Income Risk Inequality: Evidence from Spanish Administrative Records*

Manuel Arellano
CEMFI

Stéphane Bonhomme
University of Chicago

Micole De Vera
CEMFI

Laura Hospido
Banco de España and IZA

Siqi Wei
CEMFI

July 18, 2022

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 CODE: D31, E24, E31, J31.

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

*This work is part of the Global Repository of Income Dynamics (GRID) Project, set up by Fatih Guvenen, Luigi Pistaferri, and Gianluca Violante, to produce and maintain an open-access, international, database of micro statistics on income inequality and 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, the EEA-ESEM meeting and the GRID 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

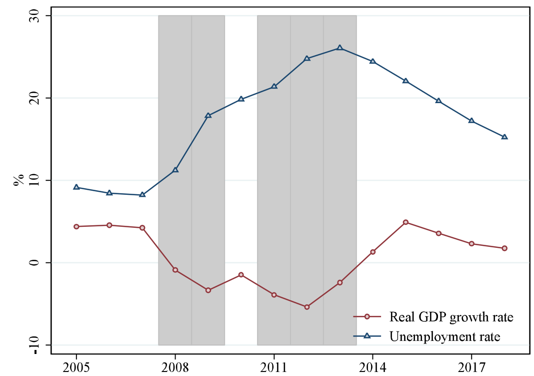
Income inequality is the focus of a large empirical literature, which now spans many countries over decades or centuries (Atkinson, 2003, Alvaredo et al., 2017). However, the measurement of cross-sectional inequality only provides an incomplete understanding of the diversity of individual income trajectories, 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., 2021, Busch et al., 2022).

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. (2021) 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 attempt to capture key determinants of the agent's information set in the data.

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.

Given 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 Repository of Income Dynamics (GRID) 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 over the life cycle and with 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., 2022, 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 $\pm 2,000$ 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 past 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 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. (2022). 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. In the appendix, we provide several complements to the paper. In Appendix A we provide additional results relating to Section 3. In Appendix B we provide robustness checks and extensions to the main results. Lastly, in Appendix C we provide additional empirical results relating to Sections 4 and 5. An Online Appendix accessible in the replication folder contains additional complementary results. Codes to access the data and replicate the results are available here and as part of the replication files.

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.

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 pre-tax 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

[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. Reported earnings include all taxable payments of the employer to the employee including overtime pay, bonuses, paid vacation, and sick leave benefits.

Section 4, we use a broader measure of earnings that also includes unemployment benefits and subsidies, and we explicitly account for the presence of zero earnings in the analysis. All earnings measures are deflated to 2018 euros using the Spanish consumer price index.

Two features of the data are worth highlighting. First, the period of observation is relatively short. As mentioned above, for the years prior to 2005, income records are top and 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 about 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,[2] 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 GRID 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 A1 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. When studying 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, while 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

[2]We observe whether the individual has declared herself as self-employed but not how much is earned in self-employment income.

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.[3] 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 focus our analysis of income risk on males only.

Descriptive statistics. We provide descriptive statistics about the samples in the appendix.[4] 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] 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.

[4] In Appendix Tables A2 and A3 we show summary statistics for the CS sample, and in Appendix Tables A4, A5, and A6 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 Online Appendix Tables S-A1, S-A2, S-A3, and S-A4 we show the same summary statistics where we convert earnings to US Dollars using the 2018 exchange rate.

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. (2021) and work inspired by their empirical methodology.

