), and (2) in the special case where $C_m = 0$.

B Robustness checks and extensions

B.1 After-tax income

In the main analysis, we have computed risk based on pre-tax income. In order to assess how the tax system may affect our calculations of income risk, in this subsection we construct a measure of after-tax income, and apply our approach to this measure.

We use administrative data on tax returns to estimate average effective tax rates with respect to gross labor income. We apply these average effective tax rates to our measure of income to calculate after-tax income. This approach follows Garcia-Miralles et al. (2019), who estimate tax functions of the Spanish personal income tax. The data source we use for this part is representative of the population of Spanish taxpayers. We have repeated cross-sections available from 2005 to 2017. We select males aged 25-55, who file individual tax returns. In the case of joint filing, it is not possible to distinguish the income that corresponds to each household members. For each year and individual in the sample, we compute gross labor income and tax liabilities. We compute effective tax rates as tax liabilities over gross labor income. In all calculations, we restrict the sample to taxpayers with positive gross labor income. We consider the income brackets set by the Government each year, and calculate average effective tax rates within each interval. We report the income brackets and corresponding average effective tax rates in Table G2. The simplified tax rules that we use here give an approximately linear relationship between before-tax and after-tax income (see Figure G3).

In Figure G5 we reproduce the main results from Section 5 using our CV calculated using after-tax income. The top left panel shows the quantiles of CV, calculated using after-tax income, over the sample period. The level and evolution of income risk are very similar when using after-tax or before-tax income (compare with Figure 9). In Table G3 we report the corresponding numbers, which are close to those in Table 3. In Figure G4 we see that the percentiles of CV based on before-tax and after-tax income are very close to each other. Moreover, the patterns of income risk over the life cycle, its relationship with income, and its persistence, are all similar when using after-tax or before-tax income.

In contrast, the presence of unemployment benefits tends to dampen both the level and

dispersion in income risk. This can be seen in Figure G6, where we reproduce the main results from Section 5 using an income measure net of unemployment benefits.

B.2 Neural network CV

There are two reasons why our income risk inferences may be incorrect: the set of predictors X_{it} may not correspond to the agent's information set, or the prediction model may be misspecified. Here we focus on the latter concern, and use a flexible prediction method, instead of the exponential specification that we have relied on so far.

In order to make the prediction more flexible, we rely on a neural network to estimate the CV. Consider the denominator of our CV, which is the conditional mean of income $\mathbb{E}(Y_{it}|X_{it})$. A feed-forward neural network with one hidden layer is

$$\mathbb{E}(Y_{it}|X_{it}) = \exp\left(\beta_0 + \sum_{m=1}^{M} \beta_m \tau(X_{it}'\alpha_m)\right),$$

where M is the number of nodes, τ is a nonlinear function, and β_m and α_m are parameters. We model the numerator of CV similarly using the same τ function, and different number of nodes and parameters.

Following the recent literature (e.g., Goodfellow et al., 2016), we take $\tau(u) = \max(u, 0)$, which corresponds to the "rectified linear unit" ReLU function. We use the Poisson loss function. To choose the number of nodes M, we perform a single-fold cross-validation strategy, using 2005-2016 as the estimation sample and 2017 as the hold-out sample. This gives $M = 8$ nodes for estimating the mean (i.e., the CV denominator), and 7 nodes for estimating the mean absolute deviation (i.e., the CV numerator).[2] We focus on one-layer specifications for parsimony,[3] but we have performed some robustness checks using additional layers and adding penalty terms.

In Figure G7 we reproduce some of our main findings, now based on neural network specifications to construct the CV risk measure. In Table G4 we report the corresponding

[2]The neural network estimation was implemented in R with the package "H2O" (H2O.ai, 2020). The Poisson loss function is minimized using a parallelized version of stochastic gradient descent. The default parameters for the number of epochs is 10 and the number of training samples per iteration is adapted, trading-off computation time and communication between parallel clusters.

[3]To reduce the variance in the predictions of the neural network of a given architecture, we run the estimation algorithm on the full sample 15 times and average the corresponding predictions. To compare estimates, we trim the 99th percentiles of the mean squared error.

numbers. Overall, the results based on the neural networks are quantitatively similar to the baseline ones. In particular, we find that income risk inequality increases in the recession, that income risk is higher for the young and the low income, and that it is highly persistent.

In Figure G8 we provide a direct comparison between the CV computed using our baseline exponential specification and the CV computed using a neural network approach. We see that the densities of the two measures agree well, and that the two CV are highly correlated, the correlation coefficient being 0.98.

B.3 Incorporating unobserved heterogeneity: grouped fixed-effects

To mimic the individual's prediction problem, it may be important to account for predictors that we as researchers do not observe. In this subsection, we augment the set of predictors as (X_{it}, ξ_i), where ξ_i is a latent component. For this purpose, we use a grouped fixed-effects approach. Following Bonhomme et al. (2021) we first group individuals into K categories, and then include the group indicators as predictors to estimate CV. We use different groups for the conditional mean (the denominator) and the conditional mean absolute deviation (the numerator). The benefit of this approach is the ability to handle incomplete models without having to specify a model for the predetermined variables X_{it}, initial conditions, and unobserved heterogeneity (Hahn and Kuersteiner, 2011, Arellano and Hahn, 2016).[4] In Section C we describe implementation, and we provide descriptive information about the groups that we estimate.

In estimation, we take 4 groups for both the numerator and denominator of the CV. We have experimented with different numbers of groups. Given the estimated groups, we estimate the conditional mean and conditional mean absolute deviation of income Y_{it} given the observed predictors X_{it} and the groups using Poisson regressions, where we account for interactions between the group indicators and a quadratic in age. As in the case with other nonlinear fixed effects estimators, the consistency of the grouped fixed-effects approach requires the number of time periods to tend to infinity together with the number of individuals. To reduce the noise in the grouping, for this analysis we restrict the sample to individuals with at least 4 observations prior to 2018.

[4] A rich parametric model of this kind is Altonji et al. (2013).

In the top left graph of Figure G11 we report the percentile of CV over the period, obtained using the prediction method that allows for group-level heterogeneity. In Table G6 we provide the numbers.[5] Compared to the risk estimates without unobserved heterogeneity (see Table 3), the specification with unobserved heterogeneity implies somewhat lower levels of risk. For example, the 10th (respectively, 90th) percentile of CV ranges between 0.05 and 0.06 (resp., 0.39 and 0.78), compared to between 0.07 and 0.08 (resp., 0.42 and 0.93) in the specification without unobserved heterogeneity. At the same time, the results remain qualitatively similar to the baseline. Moreover, the other three graphs of Figure G11 show that other main features of the risk and its relationship with age and income are preserved.

Lastly, in Figure G12 we compare the CV computed according to our baseline specification to the one computed using the model with unobserved heterogeneity. The CV densities agree with each other, although the model with heterogeneity tends to predict somewhat lower risk at the bottom of the risk distribution. The two CV measures are correlated, though not perfectly, the correlation coefficient being 0.85.

B.4 Robust CV

The CV measure that we use in the main analysis is computed as a ratio between two conditional means; see equation (1). In order to alleviate sensitivity to outliers, one may alternatively compute the following median-based counterpart:

$$\widetilde{\text{CV}}(X_{it}) = \frac{\text{median}(|Y_{it} - \text{median}(Y_{it}|X_{it})| \mid X_{it})}{\text{median}(Y_{it} \mid X_{it})}. \tag{B5}$$

This "robust" counterpart to the CV has the conditional median income in the denominator, and the conditional median absolute deviation in the numerator, where the absolute deviation is computed relative to the conditional median income (Arachchige et al., 2020). We estimate the numerator and denominator using median regressions, with all macro and micro predictors as regressors. Since both income and income absolute deviation are non-negative, in Section D we describe a method based on Buchinsky and Hahn (1998), which we use to enforce the non-negativity of the median outcomes in estimation.

[5] In Figures G9 and G10 and in Table G5, we show several descriptive statistics about the groups that we estimate. In Figure G13 and Table G7, we report the results for 6 groups.

In Figure G14 we reproduce the main findings of Section 5 using the robust CV measure given by equation (B5). The evolution of this measure during the period and over the life cycle, its relationship with income, and its persistence, are all similar to what we found using our baseline CV measure of income risk.

B.5 Beyond the CV: other statistical measures of risk

In the analysis so far, we have used the coefficient of variation as the metric to quantify risk. In this subsection we document the level and evolution of other measures of risk, which, similarly to CV, are based on the conditional distribution of income given predictors.

We start by reporting results for a quantile-based measure of dispersion. In this exercise we restrict the B sample to observations with positive income. We compute the percentile difference $P90(X_{it}) - P10(X_{it})$, where we estimate $P90(X_{it})$ and $P10(X_{it})$ using linear quantile regressions of log income on the predictors. In Figure G15 we reproduce several of the main findings that we previously documented using CV, now using the percentile measure. This measure aligns well, qualitatively, with our CV. Indeed, the top left graph in Figure G15 shows that income risk inequality increases in the recession, while the top right graph shows that risk tends to be higher for younger individuals. The bottom left graph shows an inverse relationship between risk, as measured by $P90 - P10$, and income, while the bottom right graph shows that income risk is highly persistent, especially in the bottom half of the risk distribution. Quantitatively, the risk values $P90 - P10$ are higher than the CV values.[6]

In order to directly compare the quantile-based measure of risk $P90 - P10$ to the CV, in Figure G16 we plot the histogram of the CV, along with the histogram of the percentile difference (suitably rescaled). The two histograms agree quite well. We also see that both measures of risk are highly correlated, with a correlation coefficient of 0.98.

By estimating quantile regressions, we are able to document the entire conditional distribution of log income given the predictors, not only its dispersion and location. A quantity of particular interest is the skewness. In Section 3 we showed how the skewness of log annual earnings changes decreases during the Spanish recession. In Figure G17 we re-

[6] This is to be expected. For example, under log normality of income and for small standard deviation, the ratio between the two measures is approximately $\frac{\Phi^{-1}(0.9)-\Phi^{-1}(0.1)}{\sqrt{\frac{2}{\pi}}} \approx 3.2$.

port the evolution of quantiles of a percentile-based measure of skewness of the conditional distribution of log income given the predictors. Specifically, we report Kelley's skewness, $\frac{P90(X_{it})-2P50(X_{it})+P10(X_{it})}{P90(X_{it})-P10(X_{it})}$, where we estimate $P90(X_{it})$, $P50(X_{it})$, and $P10(X_{it})$ using linear quantile regressions of log income on the predictors. The graph shows that, like the skewness of log income changes, the skewness of the conditional distribution of log income given the predictors also decreases during the recession. In addition, while the higher quantiles of this "skewness risk" vary quite substantially over the period, the lower quantiles show little variation. In Figure G18 we show how the skewness measure varies over the life cycle, relates to income, and persists over time.

C Unobserved heterogeneity: a grouped fixed-effects approach

In this section we describe how we allow for unobserved predictors. To implement the grouped fixed-effects approach of Bonhomme et al. (2021), one possibility would be to group individuals based on their mean income. However, in an unbalanced panel this approach tends to mix individual-specific heterogeneity and age, particularly when there is a strong life-cycle component to income. To account for age, we proceed in two steps.

In the first step, we maximize the Poisson log likelihood

$$\sum_{i,t} \left[Y_{it} \widetilde{X}_{it}' \beta(k_i) - \exp(\widetilde{X}_{it}' \beta(k_i)) \right], \tag{C6}$$

with respect to parameters $\beta(k)$ and group indicators k_i, where $\widetilde{X}_{it} = (1, age_{it}, age_{it}^2)'$. To implement the minimization, we use a Lloyd-like algorithm (see Bonhomme and Manresa, 2015). We start with 20 randomly chosen parameter values, and select the solution that corresponds to the highest value of the objective function. Given each starting value, we stop the algorithm when the change in the log likelihood is less than 10^{-10}. This first step gives us parameters $\widetilde{\beta}(k)$ and group indicators $\widetilde{k}_i$.

In the second step, we include all our macro and micro predictors X_{it}. We aim to maximize

$$\sum_{i,t} \left[Y_{it} X_{it}' \beta(k_i) - \exp(X_{it}' \beta(k_i)) \right], \tag{C7}$$

again with respect to parameters $\beta(k)$ and group indicators k_i. In this specification, we account for group effects in the intercept and the coefficients of age and age squared. We use

10 iterations of a Lloyd-like algorithm. The group indicators $\widetilde{k}_i$ provide a starting value to initiate the algorithm. This second step gives us parameters $\widehat{\beta}(k)$ and group indicators $\widehat{k}_i$, which are the ones we use to construct our prediction $\exp(X_{it}'\widehat{\beta}(\widehat{k}_i))$, which is the denominator of CV. We proceed similarly for the numerator. Hence, the CV coefficient is constructed using two sets of groups.

We apply the method to the B sample, restricted to individuals with at least 4 observations, using 4 groups for both the numerator and the denominator of CV. In Figure G9 we show the income profiles by age for the estimated clusters that we obtain using the above algorithm to predict income levels (the "mean clusters"). The clusters differ not only in the level of income, but also in the curvature of the profiles. In Figure G10 we show the CV profiles by age for all combinations of the mean clusters and the clusters that we obtain when predicting income absolute deviations (the "absolute deviation clusters"). We see that, while the former tends to capture differences in risk levels between individuals, the latter tends to pick up differences in age profiles. Lastly, in Table G5 we show the number of individuals in each cluster and the corresponding distribution of educational attainment. We find that the mean clusters are correlated with education, consistently with the fact that the higher educated tend to face lower risk levels. In contrast, the absolute deviation clusters are virtually independent of education.

D Enforcing non-negativity of outcomes in quantile regression

Let $Y \mid X$ be a non-negative random variable with quantile function $g(x,\tau)$. In our application, Y is income or income absolute deviation, and X includes our micro and macro predictors. Let $\Pr(Y=0|X=x)=\pi(x)$. Thus, $g(x,\tau)=0$ for $\tau \leq \pi(x)$. Also,

$$\Pr(Y \leq g(x,\tau) \mid Y>0, \tau>\pi(x), X=x) = \frac{\tau-\pi(x)}{1-\pi(x)}.$$

Denote the s-conditional quantile of Y given $Y>0, X=x$ by $Q_s(Y \mid Y>0, X=x)=\psi(x,s)$ so that

$$g(x,\tau) = \begin{cases} 0 & \text{for } \tau \leq \pi(x) \\ \psi\left(x, \frac{\tau-\pi(x)}{1-\pi(x)}\right) & \text{for } \tau > \pi(x) \end{cases}.$$

For $\pi(x) \approx 0$, we have $g(x,\tau) \approx \psi(x,\tau)$.

An estimator of $g(x,\tau)$ that enforces non-negativity is as follows. First, we obtain the estimates $\widehat{\psi}(x,s) = \exp\left[\widehat{\gamma}(s)'\varphi(x)\right]$, and, using logit, $\widehat{\pi}(x) = \Lambda\left[\widehat{\beta}'\varphi(x)\right]$ where $\Lambda(v) = \exp(v)/(1+\exp(v))$. To get $\widehat{\gamma}(s)$, we run linear quantile regressions of log income on $\varphi(x)$ in the subsample with positive income; $\widehat{\beta}$ is a logit estimate; and $\varphi(x)$ is a vector of functions of x. Finally, we compute:

$$\widehat{g}(x,\tau) = \begin{cases} 0 & \text{for } \tau \leq \widehat{\pi}(x) \\ \widehat{\psi}\left(x, \frac{\tau-\widehat{\pi}(x)}{1-\widehat{\pi}(x)}\right) & \text{for } \tau > \widehat{\pi}(x) \end{cases}.$$

To ease the computational burden, we model $\gamma(s)$ as piecewise linear interpolating splines on a grid, $[\tau_1,\tau_2],[\tau_2,\tau_3],...,[\tau_{L-1},\tau_L]$, contained in the unit interval. The intercept coefficients on $(0,\tau_1]$ and $[\tau_L,1)$ are parameterized as the quantiles of an exponential distribution on their respective supports. In practice, we use an equally spaced grid on the unit interval with $L = 11$.

E Estimating measures of location and dispersion from subjective probabilistic income expectations

For each respondent in the Spanish Survey of Household Finances (EFF) who answers the subjective probabilistic income expectations question described in Section 6, we observe the fraction of points $\hat{p}_j$ allocated to each event $j = 1,\ldots,5$ (adding up to 1). Here we present a simple approach to calculate summary measures of location and dispersion from these observations under the assumption that the underlying probabilities are normally distributed (with mean μ and standard deviation σ). That is, we assume that the underlying process for next year's log income is a random walk with a household-specific drift μ_i and normally distributed shocks with a household-specific standard deviation σ_i. The household index is omitted in the text for conciseness.

Under normality, $p_1 = \Phi(-0.1\beta-\alpha)$, $p_2 = \Phi(-0.02\beta-\alpha) - \Phi(-0.1\beta-\alpha)$, $p_3 = \Phi(0.02\beta-\alpha) - \Phi(-0.02\beta-\alpha)$, $p_4 = \Phi(0.1\beta-\alpha) - \Phi(0.02\beta-\alpha)$, $p_5 = 1-\Phi(0.1\beta-\alpha)$, for $\alpha = \mu/\sigma$ and $\beta = 1/\sigma$. Elicited probabilities $\hat{p}_j$ can be regarded as noisy measurements of p_j due to rounding and the inherent randomness in the elicitation process. If the $\hat{p}_j$ are regarded as sample frequencies from a hypothetical random sample of size m, they are the

unrestricted maximum likelihood estimates of the p_j from the multinomial log likelihood:

$$L(p) = \sum_j \hat{p}_j \log p_j.$$

Alternatively, $p = (p_1, \ldots, p_4)$ can be estimated as the posterior mean of a posterior distribution:

$$\pi(p|\hat{p}) \propto \exp\left[L(p) + \frac{1}{m}\log \pi(p)\right]$$

for some chosen prior $\pi(p)$ and value of m (e.g., $m = 10$). A conventional option is to use Jeffreys prior, $\log \pi(p) = -\frac{1}{2}\sum_{j=1}^{5} \log p_j$. This is a Dirichlet distribution of order $J = 5$, which is the conjugate prior of the multinomial, therefore the posterior is also Dirichlet with posterior means given by:

$$\tilde{p}_j = \frac{\hat{p}_j + \frac{1}{2m}}{1 + \frac{J}{2m}}.$$

Jeffreys prior adds $J/2$ observations to the likelihood with equally distributed probabilities. The modified estimator $\tilde{p}_j$ has the advantage that it takes values in the open interval $(0, 1)$ so that the inverse normal cdf transformation is still defined when $\hat{p}_j = 0$ or 1. In the present statistical framework, m measures the accuracy of the process of eliciting subjective probabilities.

We implement this approach, using a Berkson estimator that enforces the Gaussian restrictions on the posterior means $\tilde{p}_j$. This estimator is based on the inverse normal probabilities:

$$\begin{aligned}
q_1 &= \Phi^{-1}(1 - c_1) = 0.1\beta + \alpha \\
q_2 &= \Phi^{-1}(1 - c_2) = 0.02\beta + \alpha \\
q_3 &= \Phi^{-1}(1 - c_3) = -0.02\beta + \alpha \\
q_4 &= \Phi^{-1}(1 - c_4) = -0.1\beta + \alpha
\end{aligned}$$

where the c_j are the cumulative probabilities $c_j = \sum_{k=1}^{j} p_k$. Sample counterparts are

$$\tilde{c}_j = \sum\nolimits_{k=1}^{j} \tilde{p}_k = \frac{\hat{c}_j + \frac{j}{2m}}{1 + \frac{J}{2m}}$$

and $\tilde{q}_j = \Phi^{-1}(1 - \tilde{c}_j)$. To estimate μ and σ we choose a particular solution of the previous

system with an intuitive interpretation:

$$\begin{aligned}\tilde{\mu} &= \frac{(\tilde{q}_1+\tilde{q}_4)+(\tilde{q}_2+\tilde{q}_3)}{5(\tilde{q}_1-\tilde{q}_4)+25(\tilde{q}_2-\tilde{q}_3)},\\ \tilde{\sigma} &= \frac{2}{5(\tilde{q}_1-\tilde{q}_4)+25(\tilde{q}_2-\tilde{q}_3)}.\end{aligned}$$

Lastly, mimicking the fact that the maximum likelihood estimator of $\theta=(\alpha,\beta)$ is $\widehat{\theta}=\arg\max_\theta \sum_j \widehat{p}_j \ln p_j(\theta)$, an alternative method to enforce the restrictions on the posterior means $\widetilde{p}_j$ would be to compute

$$\widetilde{\theta}=\arg\max_\theta \sum_j \widetilde{p}_j \ln p_j(\theta).$$

Nevertheless, in our empirical calculations of subjective risks we relied on the Berkson estimator, which has a simple closed-form expression.

F Additional tables and figures on income inequality and dynamics

Table F1: Observations below the income threshold

	Male		Female	
	Observations	Proportion	Observations	Proportion
2005	241674	0.033	197333	0.079
2006	255047	0.050	216508	0.110
2007	268115	0.063	234212	0.131
2008	274655	0.083	245235	0.140
2009	278062	0.143	250439	0.180
2010	278195	0.171	252533	0.198
2011	272740	0.179	250954	0.200
2012	270691	0.212	248709	0.222
2013	266659	0.220	245710	0.234
2014	259429	0.193	240226	0.217
2015	251970	0.155	235180	0.185
2016	247453	0.131	232652	0.166
2017	244782	0.110	232063	0.149
2018	242451	0.094	230715	0.134

Notes: Number of non-missing observations, and proportion of observations below the income threshold $\underline{y}_t$, by year and gender. The threshold that we use in Section 3 corresponds to working part time for one quarter at the national minimum wage.

Table F2: Descriptive statistics

Year	Obs	Mean Income		Females	Age Shares %			Education Shares %			
	(×1000)	Males	Females	% Share	[25,35]	[36,45]	[46,55]	Primary	Lower Sec	Upper Sec	College
2005	415	25434	18073	43.8	44.6	33.3	22.1	14.5	36.5	28.2	20.7
2006	435	25830	18496	44.3	43.6	33.7	22.7	14.3	36.6	27.8	21.3
2007	455	26146	18951	44.8	42.8	34.1	23.1	14.1	36.7	27.5	21.6
2008	463	26045	19278	45.6	41.6	34.6	23.8	13.7	36.6	27.4	22.3
2009	444	25952	19751	46.3	39.7	35.4	25.0	12.7	36.3	27.6	23.5
2010	433	25460	19387	46.8	38.0	36.1	25.9	12.1	36.4	27.5	24.1
2011	425	24771	18724	47.3	36.4	36.9	26.8	11.6	36.4	27.4	24.6
2012	407	23807	18093	47.6	34.5	37.8	27.7	11.1	35.8	27.6	25.6
2013	396	23091	17885	47.5	32.7	38.6	28.7	11.1	35.8	27.2	25.9
2014	397	22915	17943	47.3	31.4	39.0	29.5	11.2	35.9	26.8	26.1
2015	405	23354	18316	47.4	30.3	39.2	30.5	11.2	35.9	26.5	26.4
2016	409	23788	18741	47.4	29.5	39.1	31.4	11.1	35.7	26.3	26.8
2017	415	23817	18760	47.5	29.0	38.7	32.3	11.1	35.6	26.1	27.2
2018	420	24115	19018	47.6	28.7	37.9	33.4	11.0	35.4	26.1	27.6

Notes: CS sample. Annual earnings are reported in 2018 euros.

Table F3: Percentiles of annual earnings

Year	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99
2005	1579	3525	5898	11725	17965	27069	40589	53102	89988	113145
2006	1641	3769	6282	12149	18270	27455	41043	53487	89851	112661
2007	1683	3860	6443	12334	18499	27824	41632	54110	91745	115135
2008	1666	3686	6078	11930	18596	28070	42138	54695	92447	115220
2009	1596	3203	5317	11346	18739	28659	43243	55554	94008	115996
2010	1581	3126	5145	11035	18432	28197	42346	54549	92360	114087
2011	1546	3019	4961	10565	17949	27311	40976	53007	89389	111225
2012	1454	2758	4565	10014	17351	26305	39416	51518	86040	107555
2013	1405	2528	4115	9285	16890	26004	39221	50777	84102	105918
2014	1425	2571	4170	9219	16790	25983	39086	50661	84804	106766
2015	1469	2752	4489	9591	16946	26274	39810	51511	87404	110410
2016	1526	3018	4916	10188	17299	26619	40242	52024	88405	112082
2017	1660	3261	5278	10592	17313	26461	39806	51533	87280	110445
2018	1714	3485	5661	11040	17599	26642	39717	51450	87574	110321

Notes: CS sample. Each column corresponds to a percentile of the distribution of annual earnings. The sample includes both males and females. Annual earnings are reported in 2018 euros.

Table F4: Descriptive statistics for the CS sample

(a) Panel A: Basic summary statistics

Year	Obs	Mean Income		Females	Age Shares %			Education Shares %			
	(×1000)	Males	Females	% Share	[25,35]	[36,45]	[46,55]	Primary	Lower Sec	Upper Sec	College
2005	415	30064	21363	43.8	44.6	33.3	22.1	14.5	36.5	28.2	20.7
2006	435	30532	21863	44.3	43.6	33.7	22.7	14.3	36.6	27.8	21.3
2007	455	30905	22400	44.8	42.8	34.1	23.1	14.1	36.7	27.5	21.6
2008	463	30786	22788	45.6	41.6	34.6	23.8	13.7	36.6	27.4	22.3
2009	444	30676	23347	46.3	39.7	35.4	25.0	12.7	36.3	27.6	23.5
2010	433	30094	22916	46.8	38.0	36.1	25.9	12.1	36.4	27.5	24.1
2011	425	29280	22132	47.3	36.4	36.9	26.8	11.6	36.4	27.4	24.6
2012	407	28140	21387	47.6	34.5	37.8	27.7	11.1	35.8	27.6	25.6
2013	396	27294	21141	47.5	32.7	38.6	28.7	11.1	35.8	27.2	25.9
2014	397	27087	21209	47.3	31.4	39.0	29.5	11.2	35.9	26.8	26.1
2015	405	27605	21650	47.4	30.3	39.2	30.5	11.2	35.9	26.5	26.4
2016	409	28119	22153	47.4	29.5	39.1	31.4	11.1	35.7	26.3	26.8
2017	415	28153	22174	47.5	29.0	38.7	32.3	11.1	35.6	26.1	27.2
2018	420	28504	22479	47.6	28.7	37.9	33.4	11.0	35.4	26.1	27.6

(b) Panel B: Percentiles of annual earnings

Year	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5
2005	1866	4166	6972	13860	21235	31997	47977	62769	106369	133741
2006	1940	4455	7426	14360	21596	32453	48514	63223	106207	133169
2007	1990	4562	7616	14580	21867	32889	49210	63959	108446	136093
2008	1969	4357	7184	14101	21981	33179	49809	64652	109276	136193
2009	1887	3786	6285	13411	22150	33876	51115	65667	111121	137111
2010	1869	3695	6082	13044	21788	33330	50055	64478	109173	134855
2011	1827	3568	5864	12488	21217	32283	48434	62656	105661	131472
2012	1719	3260	5396	11837	20509	31093	46591	60896	101702	127134
2013	1661	2988	4864	10976	19965	30737	46361	60020	99412	125199
2014	1684	3039	4929	10897	19847	30712	46202	59883	100242	126201
2015	1736	3253	5306	11337	20031	31056	47057	60888	103314	130508
2016	1804	3567	5811	12043	20448	31465	47567	61494	104497	132484
2017	1962	3855	6239	12520	20465	31277	47052	60913	103168	130550
2018	2026	4119	6691	13050	20802	31492	46947	60816	103516	130403

Notes: CS sample. Annual earnings are reported in 2018 U.S. dollars.

Table F5: Descriptive statistics for the LS sample

(a) Panel A: Basic summary statistics

Year	Obs	Mean Income		Females	Age Shares %			Education Shares %			
	(×1000)	Males	Females	% Share	[25,35]	[36,45]	[46,55]	Primary	Lower Sec	Upper Sec	College
2005	283	26408	19403	44.1	47.9	37.5	14.6	10.9	35.7	29.9	23.6
2006	289	27025	19756	45.1	47.2	37.9	14.9	10.3	35.8	29.5	24.3
2007	292	28027	20640	46.1	46.4	38.5	15.1	9.9	35.3	29.5	25.4
2008	286	28430	21268	46.9	44.9	39.5	15.6	9.4	35.1	29.3	26.2
2009	281	27769	21497	47.3	43.1	40.6	16.3	9.1	35.2	29.0	26.7
2010	281	26803	20921	47.6	41.6	41.5	16.9	9.0	35.3	28.7	27.0
2011	279	26119	20296	48.0	39.8	42.7	17.4	8.8	35.0	28.5	27.6
2012	273	24826	19409	48.2	37.7	44.1	18.2	8.7	34.6	28.4	28.3
2013	273	23763	18935	48.0	36.0	45.2	18.9	8.9	34.7	27.9	28.5

(b) Panel B: Percentiles of annual earnings

Year	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5
2005	1946	4803	7606	13397	19022	28316	41405	53409	88304	109627
2006	2078	5194	8068	13776	19448	28718	41791	53725	88043	109562
2007	2215	5659	8626	14295	20114	29790	43246	55692	91572	113762
2008	2294	5705	8632	14403	20578	30474	44132	56604	92450	114754
2009	2002	4730	7645	13880	20535	30581	44361	56252	92750	113096
2010	1915	4457	7195	13338	19980	29763	43040	54680	90383	109702
2011	1889	4365	6955	12897	19469	28932	41687	53160	87399	106747
2012	1692	3805	6259	12109	18579	27577	39880	51314	83540	102653
2013	1555	3283	5435	11096	17968	27030	39346	50264	81190	99764

Notes: LS sample, restricted to non-missing 1-year and 5-year changes in log earnings. Annual earnings are reported in 2018 euros.

Table F6: Descriptive statistics for the LS sample

(a) Panel A: Basic summary statistics

Year	Obs	Mean Income		Females	Age Shares %			Education Shares %			
	(×1000)	Males	Females	% Share	[25,35]	[36,45]	[46,55]	Primary	Lower Sec	Upper Sec	College
2005	283	31215	22935	44.1	47.9	37.5	14.6	10.9	35.7	29.9	23.6
2006	289	31945	23353	45.1	47.2	37.9	14.9	10.3	35.8	29.5	24.3
2007	292	33129	24397	46.1	46.4	38.5	15.1	9.9	35.3	29.5	25.4
2008	286	33606	25140	46.9	44.9	39.5	15.6	9.4	35.1	29.3	26.2
2009	281	32824	25411	47.3	43.1	40.6	16.3	9.1	35.2	29.0	26.7
2010	281	31682	24729	47.6	41.6	41.5	16.9	9.0	35.3	28.7	27.0
2011	279	30873	23991	48.0	39.8	42.7	17.4	8.8	35.0	28.5	27.6
2012	273	29345	22943	48.2	37.7	44.1	18.2	8.7	34.6	28.4	28.3
2013	273	28088	22382	48.0	36.0	45.2	18.9	8.9	34.7	27.9	28.5

(b) Panel B: Percentiles of annual earnings

Year	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5
2005	2300	5677	8991	15835	22485	33471	48942	63131	104378	129582
2006	2456	6139	9537	16284	22988	33946	49398	63505	104070	129506
2007	2618	6690	10197	16898	23775	35213	51118	65830	108241	134470
2008	2711	6743	10203	17024	24324	36021	52166	66908	109279	135642
2009	2366	5591	9037	16407	24273	36148	52437	66492	109634	133683
2010	2263	5269	8505	15766	23617	35180	50875	64633	106835	129671
2011	2233	5160	8221	15245	23013	34199	49275	62837	103309	126178
2012	2000	4498	7399	14314	21961	32597	47140	60655	98747	121339
2013	1838	3881	6424	13116	21239	31951	46508	59414	95969	117924

Notes: LS sample, restricted to non-missing 1-year and 5-year changes in log earnings. Annual earnings are reported in 2018 U.S. dollars.

Table F7: Descriptive statistics for the H sample

(a) Panel A: Basic summary statistics

Year	Obs	Mean Income		Females	Age Shares %			Education Shares %			
	(×1000)	Males	Females	% Share	[25,35]	[36,45]	[46,55]	Primary	Lower Sec	Upper Sec	College
2008	240	30037	22843	45.8	40.1	42.8	17.2	8.8	34.8	30.4	26.0
2009	242	29241	22912	46.5	38.7	43.6	17.7	8.6	34.7	30.0	26.7
2010	245	28248	22203	47.2	37.4	44.4	18.2	8.4	34.8	29.6	27.1
2011	244	27624	21558	47.7	35.6	45.5	18.9	8.2	34.5	29.5	27.8
2012	239	26153	20444	48.1	33.6	46.7	19.6	8.1	34.3	29.2	28.4
2013	238	25401	20148	48.1	31.9	47.8	20.3	8.0	34.2	28.9	28.9

(b) Panel B: Percentiles of annual earnings

Year	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5
2008	3060	7238	10291	15711	21969	32328	46233	59241	96067	118977
2009	2406	6016	9087	15206	21817	32271	46126	58584	95444	116255
2010	2306	5624	8602	14706	21184	31277	44634	56748	92561	112395
2011	2305	5600	8416	14324	20668	30384	43227	55170	89604	109596
2012	2035	4941	7677	13514	19589	28698	41087	52703	85318	104343
2013	1876	4403	7036	12772	19143	28426	40805	51963	83040	103001

Notes: H sample, restricted to non-missing 1-year and 5-year changes in log earnings. Annual earnings are reported in 2018 euros.

Table F8: Descriptive statistics for the H sample

(a) Panel A: Basic summary statistics

Year	Obs	Mean Income		Females	Age Shares %			Education Shares %			
	(×1000)	Males	Females	% Share	[25,35]	[36,45]	[46,55]	Primary	Lower Sec	Upper Sec	College
2008	240	35505	27001	45.8	40.1	42.8	17.2	8.8	34.8	30.4	26.0
2009	242	34563	27082	46.5	38.7	43.6	17.7	8.6	34.7	30.0	26.7
2010	245	33390	26244	47.2	37.4	44.4	18.2	8.4	34.8	29.6	27.1
2011	244	32652	25483	47.7	35.6	45.5	18.9	8.2	34.5	29.5	27.8
2012	239	30914	24165	48.1	33.6	46.7	19.6	8.1	34.3	29.2	28.4
2013	238	30025	23815	48.1	31.9	47.8	20.3	8.0	34.2	28.9	28.9

(b) Panel B: Percentiles of annual earnings

Year	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5
2008	3617	8556	12165	18571	25968	38213	54649	70024	113555	140634
2009	2844	7111	10741	17974	25788	38146	54522	69248	112818	137418
2010	2726	6648	10168	17382	25040	36971	52759	67078	109410	132854
2011	2724	6619	9948	16931	24430	35915	51096	65213	105915	129546
2012	2406	5840	9075	15974	23155	33922	48566	62297	100849	123337
2013	2217	5204	8317	15096	22628	33601	48233	61422	98156	121750

Notes: H sample, restricted to non-missing 1-year and 5-year changes in log earnings. Annual earnings are reported in 2018 U.S. dollars.

Table F9: Descriptive statistics for the B sample

(a) Panel A: Basic summary statistics

Year	Obs	Mean Income	Age Shares %			Education Shares %			
	(×1000)	Males	[25,35]	[36,45]	[46,55]	Primary	Lower Sec	Upper Sec	College
2006	223.091	27167	41.0	34.5	24.6	17.7	40.0	26.5	15.8
2007	233.686	27183	40.3	34.6	25.1	17.3	40.0	26.3	16.3
2008	244.511	26556	40.0	34.7	25.3	17.3	40.1	26.0	16.6
2009	249.983	25409	38.9	35.1	26.0	17.0	40.0	25.9	17.0
2010	254.180	24098	37.7	35.5	26.8	16.8	40.2	25.7	17.3
2011	253.227	22948	36.2	36.1	27.7	16.5	40.5	25.5	17.6
2012	250.995	21210	34.4	36.9	28.7	16.3	40.6	25.3	17.8
2013	247.411	20496	32.4	37.6	29.9	16.2	40.7	25.2	17.9
2014	242.084	20725	30.6	38.3	31.1	16.1	40.7	25.0	18.2
2015	235.009	21813	29.1	38.7	32.2	15.8	40.8	24.9	18.5
2016	229.111	22763	27.7	39.1	33.2	15.4	40.9	24.8	18.9
2017	225.461	23245	26.8	39.0	34.2	15.0	40.8	24.8	19.4
2018	222.442	23926	26.4	38.7	35.0	14.7	40.7	24.8	19.7

(b) Panel B: Percentiles of annual earnings

Year	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5
2006	0	5370	10109	15856	21177	31976	48039	63411	110742	142959
2007	0	3043	9165	15770	21308	32173	48351	64156	112581	144502
2008	0	1631	7630	15282	21049	31755	47730	62871	110914	140991
2009	0	0	5220	13535	20318	31042	47179	62099	109646	138144
2010	0	0	4023	11790	19478	29882	45415	59767	106035	135261
2011	0	0	2396	10330	18760	28810	43815	57951	102381	130759
2012	0	0	799	8301	17607	27080	41273	54954	96155	122129
2013	0	0	328	6943	16917	26611	41087	54167	93873	119257
2014	0	0	689	7238	17104	26881	41376	54529	95231	121736
2015	0	0	1705	8743	17827	27686	42557	56121	99167	127645
2016	0	0	2548	10401	18564	28495	43530	57297	102118	131781
2017	0	49	3666	11802	18872	28702	43372	56937	101672	130799
2018	0	697	4768	12903	19373	29075	43536	57117	102123	132345

Notes: B sample. Annual earnings are reported in 2018 euros. Only males.

Table F10: Descriptive statistics for the B sample

(a) Panel A: Basic summary statistics

Year	Obs	Mean Income	Age Shares %			Education Shares %			
	(×1000)	Males	[25,35]	[36,45]	[46,55]	Primary	Lower Sec	Upper Sec	College
2006	223.091	32112	41.0	34.5	24.6	17.7	40.0	26.5	15.8
2007	233.686	32131	40.3	34.6	25.1	17.3	40.0	26.3	16.3
2008	244.511	31391	40.0	34.7	25.3	17.3	40.1	26.0	16.6
2009	249.983	30035	38.9	35.1	26.0	17.0	40.0	25.9	17.0
2010	254.180	28484	37.7	35.5	26.8	16.8	40.2	25.7	17.3
2011	253.227	27125	36.2	36.1	27.7	16.5	40.5	25.5	17.6
2012	250.995	25070	34.4	36.9	28.7	16.3	40.6	25.3	17.8
2013	247.411	24227	32.4	37.6	29.9	16.2	40.7	25.2	17.9
2014	242.084	24497	30.6	38.3	31.1	16.1	40.7	25.0	18.2
2015	235.009	25784	29.1	38.7	32.2	15.8	40.8	24.9	18.5
2016	229.111	26906	27.7	39.1	33.2	15.4	40.9	24.8	18.9
2017	225.461	27476	26.8	39.0	34.2	15.0	40.8	24.8	19.4
2018	222.442	28281	26.4	38.7	35.0	14.7	40.7	24.8	19.7

(b) Panel B: Percentiles of annual earnings

Year	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5
2006	0	6348	11950	18743	25032	37797	56784	74954	130901	168982
2007	0	3597	10834	18641	25187	38030	57152	75835	133075	170806
2008	0	1927	9019	18064	24881	37535	56419	74316	131104	166656
2009	0	0	6171	15999	24017	36692	55767	73403	129605	163291
2010	0	0	4756	13936	23024	35321	53682	70646	125337	159883
2011	0	0	2832	12210	22175	34055	51790	68500	121018	154562
2012	0	0	945	9812	20811	32009	48786	64957	113658	144361
2013	0	0	387	8207	19997	31455	48566	64027	110961	140965
2014	0	0	814	8556	20218	31774	48908	64455	112566	143896
2015	0	0	2016	10334	21073	32726	50304	66336	117219	150881
2016	0	0	3012	12294	21943	33682	51453	67728	120707	155770
2017	0	58	4333	13951	22307	33927	51268	67301	120179	154609
2018	0	824	5636	15252	22900	34367	51461	67515	120712	156436

Notes: B sample. Annual earnings are reported in 2018 U.S. dollars. Only males.

Table F11: Descriptive statistics for the B sample, with positive income

(a) Panel A: Basic summary statistics

Year	Obs	Mean Income	Age Shares %			Education Shares %			
	(×1000)	Males	[25,35]	[36,45]	[46,55]	Primary	Lower Sec	Upper Sec	College
2006	218.712	27711	40.9	34.5	24.6	17.6	40.0	26.6	15.8
2007	226.159	28087	40.2	34.7	25.1	17.1	40.2	26.4	16.4
2008	235.093	27620	39.8	34.8	25.4	17.0	40.2	26.1	16.7
2009	236.784	26826	38.4	35.3	26.3	16.5	40.1	26.1	17.3
2010	239.605	25564	37.2	35.7	27.1	16.3	40.4	25.8	17.6
2011	236.440	24577	35.6	36.2	28.2	15.8	40.6	25.7	17.9
2012	229.793	23167	33.5	36.8	29.6	15.4	40.5	25.8	18.3
2013	224.752	22563	31.6	37.5	30.9	15.1	40.7	25.7	18.5
2014	221.923	22607	30.1	38.2	31.7	15.2	40.8	25.4	18.7
2015	219.181	23388	28.7	38.8	32.4	15.0	40.9	25.2	18.9
2016	215.660	24183	27.5	39.2	33.2	14.7	40.9	25.1	19.2
2017	214.367	24448	26.7	39.1	34.2	14.5	40.8	25.1	19.6
2018	213.275	24954	26.3	38.8	34.9	14.2	40.7	25.1	20.0

(b) Panel B: Percentiles of annual earnings

Year	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5
2006	1987	7403	11279	16162	21459	32301	48462	63857	111662	143935
2007	1644	7194	11285	16297	21779	32723	49028	64957	114061	145903
2008	1363	6242	10351	16010	21597	32416	48457	63754	112489	142797
2009	1047	5228	8377	14860	21097	31980	48220	63369	111545	140805
2010	1026	4402	6465	13631	20268	30855	46441	61151	107852	138024
2011	621	3280	5458	12701	19661	29917	45042	59569	104753	134212
2012	444	2664	5136	11486	18737	28388	42760	56862	99211	126801
2013	359	2335	4627	10309	18208	28116	42636	56154	97214	123367
2014	344	2184	4422	10157	18277	28228	42866	56400	98274	125706
2015	433	2466	4730	11140	18738	28809	43781	57660	101636	131751
2016	500	2852	5299	12401	19347	29465	44567	58613	104174	134866
2017	633	3445	6169	13222	19507	29513	44273	58038	103296	132568
2018	773	4052	7240	13949	19895	29760	44257	58048	103687	134422

Notes: B sample, positive income. Annual earnings are reported in 2018 euros. Only males.