3.1 Income inequality

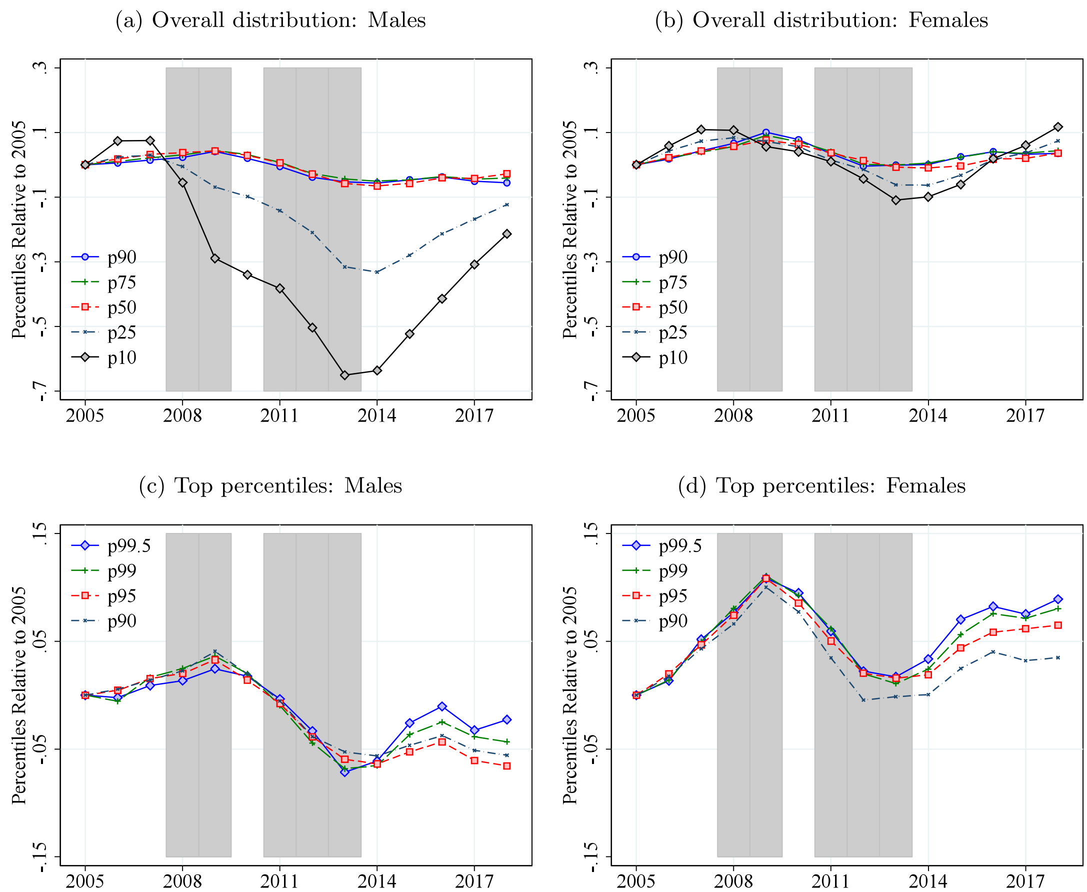
In Figure 2 we start by showing percentiles of log real earnings, by gender, from 2005 to 2018, taking 2005 as the reference year.[5] In the top two graphs, we show the 10th, 25th, median, 75th, and 90th percentiles for males and females, respectively. While the evolution of earnings percentiles over 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

[5] In Appendix Figure A1 we show the original percentiles, without normalizing them to zero in 2005.