Table F12: Descriptive statistics for the B sample, with positive income

(a) Panel A: Basic summary statistics

Year	Obs	Mean Income	Age Shares %			Education Shares %			
	(×1000)	Males	[25,35]	[36,45]	[46,55]	Primary	Lower Sec	Upper Sec	College
2006	218.712	32755	40.9	34.5	24.6	17.6	40.0	26.6	15.8
2007	226.159	33200	40.2	34.7	25.1	17.1	40.2	26.4	16.4
2008	235.093	32648	39.8	34.8	25.4	17.0	40.2	26.1	16.7
2009	236.784	31709	38.4	35.3	26.3	16.5	40.1	26.1	17.3
2010	239.605	30217	37.2	35.7	27.1	16.3	40.4	25.8	17.6
2011	236.440	29051	35.6	36.2	28.2	15.8	40.6	25.7	17.9
2012	229.793	27384	33.5	36.8	29.6	15.4	40.5	25.8	18.3
2013	224.752	26670	31.6	37.5	30.9	15.1	40.7	25.7	18.5
2014	221.923	26723	30.1	38.2	31.7	15.2	40.8	25.4	18.7
2015	219.181	27646	28.7	38.8	32.4	15.0	40.9	25.2	18.9
2016	215.660	28585	27.5	39.2	33.2	14.7	40.9	25.1	19.2
2017	214.367	28898	26.7	39.1	34.2	14.5	40.8	25.1	19.6
2018	213.275	29496	26.3	38.8	34.9	14.2	40.7	25.1	20.0

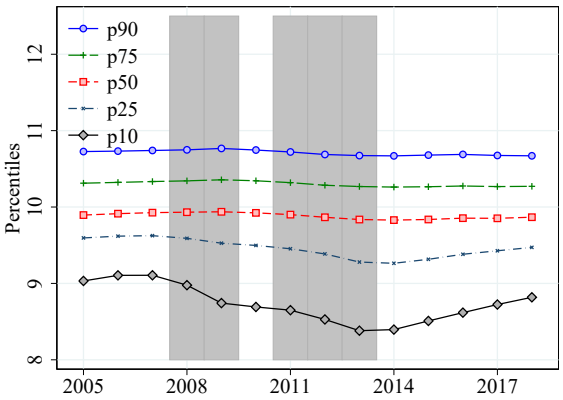
(b) Panel B: Percentiles of annual earnings

Year	P1	P5	P10	P25	P50	P75	P90	P95	P99	P99.5
2006	2348	8750	13332	19104	25365	38181	57284	75481	131989	170135
2007	1944	8503	13339	19264	25743	38680	57953	76781	134824	172463
2008	1612	7378	12236	18924	25529	38316	57277	75359	132966	168791
2009	1238	6180	9902	17565	24937	37801	56998	74904	131850	166437
2010	1213	5203	7642	16112	23957	36472	54895	72283	127485	163148
2011	734	3877	6452	15013	23240	35363	53241	70412	123822	158644
2012	525	3149	6071	13576	22147	33555	50543	67213	117270	149883
2013	424	2760	5469	12186	21523	33234	50397	66376	114910	145824
2014	407	2582	5227	12006	21604	33366	50670	66667	116163	148589
2015	512	2914	5591	13168	22149	34053	51750	68155	120138	155734
2016	592	3371	6264	14658	22869	34829	52679	69282	123137	159416
2017	748	4072	7292	15629	23058	34886	52332	68602	122099	156700
2018	914	4790	8558	16488	23517	35177	52313	68614	122561	158891

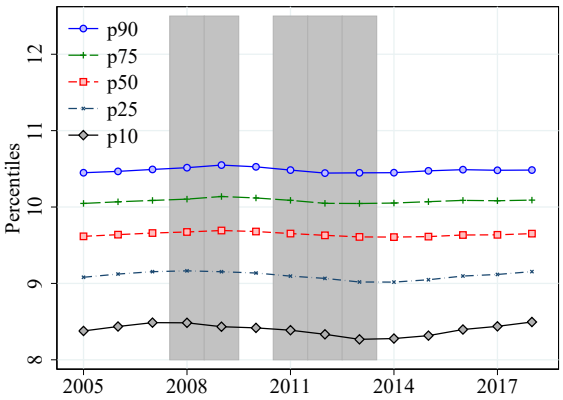
Notes: B sample, positive income. Annual earnings are reported in 2018 U.S. dollars. Only males.

Figure F1: Percentiles of the distribution of log annual earnings, no normalization

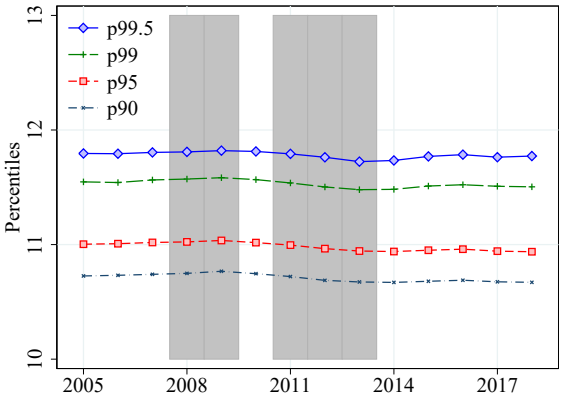
(a) Overall distribution: Males



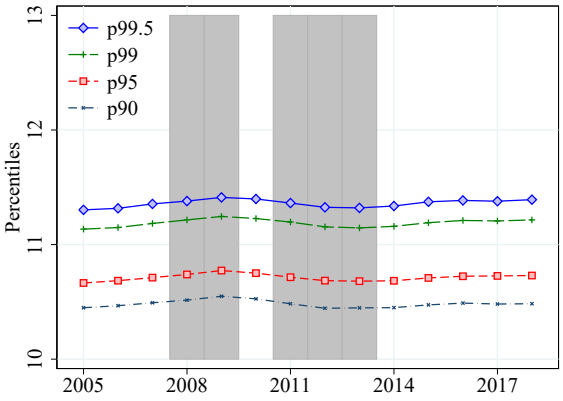
(b) Overall distribution: Females



(c) Top percentiles: Males



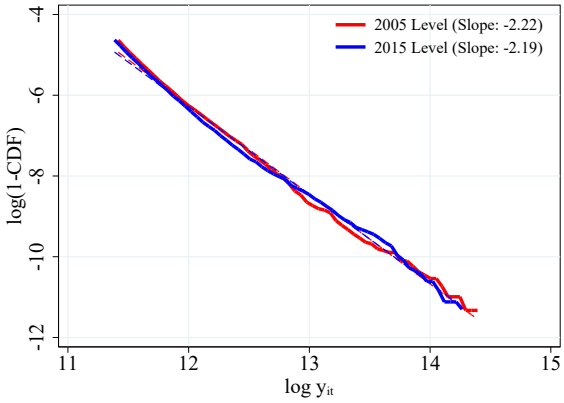
(d) Top percentiles: Females



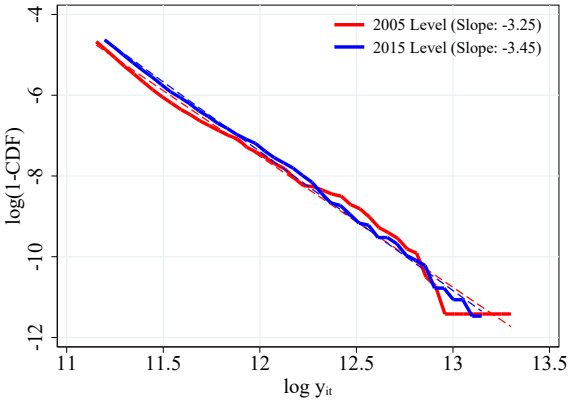
Notes: CS sample, percentiles of log annual earnings, by gender. The shaded areas indicate recession years.

Figure F2: Top income inequality: Pareto tail at top 1%

(a) Males



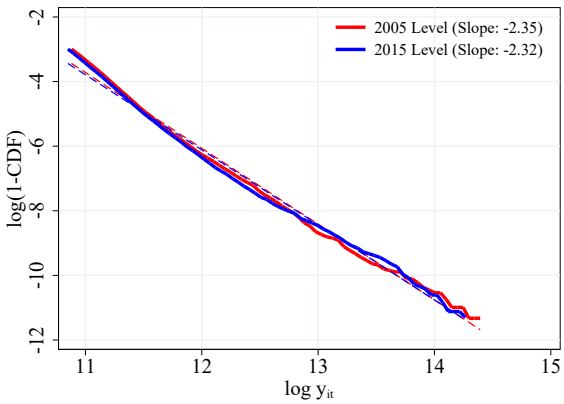
(b) Females



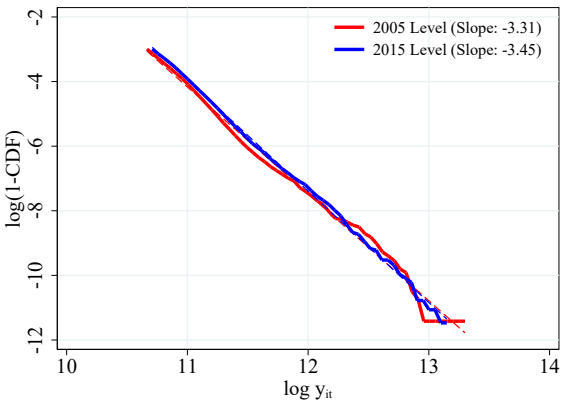
Notes: CS sample.

Figure F3: Top income inequality: Pareto tail at top 5%

(a) Males

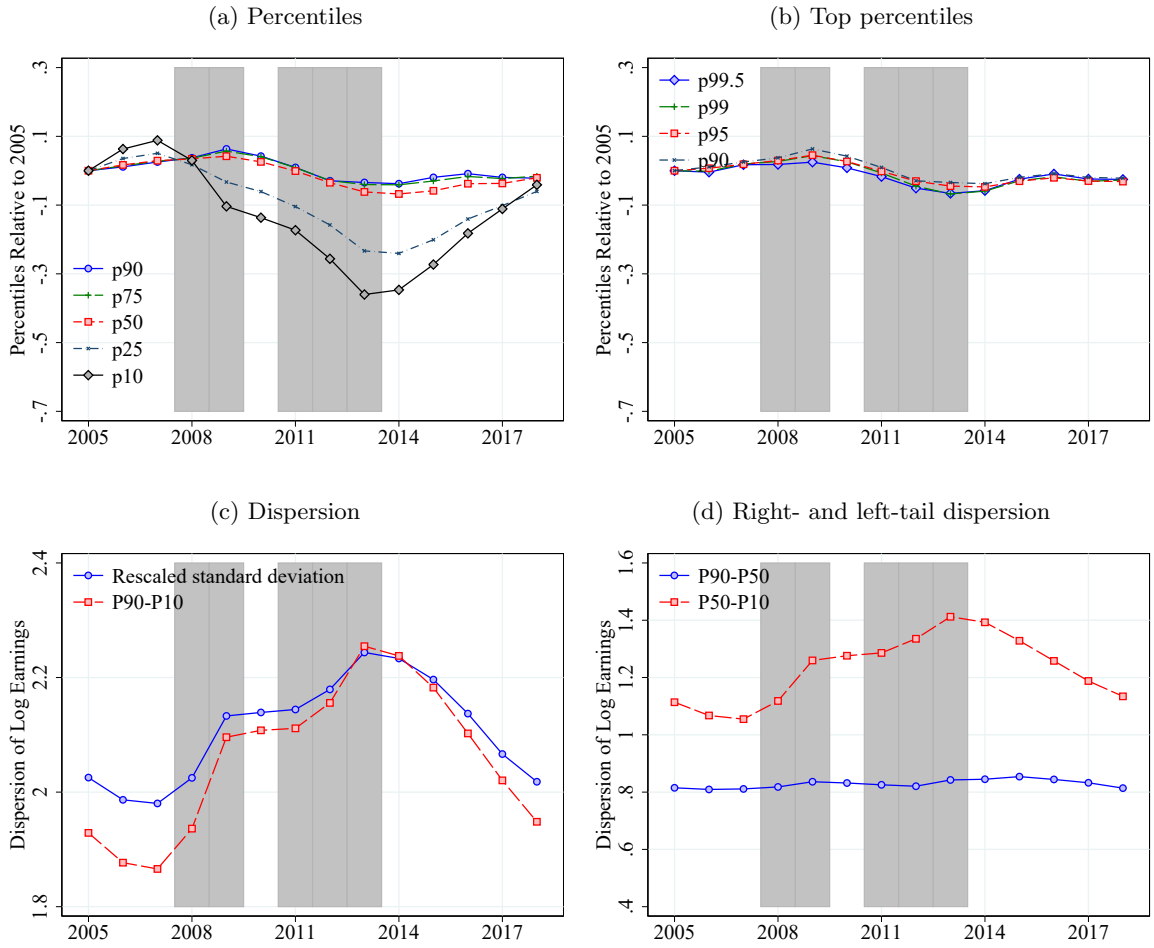


(b) Females



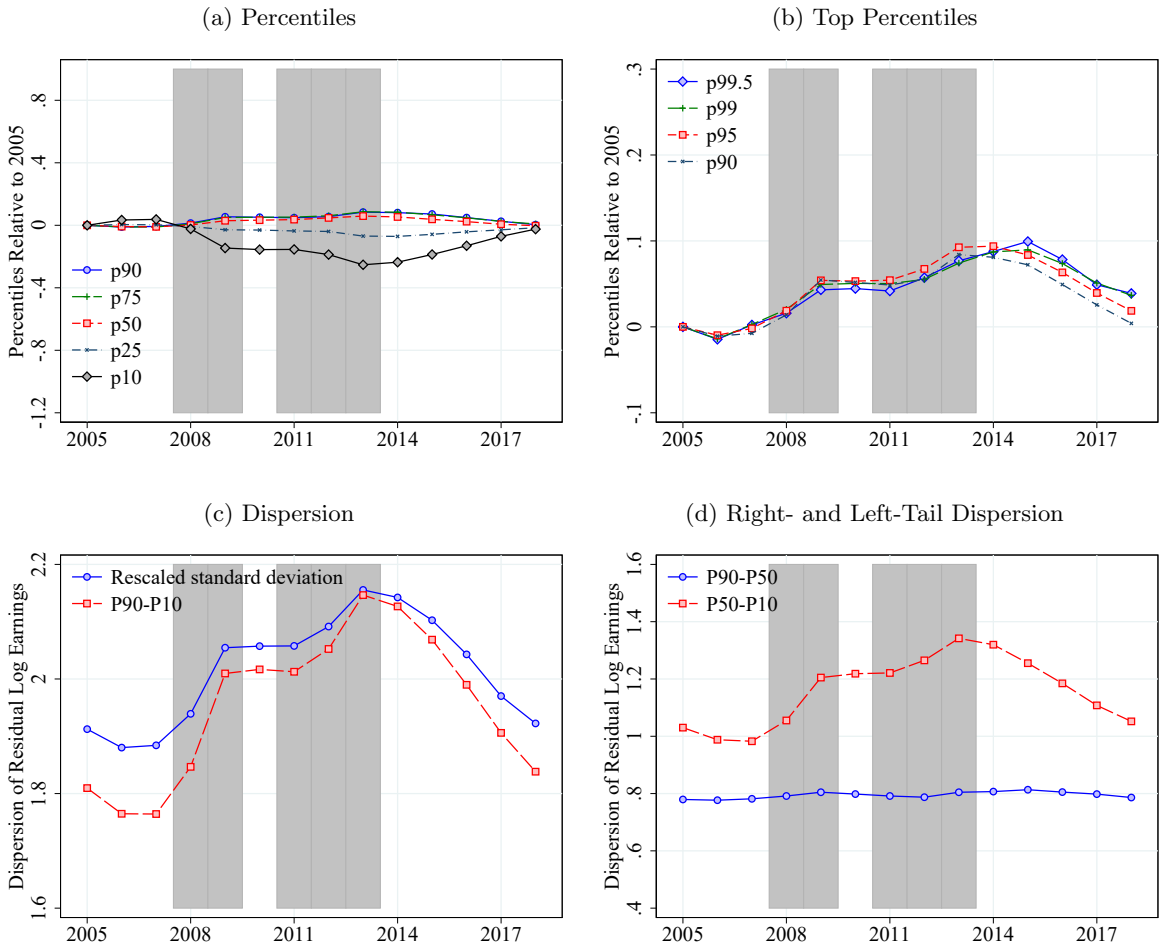
Notes: CS sample.

Figure F4: Distribution of earnings in the population (males & females)



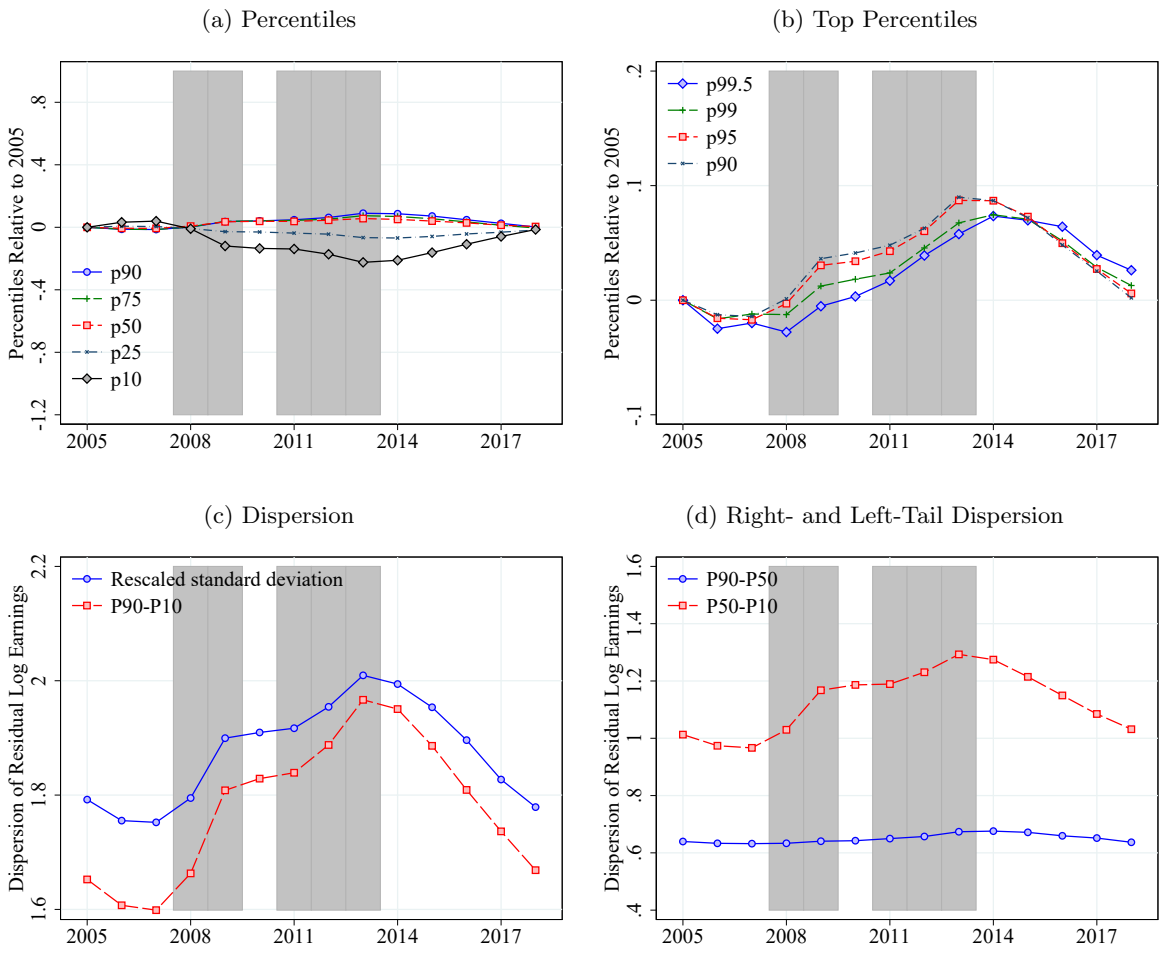
Notes: CS sample, percentiles of log annual earnings. In the top graphs, all percentiles are normalized to 0 in 2005. The shaded areas indicate recession years.

Figure F5: Distribution of residual earnings in the population (males & females) after controlling for age



Notes: CS sample, percentiles of residualized log annual earnings, after controlling for age. All percentiles are normalized to 0 in 2005. The shaded areas indicate recession years.

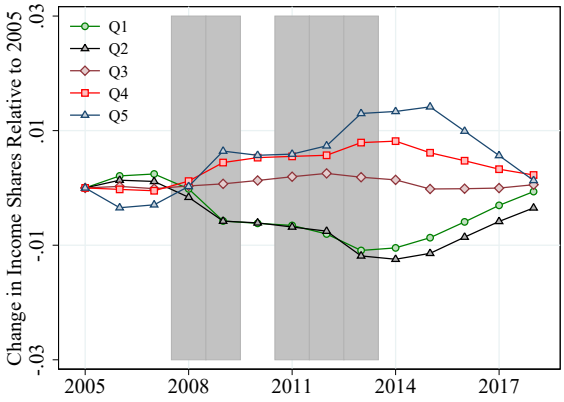
Figure F6: Distribution of residual earnings in the population (males & females) after controlling for age and education



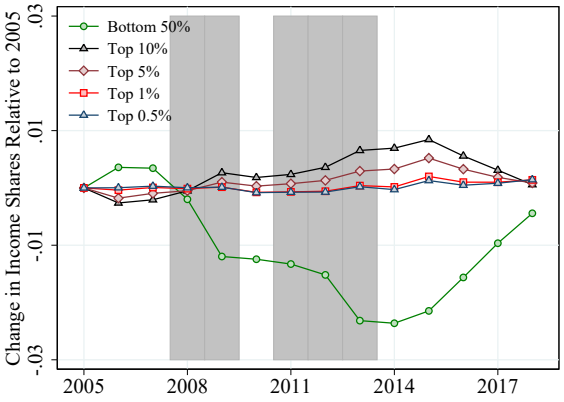
Notes: CS sample, percentiles of residualized log annual earnings, after controlling for age and education. All percentiles are normalized to 0 in 2005. The shaded areas indicate recession years.

Figure F7: Changes in income shares relative to 2005

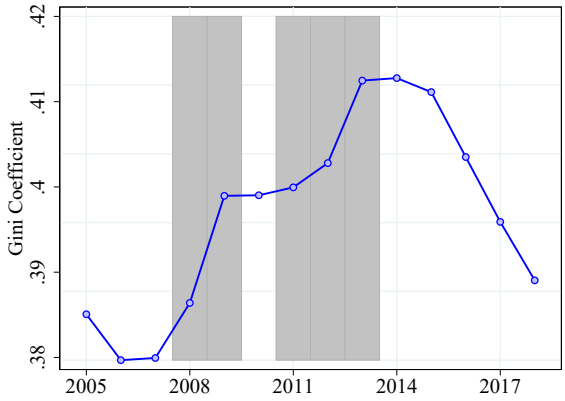
(a) Income Shares of Quintiles



(b) Selected Income Shares



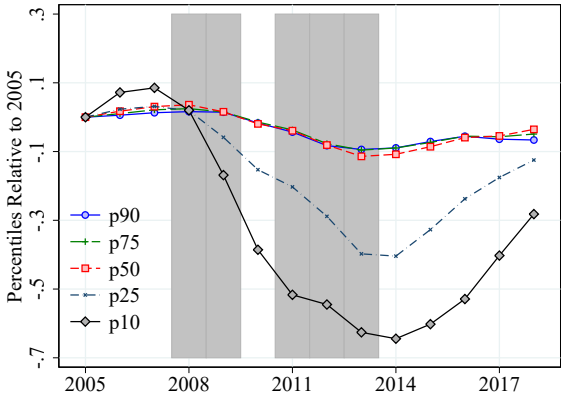
(c) Gini Coefficient



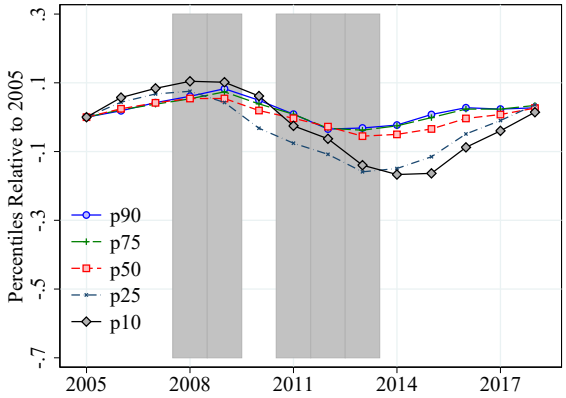
Notes: CS sample.

Figure F8: Evolution of income percentiles and inequality, earnings & unemployment benefits

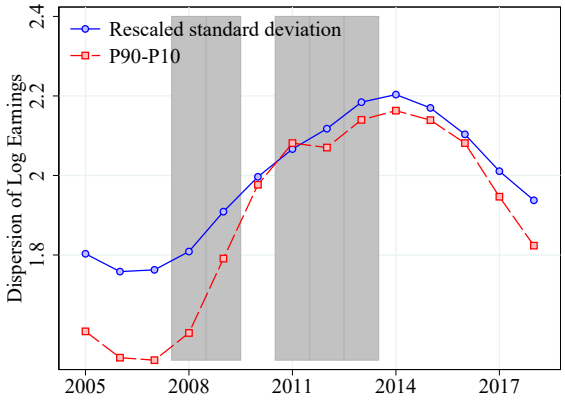
(a) Overall distribution: Males



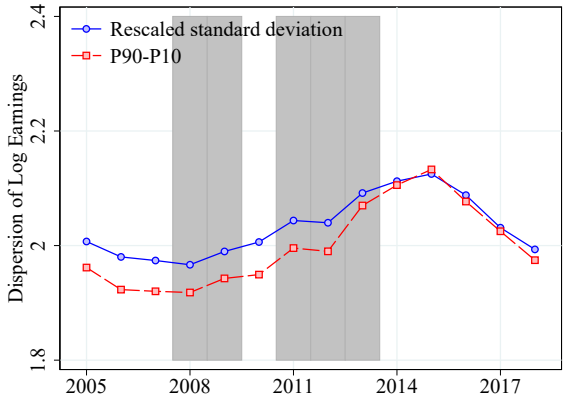
(b) Overall distribution: Females



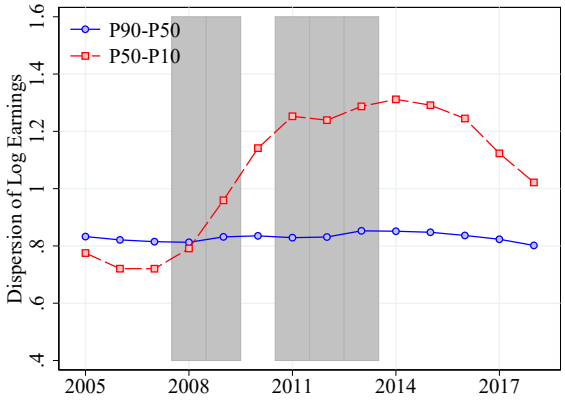
(c) Inequality: Males



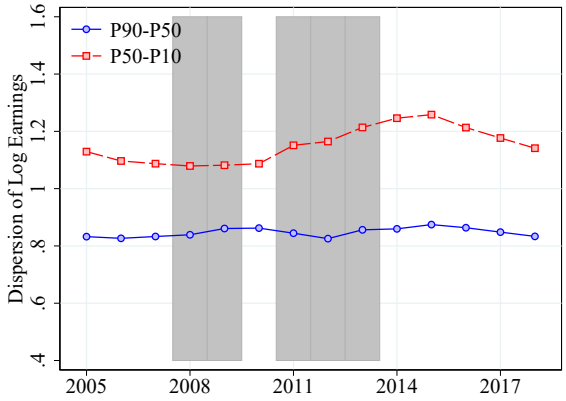
(d) Inequality: Females



(e) Upper & lower inequality: Males



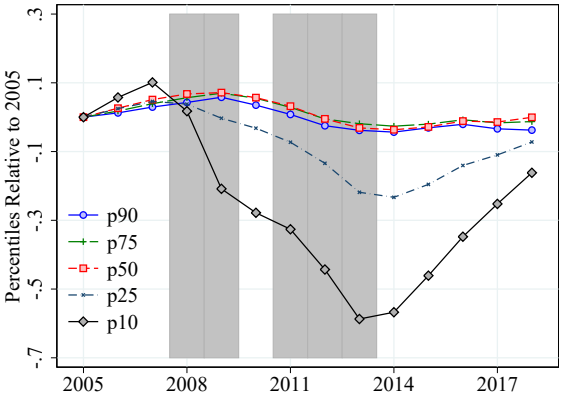
(f) Upper & lower inequality: Females



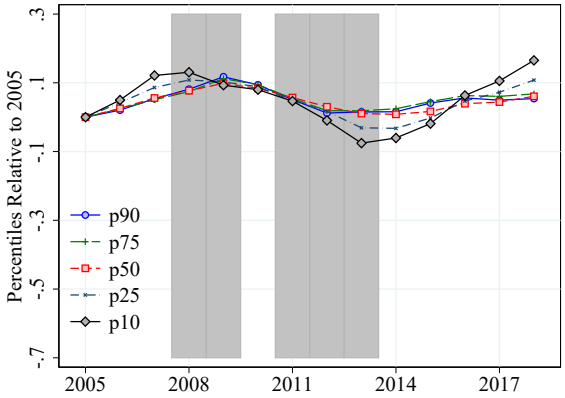
Notes: CS sample, earnings and unemployment benefits. Top panel: percentiles of log annual income, by gender. All percentiles are normalized to 0 in 2005. Middle and bottom panels: P90-P10 difference and rescaled standard deviation, and P90-P50 and P50-P10 percentile differences. The shaded areas indicate recession years.

Figure F9: Evolution of earnings percentiles and inequality, no immigrants

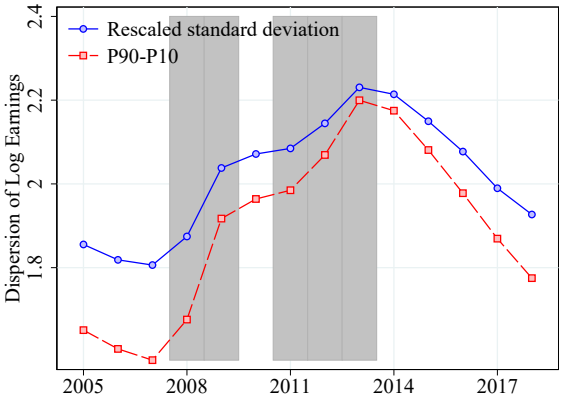
(a) Overall distribution: Males



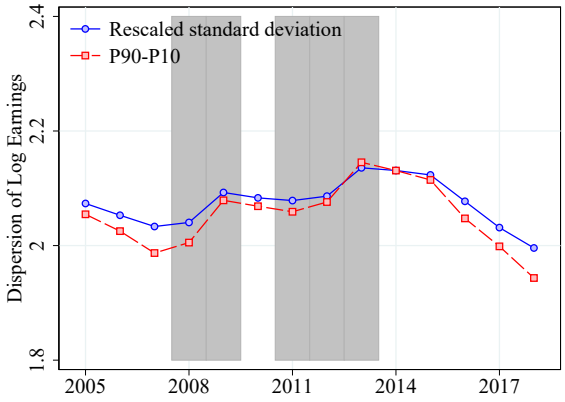
(b) Overall distribution: Females



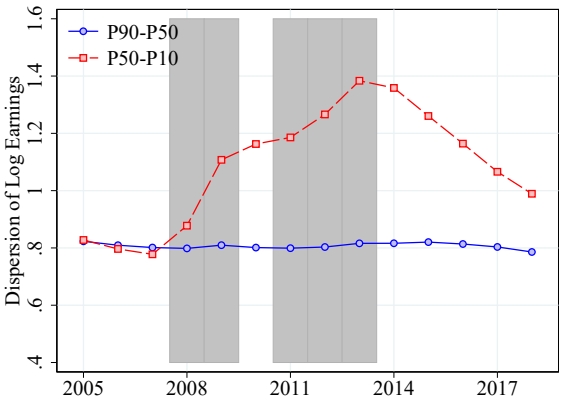
(c) Inequality: Males



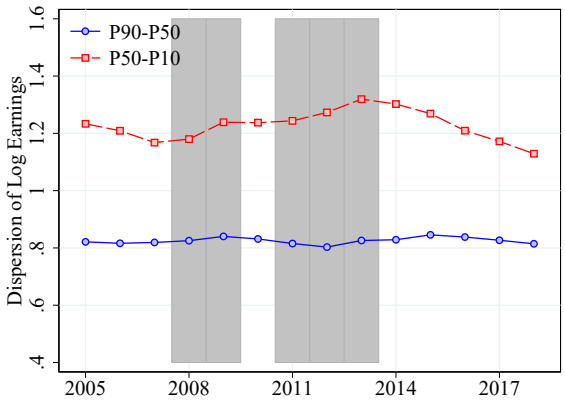
(d) Inequality: Females



(e) Upper & lower inequality: Males

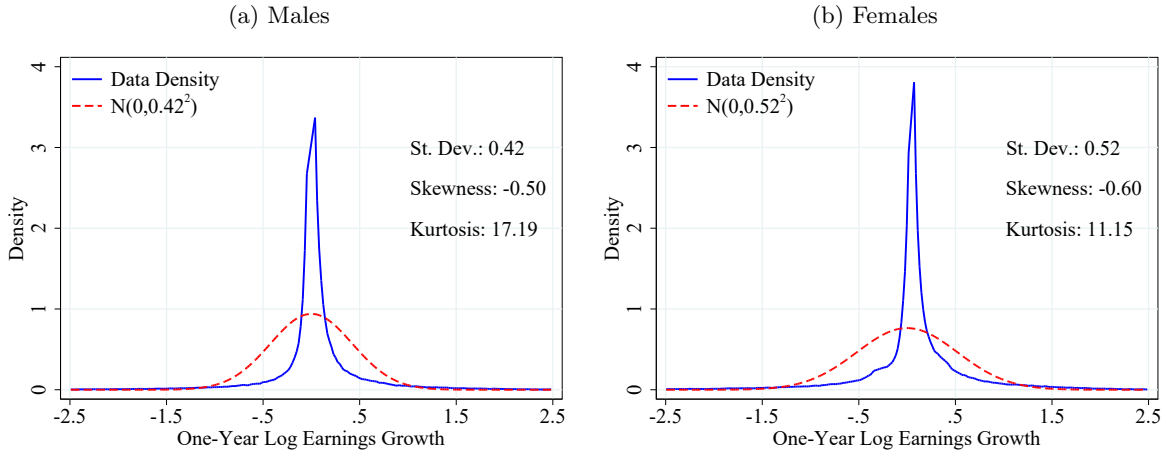


(f) Upper & lower inequality: Females



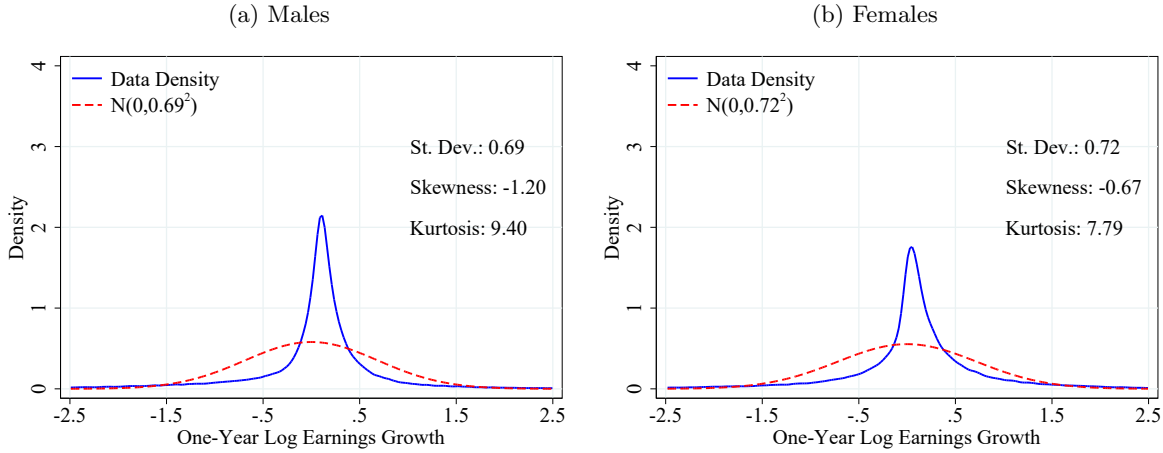
Notes: CS sample, no immigrants. Top panel: percentiles of log annual earnings, by gender. All percentiles are normalized to 0 in 2005. Middle and bottom panels: P90-P10 difference and rescaled standard deviation, and P90-P50 and P50-P10 percentile differences. The shaded areas indicate recession years.

Figure F10: Empirical densities of one-year log earnings changes



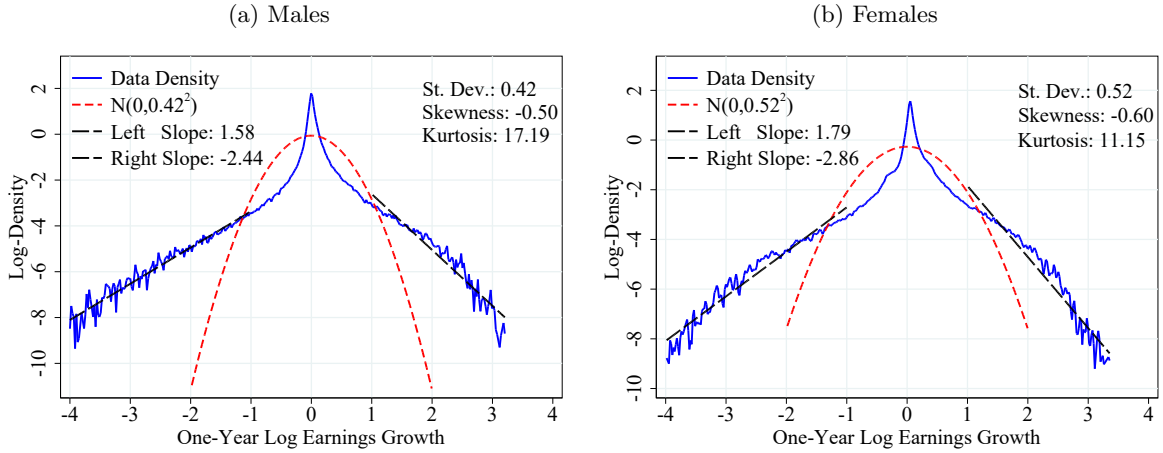
Notes: LS sample, one-year changes in log residual annual earnings. For 2005.

Figure F11: Empirical densities of five-year log earnings changes



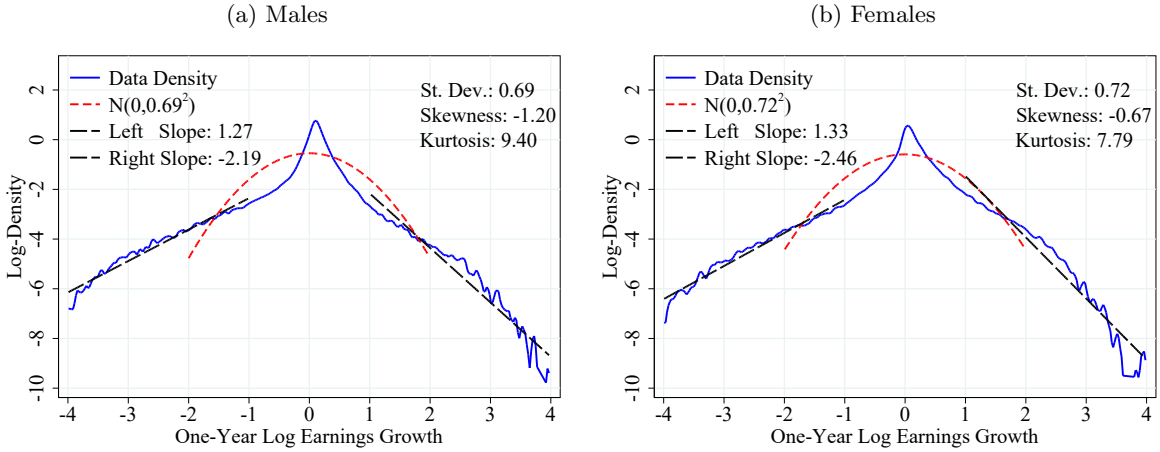
Notes: LS sample, five-year changes in log residual annual earnings. For 2005.

Figure F12: Empirical log densities of one-year log earnings changes



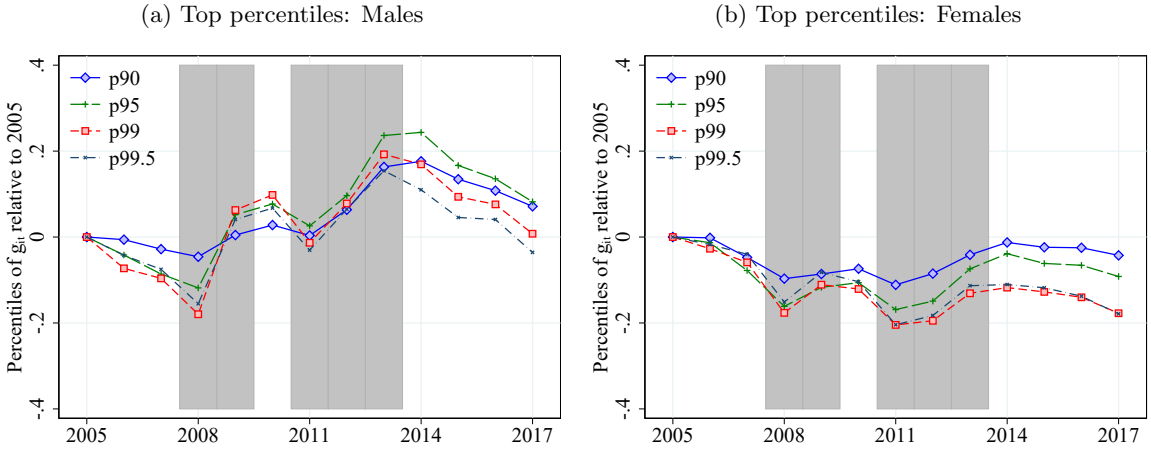
Notes: LS sample, one-year changes in log residual annual earnings. For 2005.

Figure F13: Empirical log densities of five-year log earnings changes



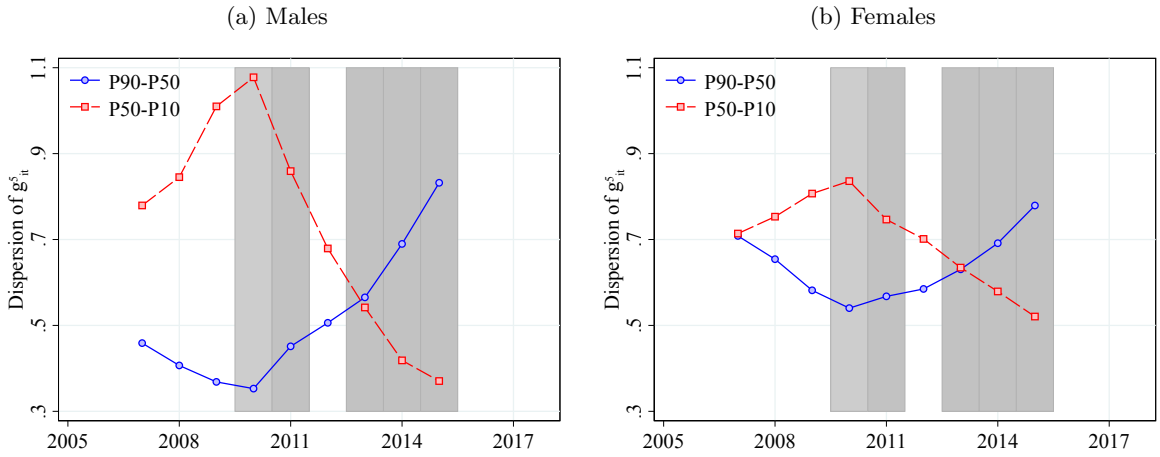
Notes: LS sample, five-year changes in log residual annual earnings. For 2005.

Figure F14: Top percentiles of one-year changes in log earnings



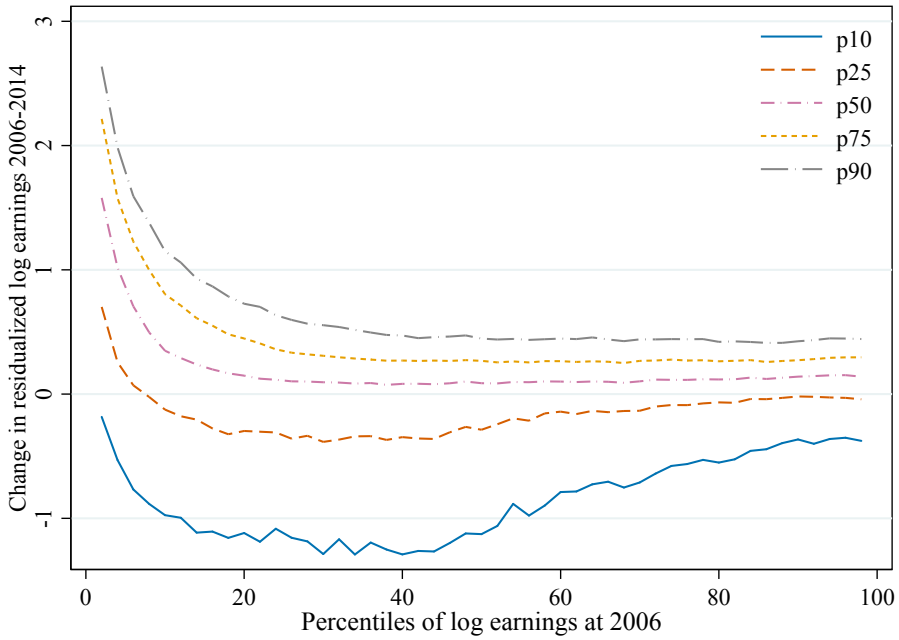
Notes: LS sample, one-year changes in residualized log earnings. All percentiles are normalized to 0 in 2005. The shaded areas indicate recession years.