Figure 2: Percentiles of the distribution of log annual earnings

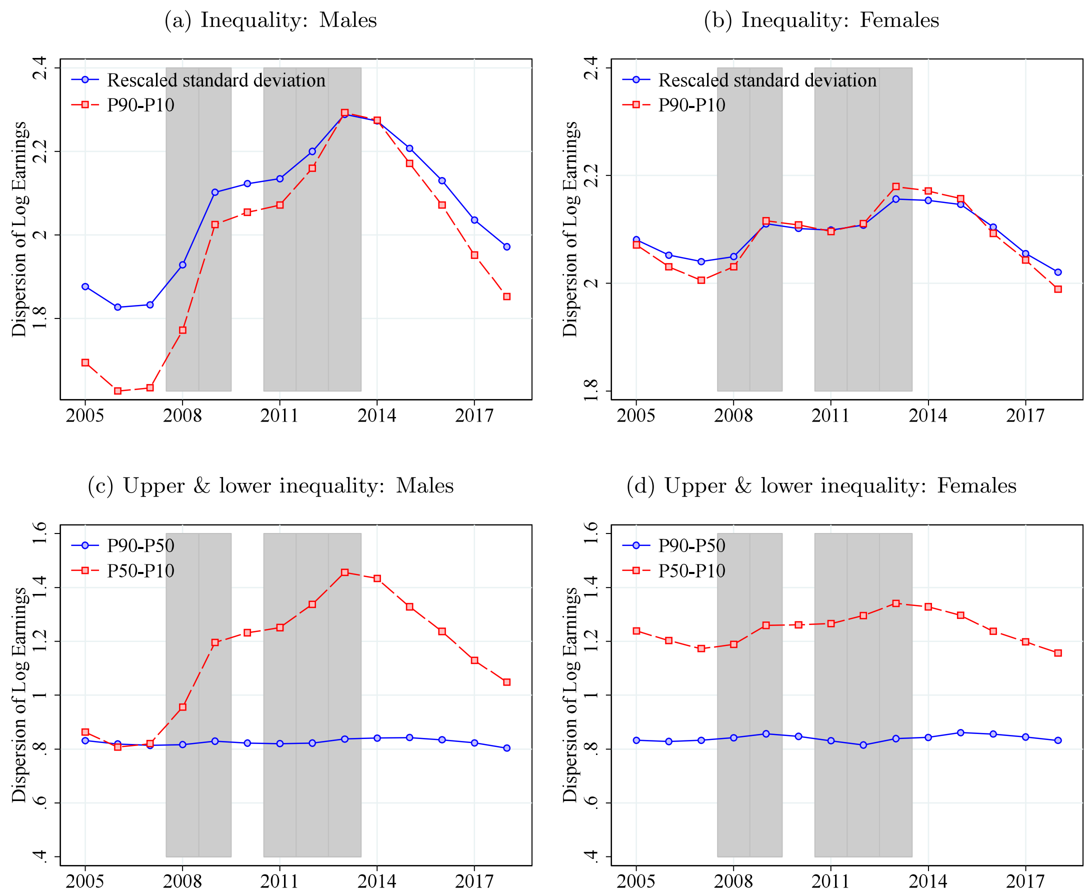


Notes: CS sample, percentiles of log annual earnings, by gender. All percentiles are normalized to 0 in 2005. The shaded areas indicate recession years.

the UK (Piketty and Saez, 2013).[6] 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 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.

[6]In Appendix Figures A2 and A3 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 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.

In Figure 3 we show various measures of inequality, by gender and over time.[7] 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 of inequality give a consistent message. For males, inequality increases substantially with

[7] In Appendix Figure A4 we show the evolution in the overall population, pooling both genders. In Online Appendix Figures S-A1 and S-A2 we show the results controlling for age, and for age and education, respectively.

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.[8]

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.[9]