Figure F15: Dispersion of five-year earnings changes



Notes: LS sample, five-year changes in residualized log earnings. The shaded areas indicate recession years.

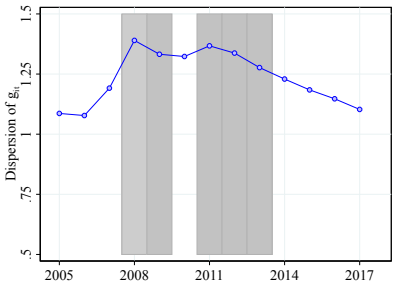
Figure F16: Log earnings changes between 2006 and 2014 against initial earnings



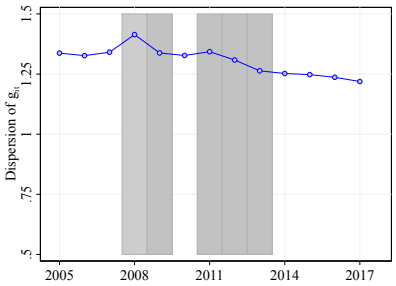
Notes: Non-missing observations on residualized log earnings in 2006 and 2014. On the x-axis we report percentiles of log earnings in 2006. On the y-axis we report the changes in residualized log earnings between 2006 and 2014.

Figure F17: Skewness and kurtosis of one-year log earnings changes

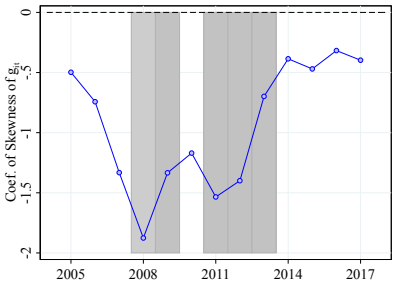
(a) Standard Deviation: Males



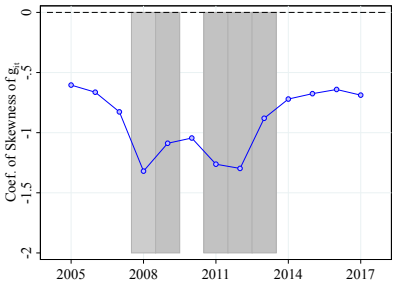
(b) Standard Deviation: Females



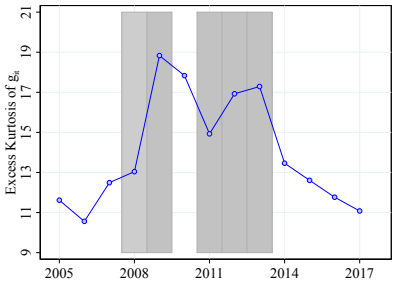
(c) Coef. of Skewness: Males



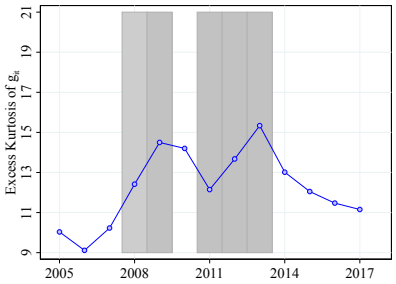
(d) Coef. of Skewness: Females



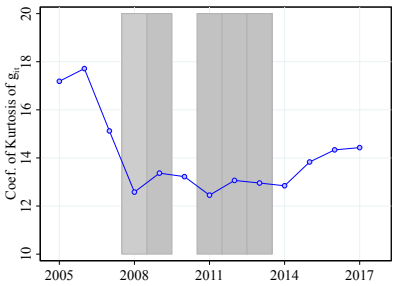
(e) Excess Crow-Siddiqui Kurtosis: Males



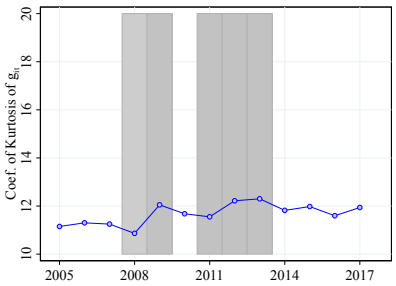
(f) Excess Crow-Siddiqui Kurtosis: Females



(g) Coef. of Kurtosis: Males



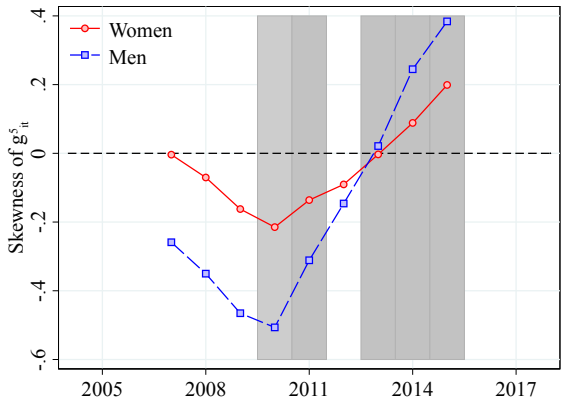
(h) Coef. of Kurtosis: Females



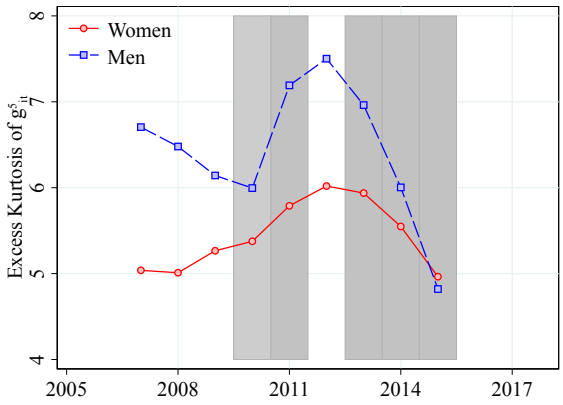
Notes: LS sample, one-year changes in residualized log earnings. The standard deviation is rescaled using a scaling factor of 2.56. Excess Crow-Siddiqui kurtosis is defined as $\frac{P97.5-P2.5}{P75-P25} - 2.91$*, so that it would be zero for the normal distribution. The shaded areas indicate recession years.*

Figure F18: Skewness and kurtosis of five-year earnings changes

(a) Skewness

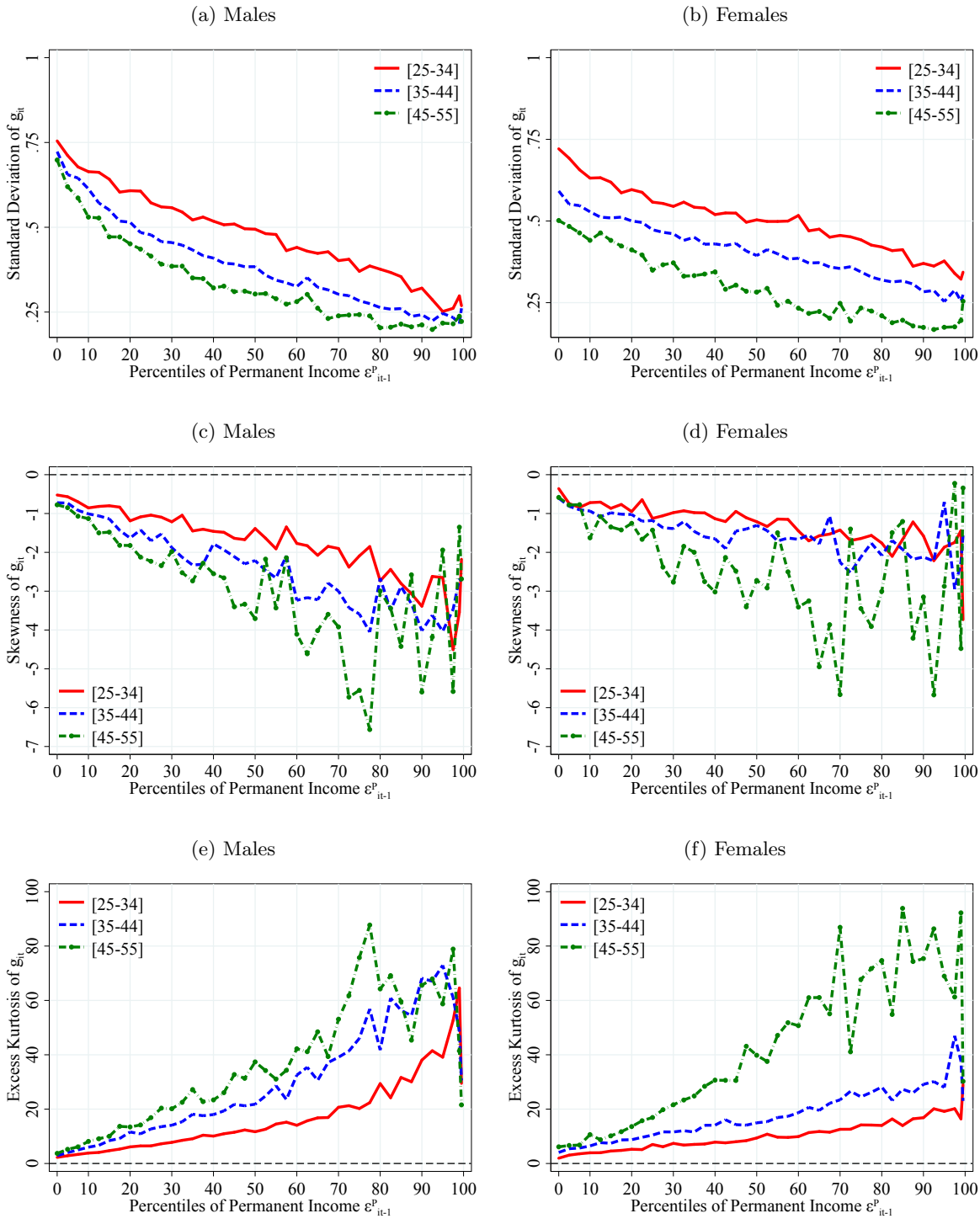


(b) Excess Kurtosis



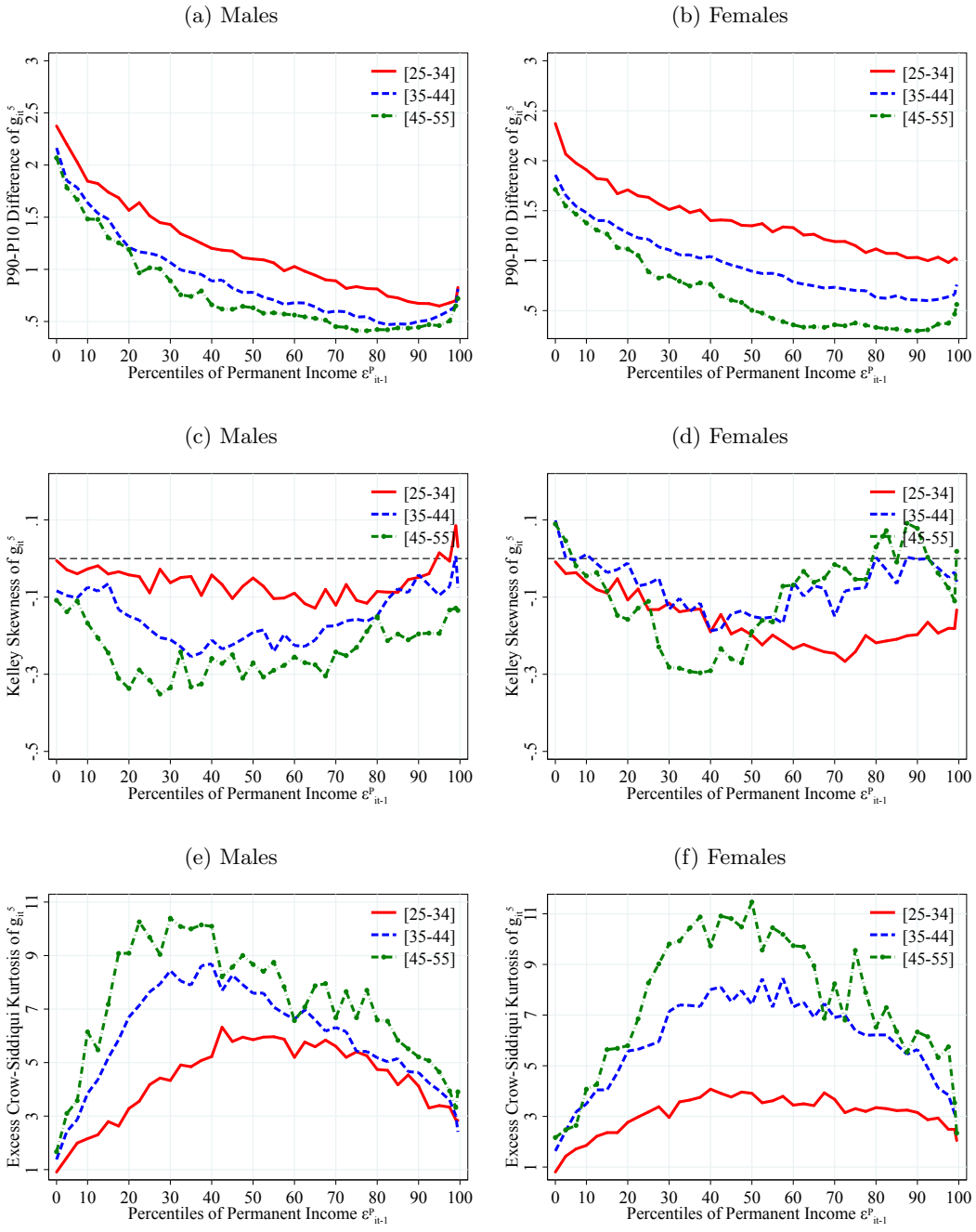
Notes: LS sample, five-year changes in residualized log earnings. The shaded areas indicate recession years.

Figure F19: Standardized moments of one-year earnings changes



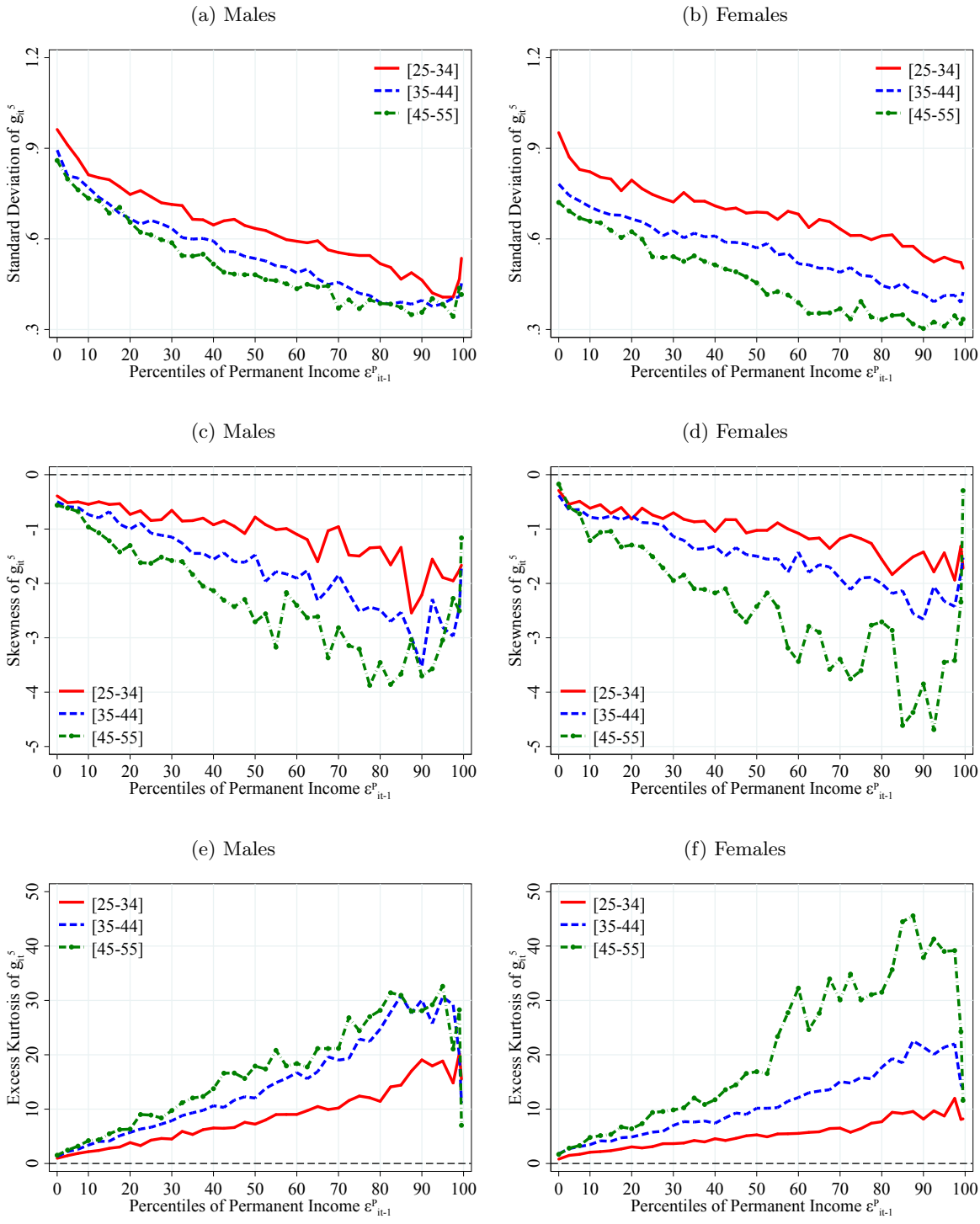
Notes: See the notes to Figure 6. Moment-based measures.

Figure F20: Dispersion, skewness and kurtosis of five-year earnings changes



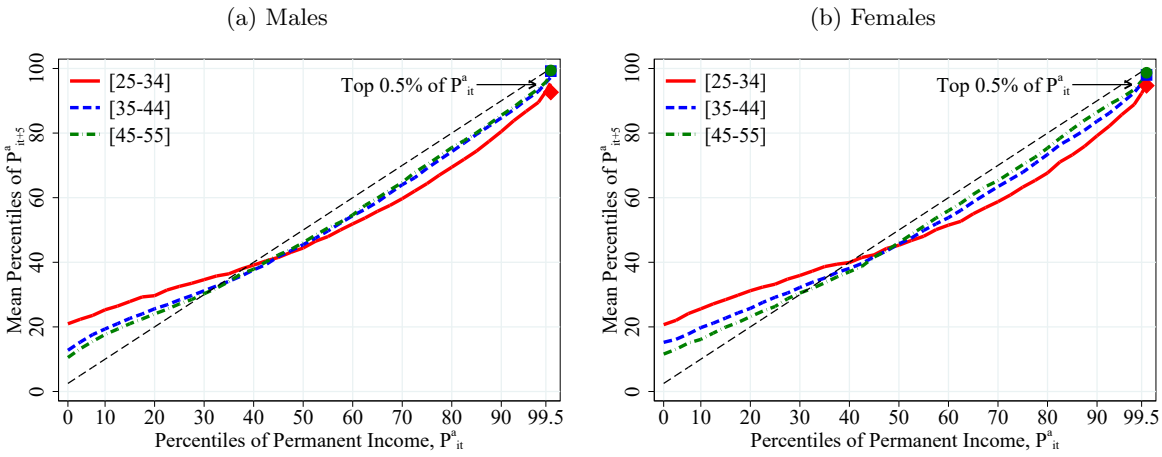
Notes: H sample, five-year changes in residualized log earnings. On the x-axis we report percentiles of residualized log permanent earnings ε^P_{it-1}. In the top panel we show the P90-P10 percentile difference, in the middle panel we show Kelley skewness, and in the bottom panel we show excess Crow-Siddiqui kurtosis. The various curves on the graphs corresponds to various age groups: between 25 and 34 years, between 35 and 44, and between 45 and 55 years, respectively.

Figure F21: Standardized moments of five-year earnings changes



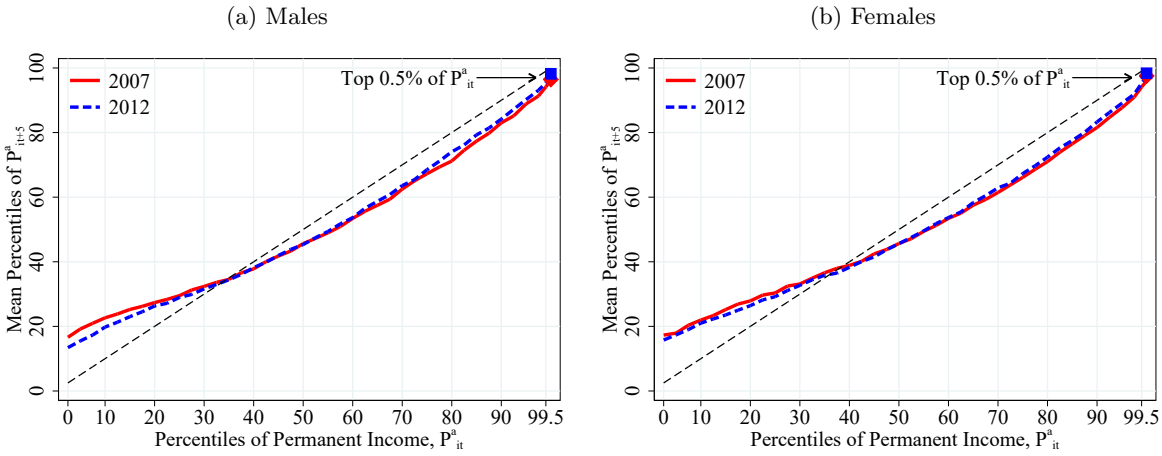
Notes: See the notes to Figure F20. Moment-based measures.

Figure F22: Evolution of mobility over the life cycle



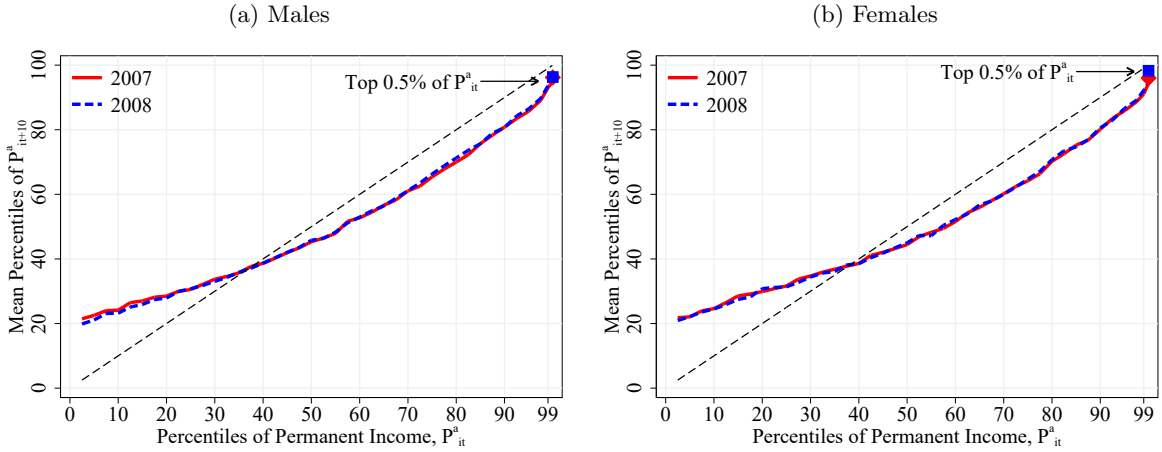
Notes: See the notes to Figure 8.

Figure F23: Evolution of mobility over time



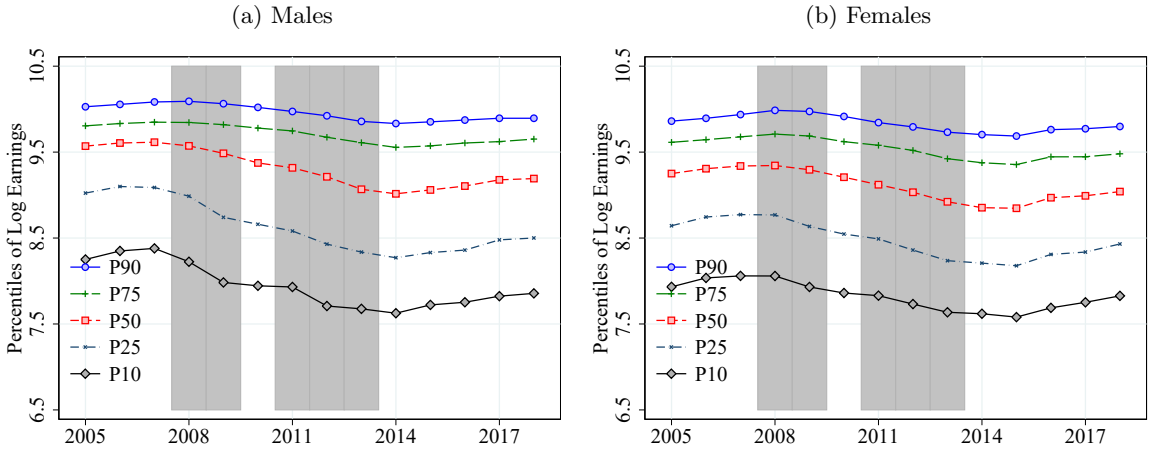
Notes: See the notes to Figure 8.

Figure F24: Evolution of 10-year mobility over time



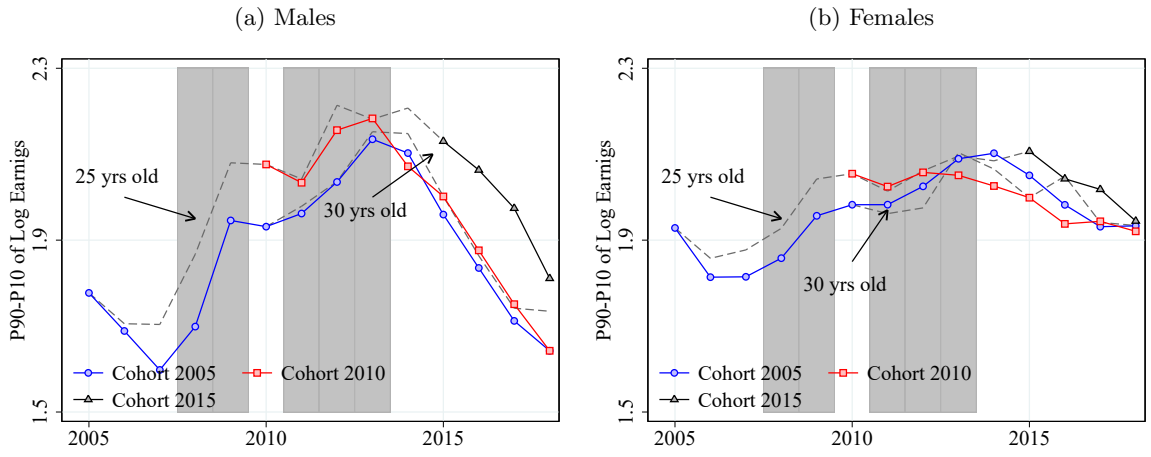
Notes: See the notes to Figure 8.

Figure F25: Overall distribution of workers at age 25



Notes: CS sample, log annual earnings. The sample is restricted to age-25 workers only. The shaded areas indicate recession years.

Figure F26: Earnings inequality by cohort

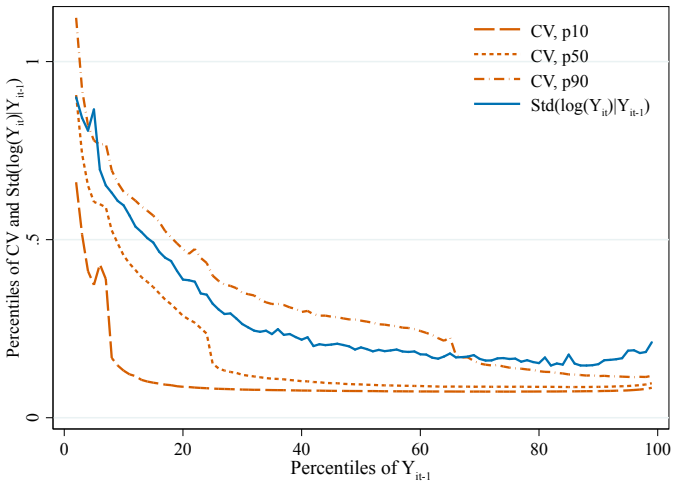


Notes: CS sample, log annual earnings. The different curves correspond to different cohorts of workers. The shaded areas indicate recession years.

G Additional tables and figures on income risk inequality

G.1 Baseline CV measure

Figure G1: Two measures of risk, $\mathrm{CV}(\mathrm{X_{it}})$ and $\mathrm{Std}(\log(Y_{it}) \,|\, Y_{it-1})$



Notes: B sample, with positive income. Exponential specification using all macro and micro predictors. Selected quantiles of the distribution of $\mathrm{CV}(\mathrm{X_{it}})$ *given income* Y_{it-1}*, and binned estimate of* $\mathrm{Std}(\log(\mathrm{Y_{it}}) \,|\, \mathrm{Y_{it-1}})$*, rescaled for comparability.*

Table G1: Poisson regression results

	(1) Income levels	(2) Income absolute deviations
Log lagged income	0.525	1.025
	(0.107)	(0.221)
Log lagged income squared	0.532	-0.0107
	(0.0864)	(0.167)
Log lagged income cubed	0.0299	-0.0251
	(0.0291)	(0.0363)
Indicator that lagged income is zero	-4.798	2.732
	(2.557)	(1.862)
Age	-0.0666	-0.0251
	(0.00875)	(0.0344)
Age squared	0.000615	-0.000444
	(0.000102)	(0.000415)
Education: lower secondary	0.102	-0.0280
	(0.0183)	(0.0712)
Education: upper secondary	0.315	0.347
	(0.0204)	(0.0941)
Education: college	0.733	0.859
	(0.0257)	(0.138)
Days worked in past year	-0.00134	-0.00382
	(0.000315)	(0.000937)
Indicator that out-of-work income is zero	-1.409	-2.475
	(0.181)	(0.709)
Log out-of-work income in past year	-0.997	-1.393
	(0.0917)	(0.359)
Indicator that worked full year in past year	-0.0214	0.0136
	(0.0416)	(0.201)
Indicator that worked full year in past two years	0.00279	-0.414
	(0.0343)	(0.207)
Indicator that worked full year in past three years	0.0797	0.536
	(0.0242)	(0.212)
Permanent contract in past year	-0.0149	0.173
	(0.0236)	(0.120)
Full-time contract in past year	0.219	0.225
	(0.0359)	(0.158)
Intercept	11.11	11.08
	(0.185)	(0.712)
N	3111191	3111191

Notes: B sample. Robust standard errors in parentheses, clustered at the individual level. In column (2), the dependent variable is $|Y_{it} - \exp(X'_{it}\widehat{\beta})|$*, where* $\widehat{\beta}$ *is shown in column (1). The standard errors in column (2) do not account for the fact that* $\widehat{\beta}$ *is estimated. Macro predictors and interactions with age and age squared are included in the regressions, but omitted from the table for conciseness.*

G.2 After-tax income

Table G2: Effective Tax Rates

year	base_1	tau_1	base_2	tau_2	base_3	tau_3	base_4	tau_4	base_5	tau_5	base_6	tau_6
2005	0	8.72	4080.0	4.42	14076.0	10.28	26316.0	16.74	45900.0	26.43	.	.
2006	0	11.29	4161.6	4.71	14357.5	10.52	26842.3	17.03	46818.0	26.89	.	.
2007	0	5.68	17360.0	12.32	32360.0	18.49	52360.0	27.5	.	.	.	.
2008	0	3.68	17707.2	10.51	33007.2	17.38	53407.2	26.83	.	.	.	.
2009	0	3.31	17707.2	10.45	33007.2	17.28	53407.2	26.66	.	.	.	.
2010	0	4.17	17707.2	12.25	33007.2	18.34	53407.2	27.03	.	.	.	.
2011	0	3.74	17707.2	12.21	33007.2	18.25	53407.2	26.42	120000.2	34.59	175000.2	39.58
2012	0	3.75	17707.2	12.93	33007.2	19.47	53407.2	28.28	120000.2	37.61	175000.2	43.31
2013	0	3.62	17707.2	13.12	33007.2	19.55	53407.2	28.32	120000.2	37.75	175000.2	43.80
2014	0	3.48	17707.2	13.12	33007.2	19.52	53407.2	28.16	120000.2	37.52	175000.2	43.68
2015	0	1.68	12450.0	5.95	20200.0	12.66	34000.0	18.81	60000.0	28.68	.	.
2016	0	2.04	12450.0	6.02	20200.0	12.89	35200.0	19.22	60000.0	28.62	.	.
2017	0	2.47	12450.0	6.20	20200.0	12.79	35200.0	19.38	60000.0	28.92	.	.
2018	0	2.47	12450.0	6.20	20200.0	12.79	35200.0	19.38	60000.0	28.92	.	.

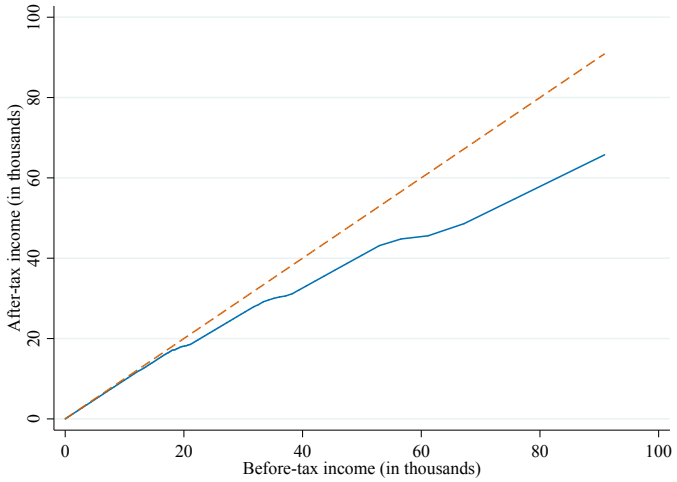
Notes: "base_x" represents the upper and lower bounds of income brackets. "tau_x" represents the average effective tax rates. All income bracket bounds are in nominal euros. Average effective tax rates are in percent.

Table G3: Income risk over the period, in numbers, after-tax income

	All	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
P90/P10	9.17	5.18	5.65	5.72	6.76	9.16	9.89	10.95	12.68	13.32	11.91	10.37	9.31	8.25
P90/P50	6.41	3.97	4.20	4.19	4.92	6.41	6.70	7.23	8.17	8.46	7.75	6.93	6.52	5.81
P50/P10	1.43	1.30	1.34	1.36	1.37	1.43	1.48	1.51	1.55	1.58	1.54	1.50	1.43	1.42
p10	0.07	0.08	0.08	0.08	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.08	0.07
p25	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
p50	0.11	0.10	0.10	0.10	0.11	0.10	0.10	0.11	0.11	0.11	0.11	0.11	0.11	0.10
p75	0.32	0.23	0.24	0.24	0.27	0.34	0.36	0.40	0.45	0.46	0.42	0.35	0.33	0.29
p90	0.67	0.40	0.43	0.43	0.52	0.67	0.70	0.80	0.90	0.93	0.85	0.75	0.70	0.61

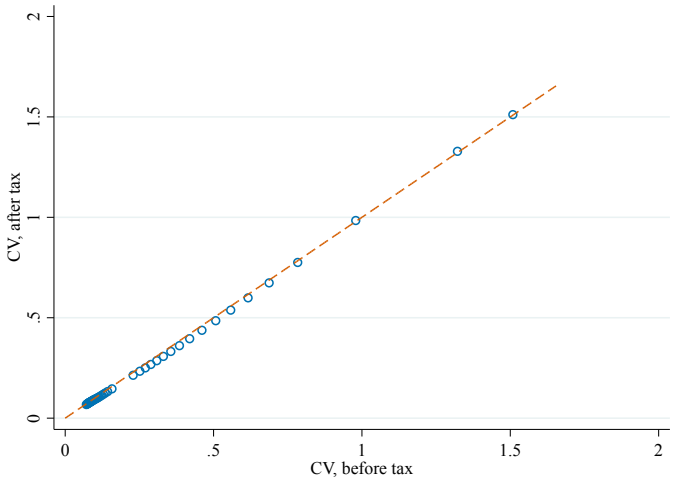
Notes: B sample. Exponential specification, using all macro and micro predictors. After-tax income.

Figure G3: Before-tax and after-tax income



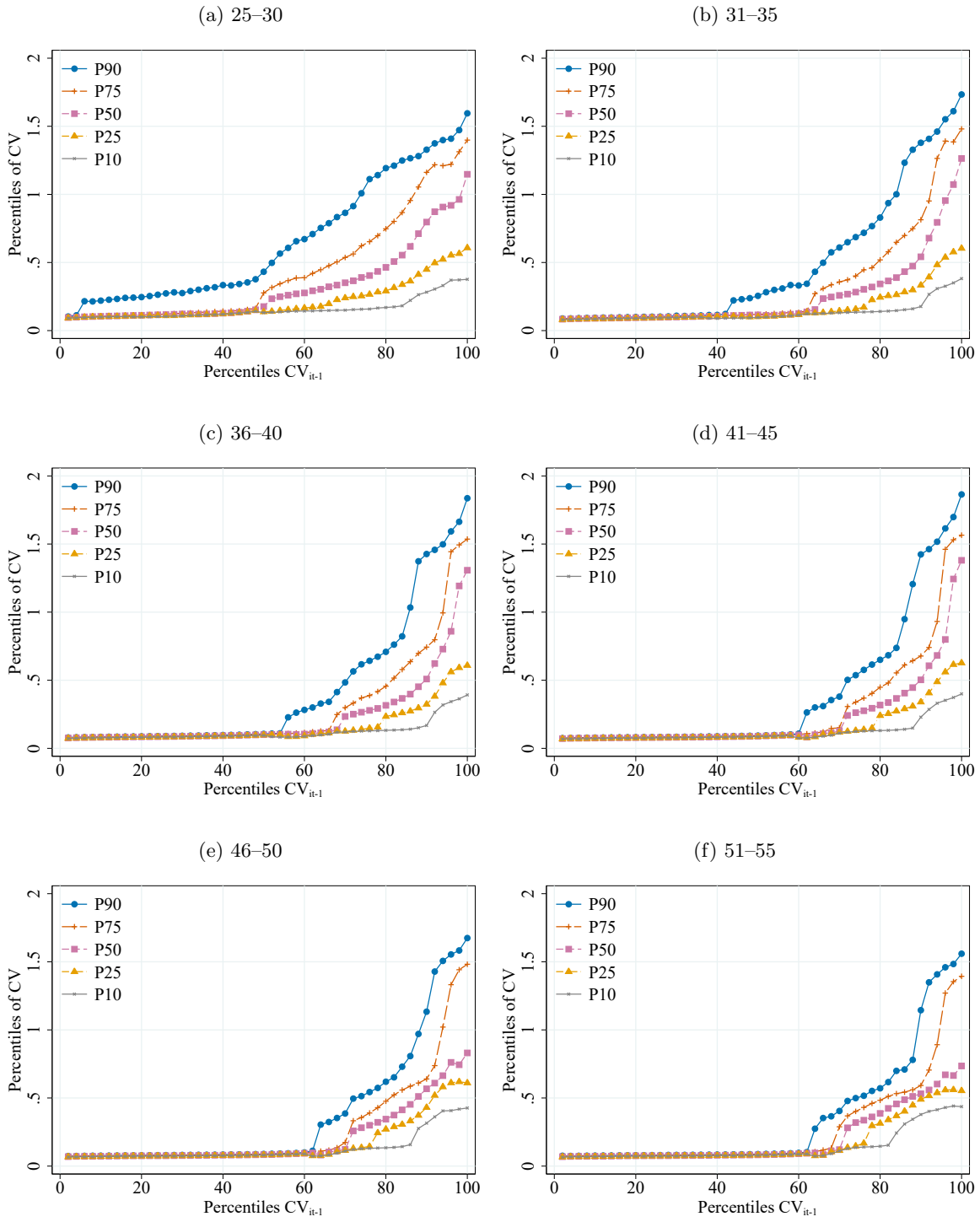
Notes: Income in thousands of euros. See the text for how we construct after-tax income.

Figure G4: Quantile-quantile plot of before-tax and after-tax CV



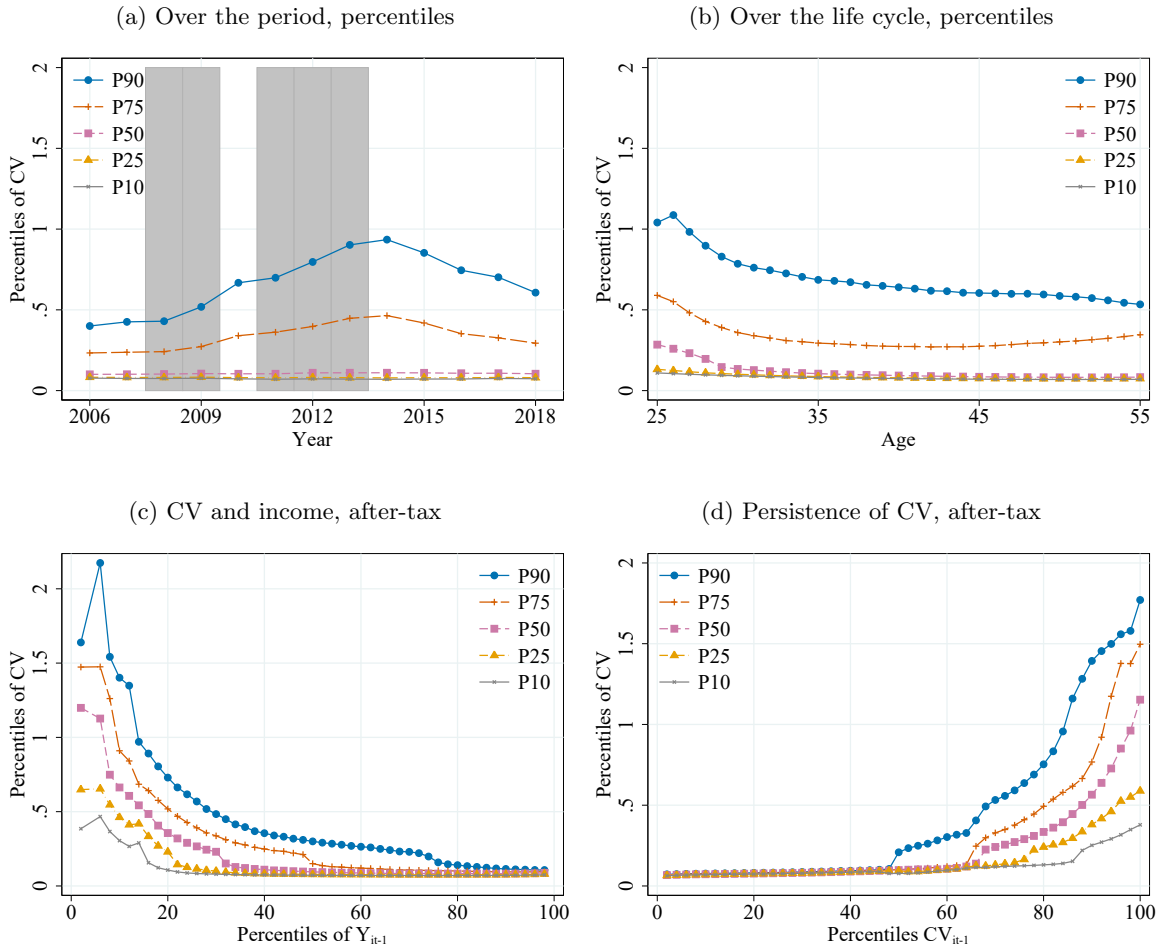
Notes: B sample. On the x-axis we report the percentiles of CV based on before-tax income. On the y-axis we report the percentiles of CV based on after-tax income.

Figure G2: CV persistence, by age



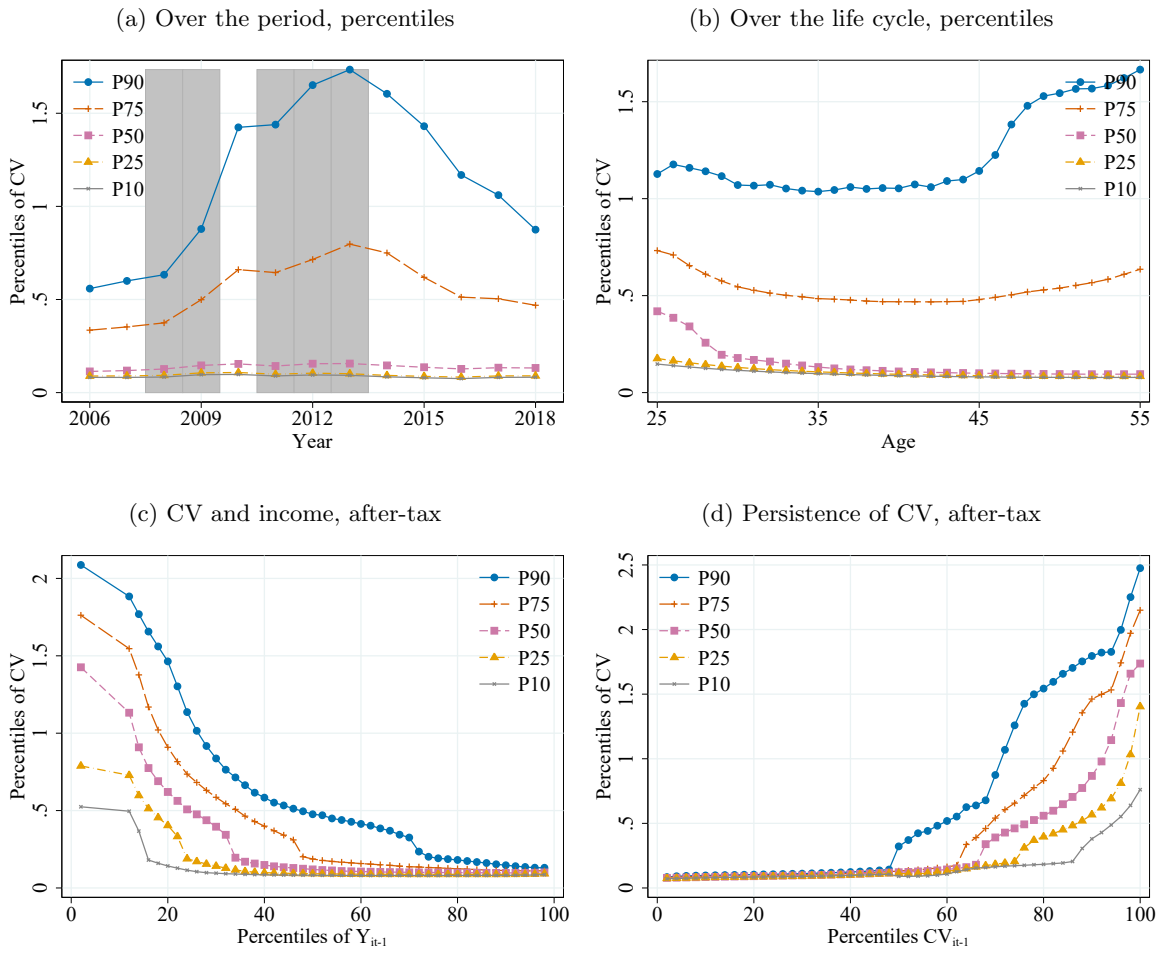
Notes: B sample. Exponential specification, using all macro and micro predictors.

Figure G5: CV, after-tax income



Notes: B sample. Exponential specification, using all macro and micro predictors. After-tax income. The shaded areas indicate recession years.

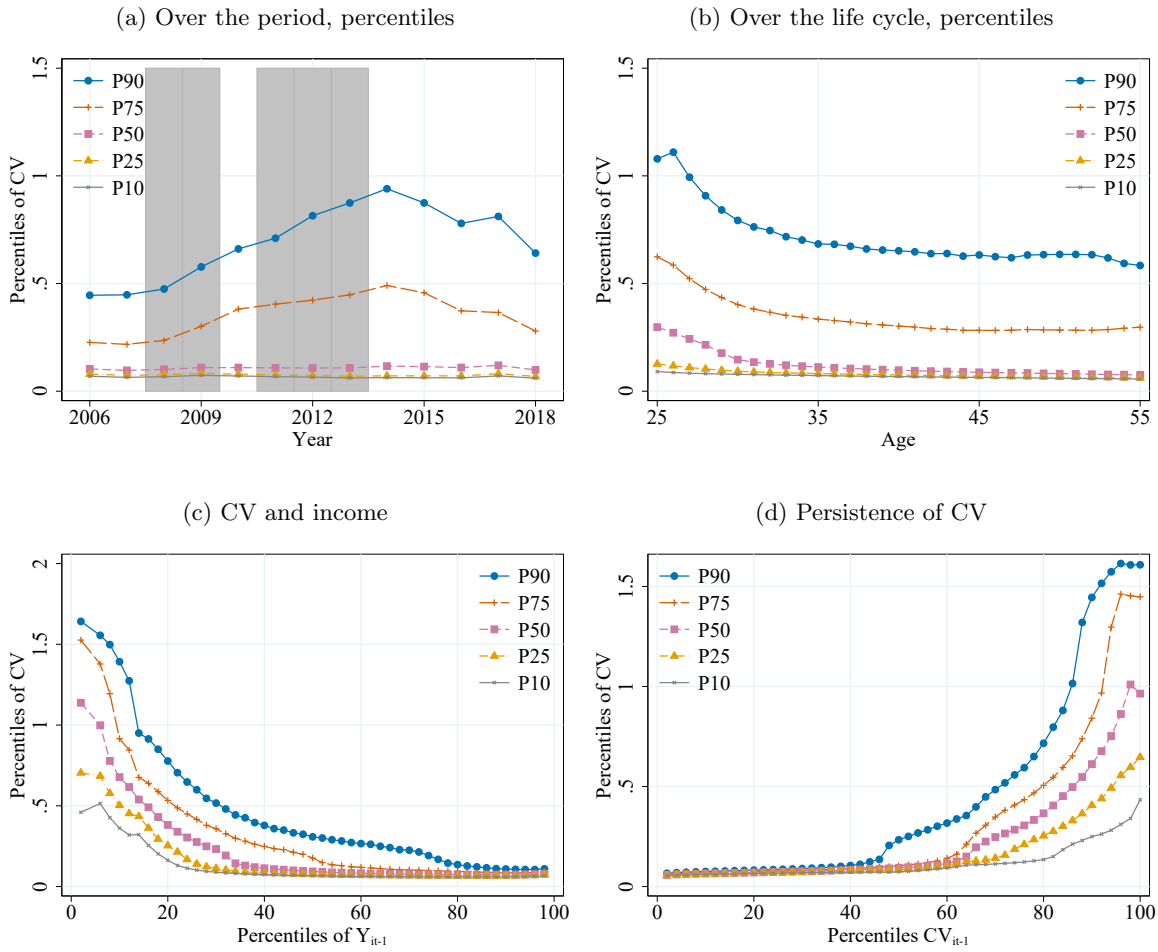
Figure G6: CV, income net of unemployment benefits



Notes: B sample, income net of unemployment benefits. Exponential specification, using all macro and micro predictors. After-tax income. The shaded areas indicate recession years.

G.3 Neural network specification

Figure G7: CV, neural network specification



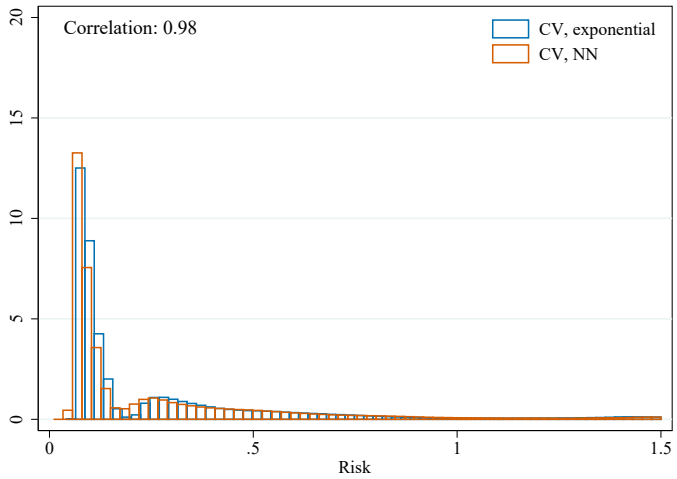
Notes: B sample. Neural network specification with Poisson loss. One layer with 8 nodes for the conditional mean and 7 nodes for the conditional mean absolute deviation. The shaded areas indicate recession years.

Table G4: Income risk over the period, in numbers, neural network specification

	All	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
P90/P10	10.82	6.41	6.96	7.04	7.85	9.33	10.52	12.51	14.16	14.93	13.98	12.59	11.43	10.63
P90/P50	6.63	4.31	4.64	4.69	5.27	6.03	6.55	7.55	8.09	8.07	7.66	7.10	6.76	6.49
P50/P10	1.63	1.49	1.50	1.50	1.49	1.55	1.61	1.66	1.75	1.85	1.83	1.77	1.69	1.64
p10	0.07	0.07	0.06	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06	0.06	0.07	0.06
p25	0.08	0.08	0.07	0.08	0.08	0.08	0.08	0.07	0.07	0.07	0.07	0.07	0.08	0.07
p50	0.11	0.10	0.10	0.10	0.11	0.11	0.11	0.11	0.11	0.12	0.11	0.11	0.12	0.10
p75	0.34	0.23	0.22	0.24	0.30	0.38	0.40	0.42	0.45	0.49	0.46	0.37	0.37	0.28
p90	0.71	0.45	0.45	0.47	0.58	0.66	0.71	0.81	0.87	0.94	0.87	0.78	0.81	0.64

Notes: B sample. Neural network specification with Poisson loss. One layer with 8 nodes for the conditional mean and 7 nodes for the conditional mean absolute deviation.

Figure G8: Two specifications of CV, exponential and neural network



Notes: B sample, using all macro and micro predictors. The correlation coefficient is computed after trimming the 99th percentiles of both CV measures.

G.4 Specification with unobserved heterogeneity

Table G5: Education Distribution of Individuals by Cluster

(a) Mean clusters

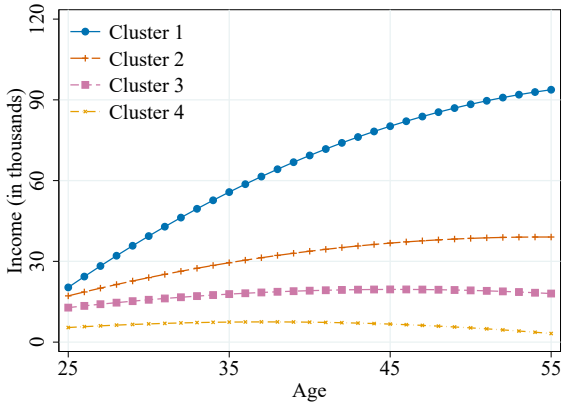
		Education (Proportion of Indivs.)			
Cluster	# of Indiv.	Primary	Lower Sec.	Upper Sec.	College
1	20002	0.16	0.33	0.27	0.24
2	96641	0.17	0.39	0.25	0.19
3	132677	0.16	0.43	0.25	0.16
4	56827	0.23	0.40	0.23	0.14

(b) Absolute deviation clusters

		Education (Proportion of Indivs.)			
Cluster	# of Indiv.	Primary	Lower Sec.	Upper Sec.	College
1	73778	0.17	0.43	0.24	0.17
2	100890	0.17	0.39	0.25	0.19
3	51268	0.20	0.42	0.23	0.15
4	80211	0.17	0.39	0.27	0.17

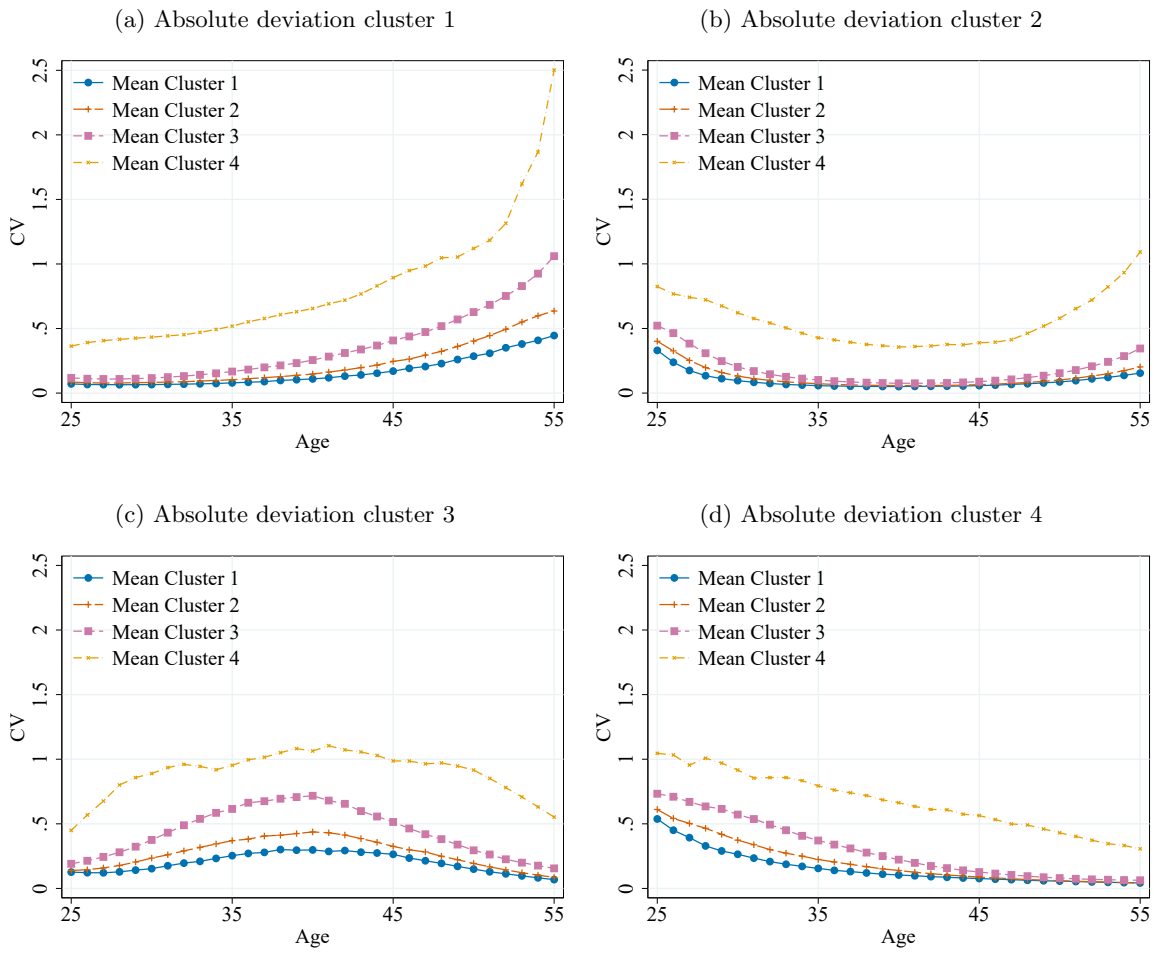
Notes: B sample, individuals with at least 4 observations prior to 2018. Specification with unobserved heterogeneity, 4 groups. The mean clusters correspond to the prediction of income levels, the absolute deviation clusters correspond to the prediction of income absolute deviations.

Figure G9: Predicted age income profiles for the estimated mean groups ($K = 4$)



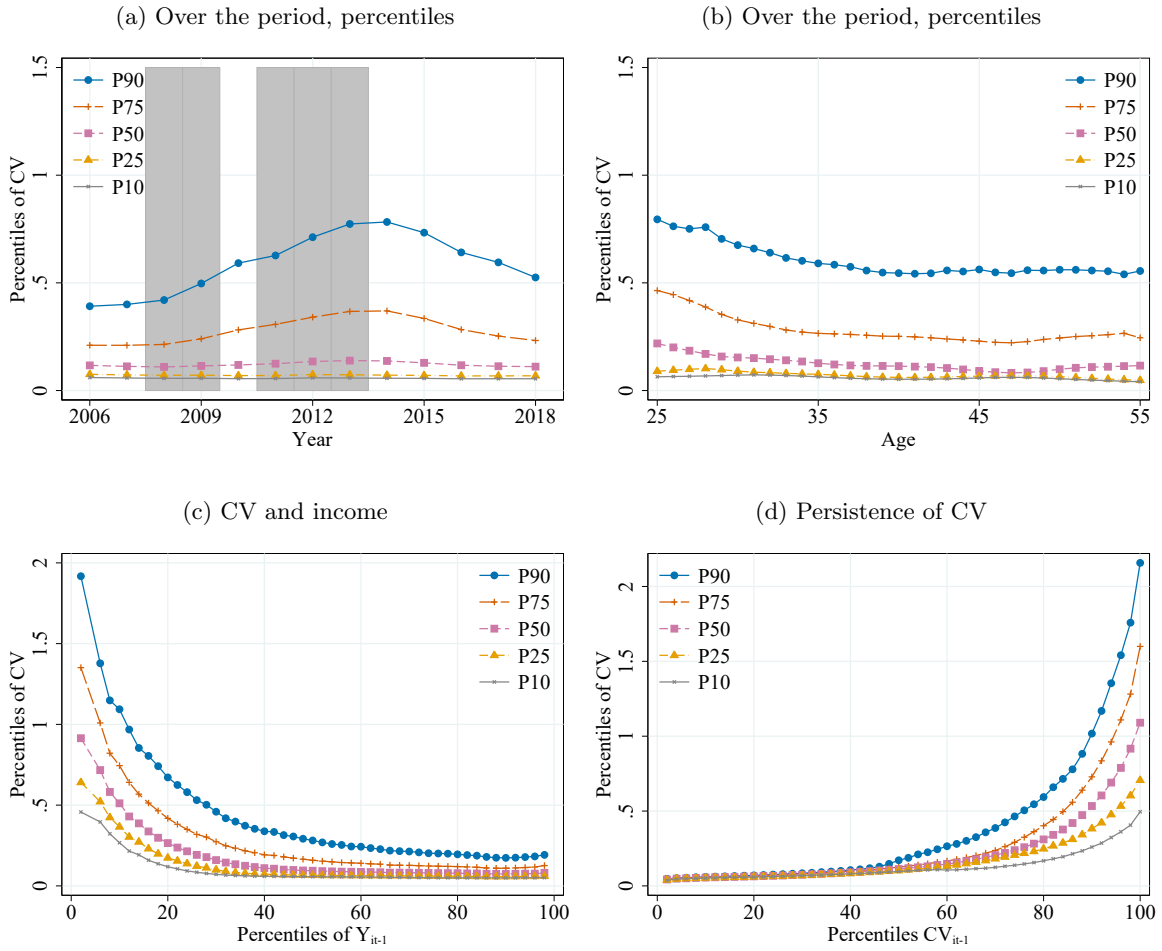
Notes: B sample, individuals with at least 4 observations prior to 2018. Specification with unobserved heterogeneity, 4 groups. The mean clusters correspond to the prediction of income levels, the absolute deviation clusters correspond to the prediction of income absolute deviations.

Figure G10: Average CV over age, by mean cluster and absolute deviation cluster



Notes: B sample, individuals with at least 4 observations prior to 2018. Specification with unobserved heterogeneity, 4 groups. The mean clusters correspond to the prediction of income levels, the absolute deviation clusters correspond to the prediction of income absolute deviations.

Figure G11: CV, specification with unobserved heterogeneity



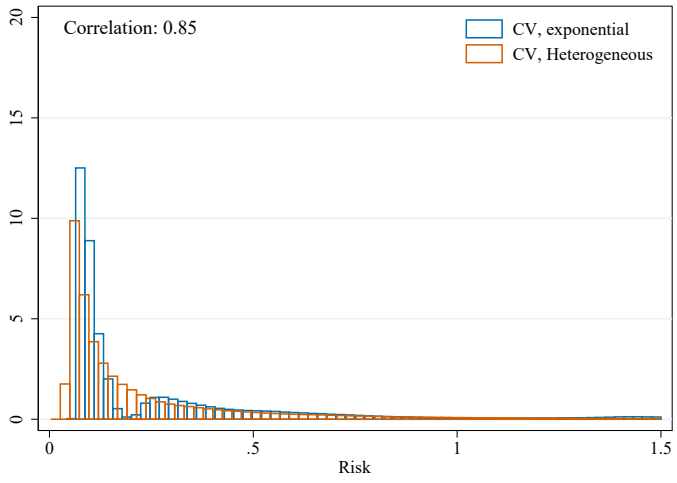
Notes: B sample, individuals with at least 4 observations prior to 2018. Exponential specification, using all macro and micro predictors, and unobserved heterogeneity. 4 groups. The shaded areas indicate recession years.

Table G6: Income risk over the period, in numbers, specification with unobserved heterogeneity

	All	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
P90/P10	10.59	6.48	6.86	7.41	8.73	10.72	11.25	12.09	13.29	13.72	12.97	11.76	10.86	9.60
P90/P50	5.00	3.35	3.56	3.83	4.35	4.98	5.03	5.28	5.56	5.69	5.70	5.46	5.27	4.74
P50/P10	2.12	1.93	1.93	1.94	2.01	2.15	2.24	2.29	2.39	2.41	2.28	2.15	2.06	2.03
p10	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.05	0.05
p25	0.07	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
p50	0.12	0.12	0.11	0.11	0.11	0.12	0.12	0.13	0.14	0.14	0.13	0.12	0.11	0.11
p75	0.27	0.21	0.21	0.21	0.24	0.28	0.31	0.34	0.37	0.37	0.33	0.28	0.25	0.23
p90	0.60	0.39	0.40	0.42	0.50	0.59	0.63	0.71	0.77	0.78	0.73	0.64	0.60	0.53

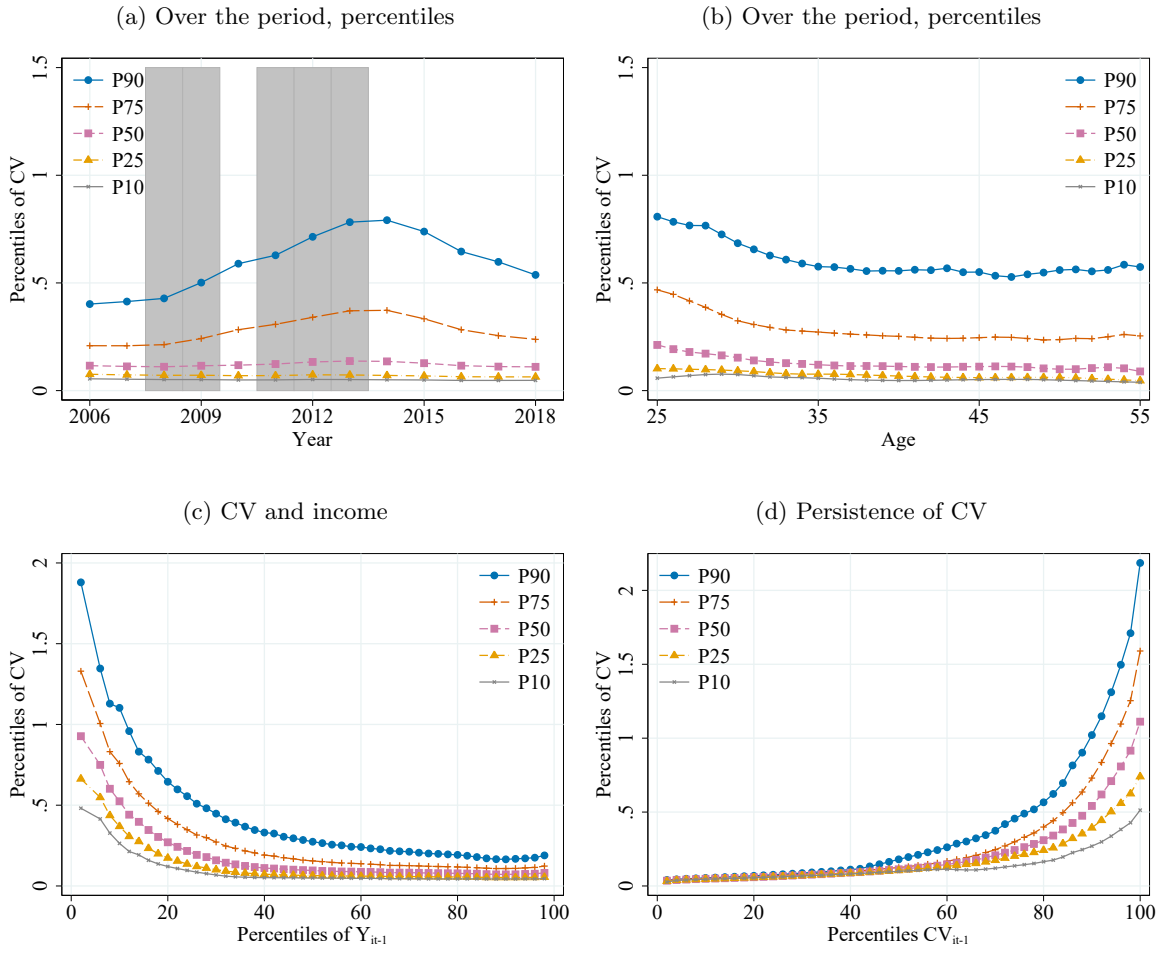
Notes: B sample, individuals with at least 4 observations prior to 2018. Exponential specification, using all macro and micro predictors and unobserved heterogeneity. 4 groups.

Figure G12: Two specifications of CV, with and without unobserved heterogeneity



Notes: B sample, individuals with at least 4 observations prior to 2018. Exponential specification, using all macro and micro predictors, and unobserved heterogeneity in the right graph (4 groups). The correlation coefficient is computed after trimming the 99th percentiles of both CV measures.

Figure G13: CV, specification with unobserved heterogeneity, 6 groups



Notes: B sample, individuals with at least 4 observations prior to 2018. Exponential specification, using all macro and micro predictors, and unobserved heterogeneity. 6 groups.

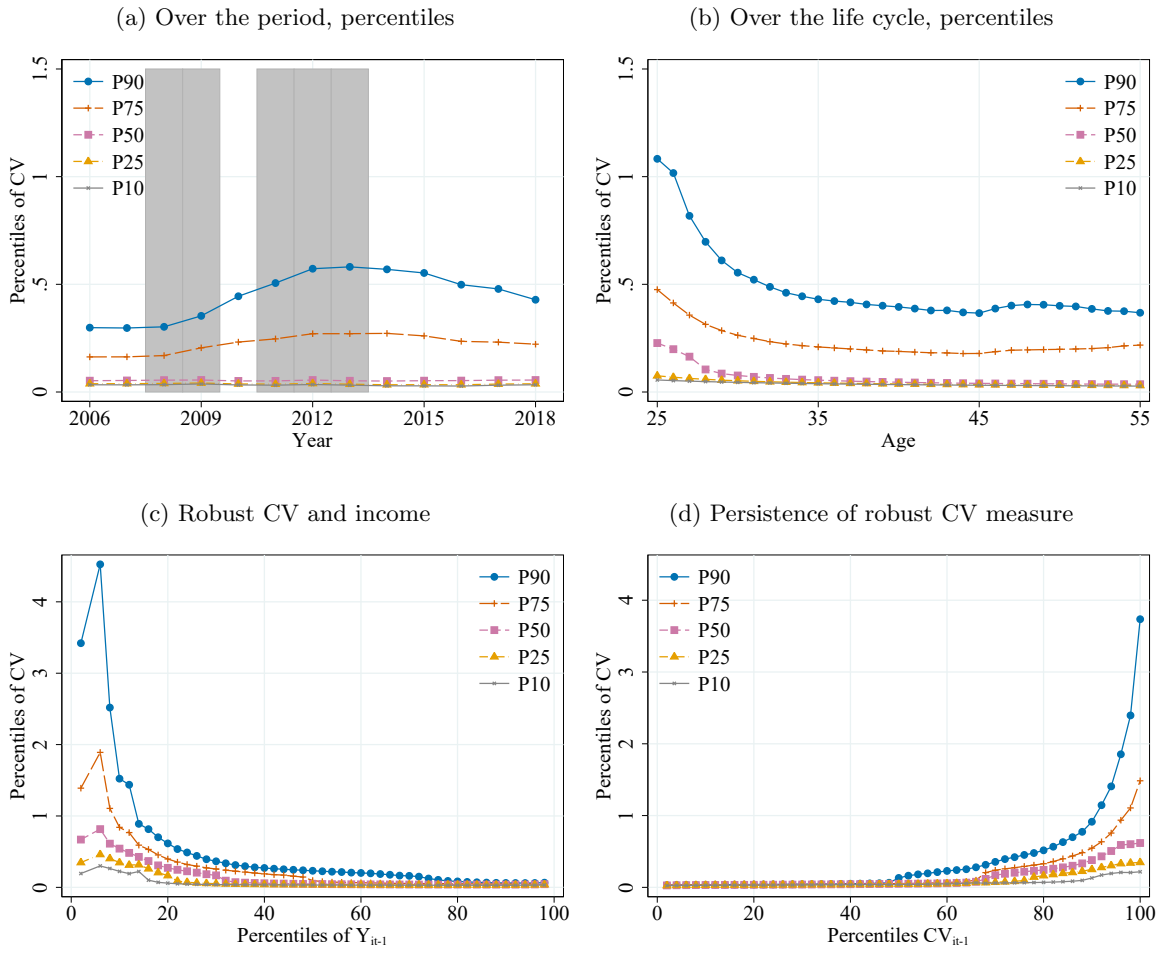
Table G7: Income risk over the period, in numbers, specification with unobserved heterogeneity, 6 groups

	All	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
P90/P10	11.99	7.33	7.81	8.34	9.76	11.90	12.62	13.77	15.23	15.75	14.94	13.65	12.64	11.28
P90/P50	5.04	3.47	3.68	3.87	4.34	4.99	5.09	5.36	5.70	5.83	5.80	5.57	5.36	4.87
P50/P10	2.38	2.11	2.12	2.16	2.25	2.39	2.48	2.57	2.67	2.70	2.58	2.45	2.36	2.32
p10	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
p25	0.07	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.06
p50	0.12	0.12	0.11	0.11	0.12	0.12	0.12	0.13	0.14	0.14	0.13	0.12	0.11	0.11
p75	0.27	0.21	0.21	0.21	0.24	0.28	0.31	0.34	0.37	0.37	0.33	0.28	0.26	0.24
p90	0.60	0.40	0.41	0.43	0.50	0.59	0.63	0.71	0.78	0.79	0.74	0.65	0.60	0.54

Notes: B sample, individuals with at least 4 observations prior to 2018. Exponential specification, using all macro and micro predictors and unobserved heterogeneity. 6 groups.

G.5 Robust CV

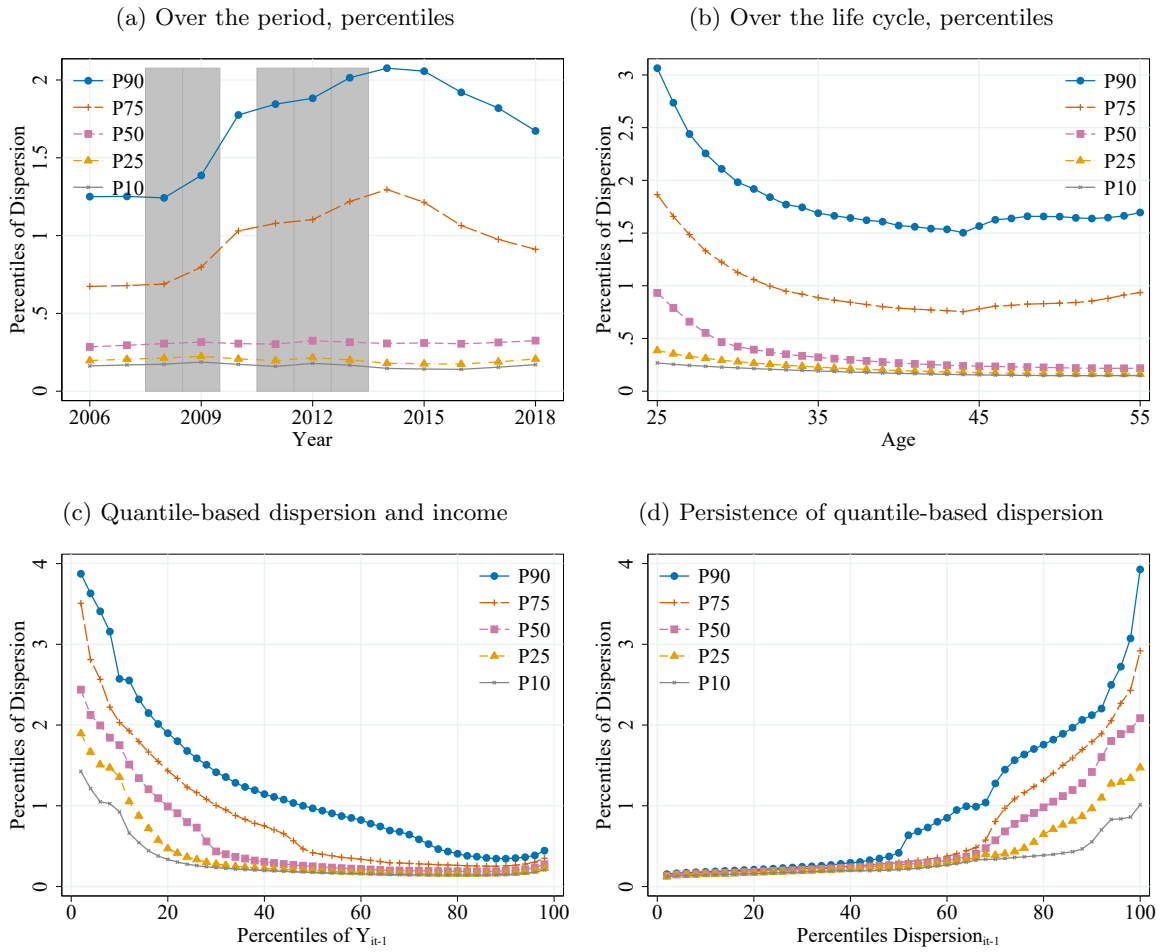
Figure G14: Robust CV measure



Notes: B sample. Robust CV measure, see equation (B5). The shaded areas indicate recession years.

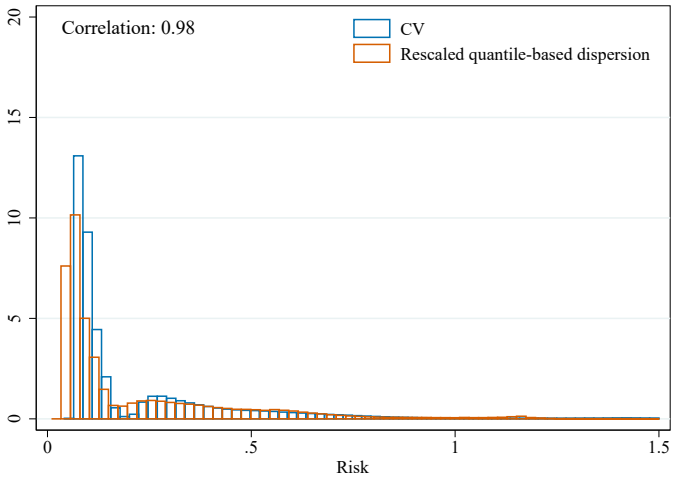
G.6 Beyond the CV

Figure G15: Quantile-based dispersion



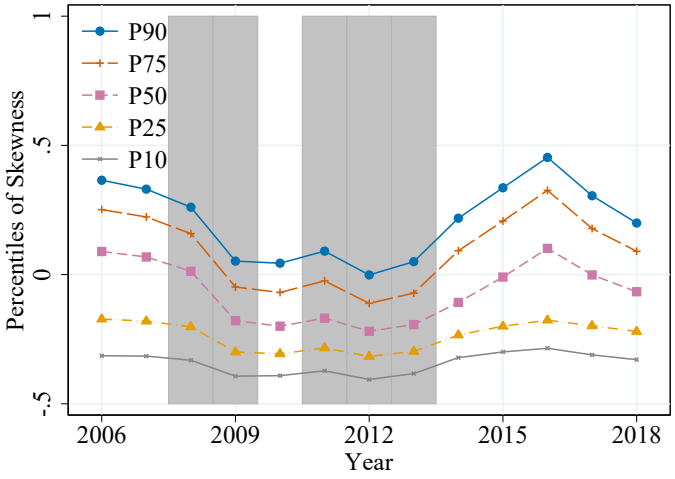
Notes: B sample, with positive income. Quantile-based measure of dispersion risk, $P90(X_{it}) - P10(X_{it})$, where $P90(X_{it})$ and $P10(X_{it})$ are estimated using linear quantile regressions of log income on all macro and micro predictors. The shaded areas indicate recession years.

Figure G16: Comparing CV and quantile-based dispersion



Notes: B sample, with positive income. CV is our CV measure (i.e., conditional mean absolute deviation divided by conditional mean). Quantile-based dispersion is $P90(X_{it}) - P10(X_{it})$ *(rescaled), where* $P90(X_{it})$ *and* $P10(X_{it})$ *are estimated using linear quantile regressions of log income on all macro and micro predictors. The correlation coefficient is computed after trimming the 99th percentiles of both CV measures.*

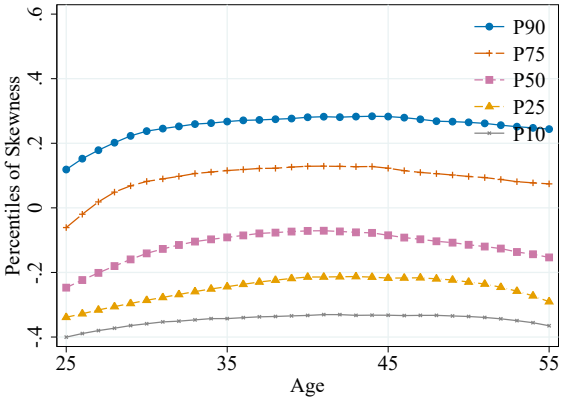
Figure G17: Quantile-based skewness over the period



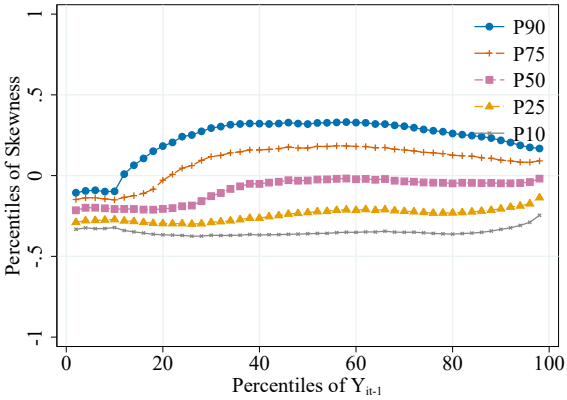
Notes: B sample, with positive income. Quantile-based measure of skewness risk, $\frac{P90(X_{it})-2P50(X_{it})+P10(X_{it})}{P90(X_{it})-P10(X_{it})}$, *where* $P90(X_{it})$, $P50(X_{it})$, *and* $P10(X_{it})$ *are estimated using linear quantile regressions of log income on all macro and micro predictors. The shaded areas indicate recession years.*

Figure G18: Quantile-based skewness, additional results

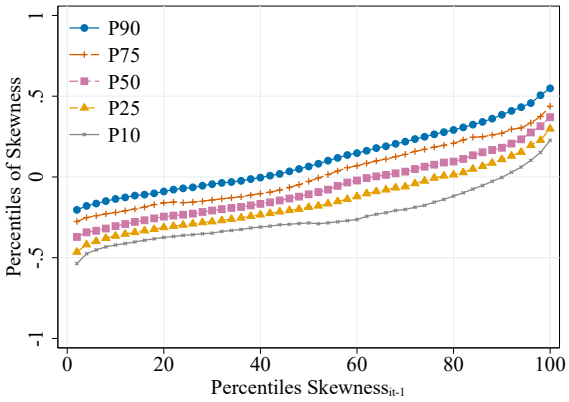
(a) Over the life cycle, percentiles



(b) Skewness and income



(c) Persistence of quantile-based skewness



Notes: B sample, with positive income. Quantile-based measure of skewness risk, $\frac{P90(X_{it})-2P50(X_{it})+P10(X_{it})}{P90(X_{it})-P10(X_{it})}$, *where* $P90(X_{it})$, $P50(X_{it})$, *and* $P10(X_{it})$ *are estimated using linear quantile regressions of log income on all macro and micro predictors.*

G.7 The income risk for civil servants

Table G8: Income risk over the period, in numbers, civil servants

	All	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
P90/P10	1.65	1.73	1.76	1.76	1.73	1.72	1.69	1.63	1.58	1.49	1.44	1.47	1.45	1.44
P90/P50	1.43	1.51	1.53	1.52	1.50	1.48	1.47	1.42	1.38	1.31	1.27	1.30	1.28	1.28
P50/P10	1.16	1.15	1.15	1.16	1.15	1.16	1.16	1.15	1.14	1.14	1.14	1.14	1.13	1.13
p10	0.08	0.08	0.08	0.08	0.08	0.08	0.07	0.08	0.07	0.07	0.07	0.07	0.08	0.08
p25	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
p50	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.08	0.08	0.08	0.08	0.09	0.09
p75	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.09	0.09	0.09	0.09	0.09	0.09
p90	0.12	0.14	0.14	0.14	0.14	0.13	0.13	0.12	0.12	0.11	0.10	0.11	0.11	0.11

Notes: B sample, restricted to civil servants under permanent contracts. Exponential specification, using all macro and micro predictors.

References

Altonji, Joseph G., Anthony A. Smith Jr., and Ivan Vidangos (2013) "Modeling Earnings Dynamics," *Econometrica*, 81 (4), 1395–1454, https://doi.org/10.3982/ECTA8415.

Arachchige, Chandima N. P. G., Luke A. Prendergast, and Robert G. Staudte (2020) "Robust analogs to the coefficient of variation," *Journal of Applied Statistics*, 1–23, 10.1080/02664763.2020.1808599.

Arellano, Manuel and Jinyong Hahn (2016) "A likelihood-Based Approximate Solution to the Incidental Parameter Problem in Dynamic Nonlinear Models with Multiple Effects," *Global Economic Review*, 45 (3), 251–274, 10.1080/1226508X.2016.1211811.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2021) "Discretizing unobserved heterogeneity," Working Paper.

Bonhomme, Stéphane and Elena Manresa (2015) "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83 (3), 1147–1184, https://doi.org/10.3982/ECTA11319.

Buchinsky, Moshe and Jinyong Hahn (1998) "An Alternative Estimator for the Censored Quantile Regression Model," *Econometrica*, 66 (3), 653–671, http://www.jstor.org/stable/2998578.

Garcia-Miralles, Esteban, Nezih Guner, and Roberto Ramos (2019) "The Spanish personal income tax: facts and parametric estimates," *SERIEs*, 10, 439–477, 10.1007/s13209-019-0197-5.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016) *Deep Learning*: MIT Press, http://www.deeplearningbook.org.

Guvenen, Fatih, Serdar Ozkan, and Jae Song (2014) "The Nature of Countercyclical Income Risk," *Journal of Political Economy*, 122 (3), 621–660, 10.1086/675535.

H2O.ai (2020) "h2o: R Interface for H2O," http://www.h2o.ai, R package version 3.30.1.3.

Hahn, Jinyong and Guido Kuersteiner (2011) "Bias reduction for dynamic non-linear panel data models with fixed effects," *Econometric Theory*, 27 (6), 1152–1191, http://www.jstor.org/stable/41300604.