3.2 Income changes

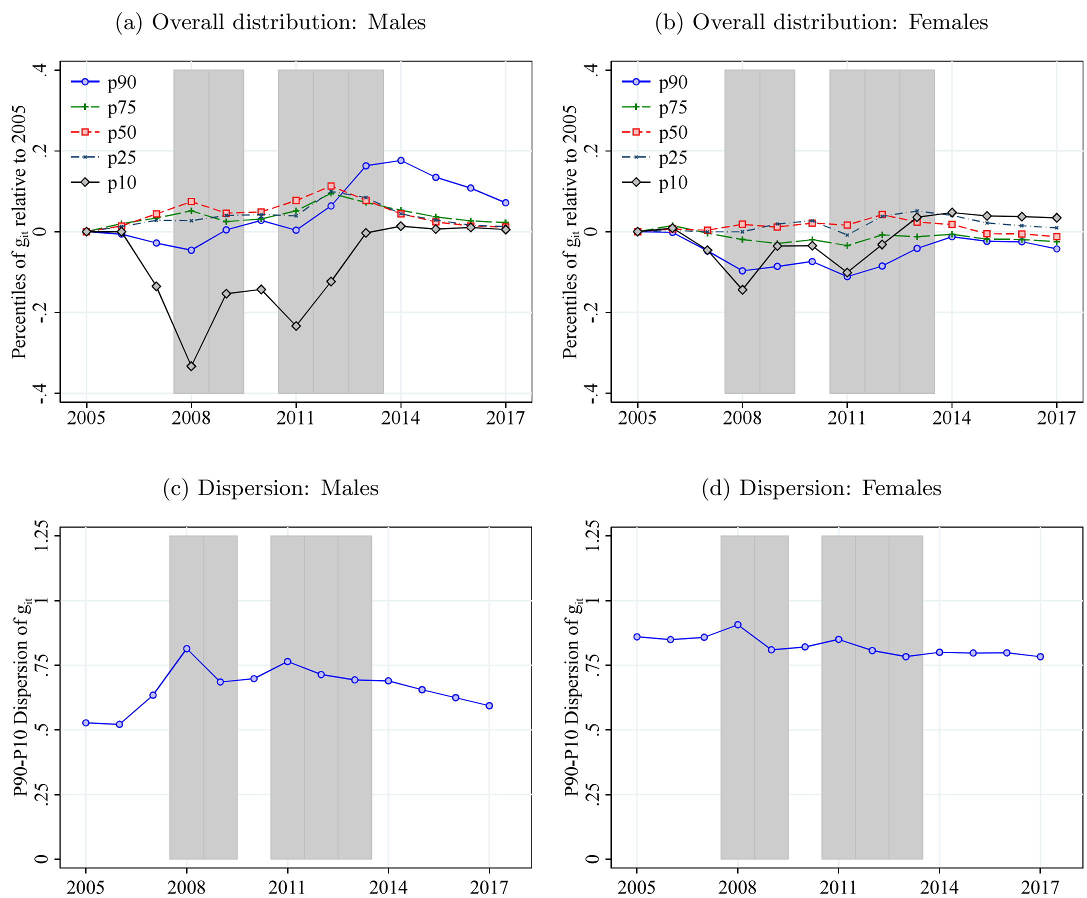
We next turn to the distribution of earnings changes and its evolution. For this purpose, we first focus on the LS sample, and construct residualized log earnings $\varepsilon_{it} = \log y_{it} - x_{it}'\widehat{\beta}$, where x_{it} includes fully-saturated interactions of age dummies, gender and year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

[8] In Appendix Figure A5 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.

[9] See Online Appendix Figure S-A3. Another notable aspect of the Spanish economy in this period is the increase in the percentage of immigrants. In Online Appendix Figure S-A4 we report earnings percentiles and inequality in a sample without immigrants, and find similar results to the ones based on the sample with immigrants.

We will also refer to multiple-year changes such as $g_{it}^5 = \Delta^5 \varepsilon_{it} = \varepsilon_{it+5} - \varepsilon_{it}$.

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.

In Figure 4 we start by documenting the evolution over time of percentiles of one-year log earnings changes.[10] 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, the 10th percentile decreases sharply around the recession. Moreover, as the comparison to

[10] In Appendix Figure A6 we show the densities of one-year and five-year log annual earnings changes, respectively. In Appendix Figure A7 we show the corresponding log densities.

the right graph shows, this evolution is not as pronounced for females.[11] In the lower panel of Figure 4 we show the P90-P10 percentile difference of log earnings changes.[12] We find that the dispersion of log earnings changes increases at the beginning of the recession, especially for males.[13]

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.[14]

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., 2022). 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. (2022) 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.[15]

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

[11] In Appendix Figure A8 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. In Appendix Figure A9, we show results pooling both genders together.

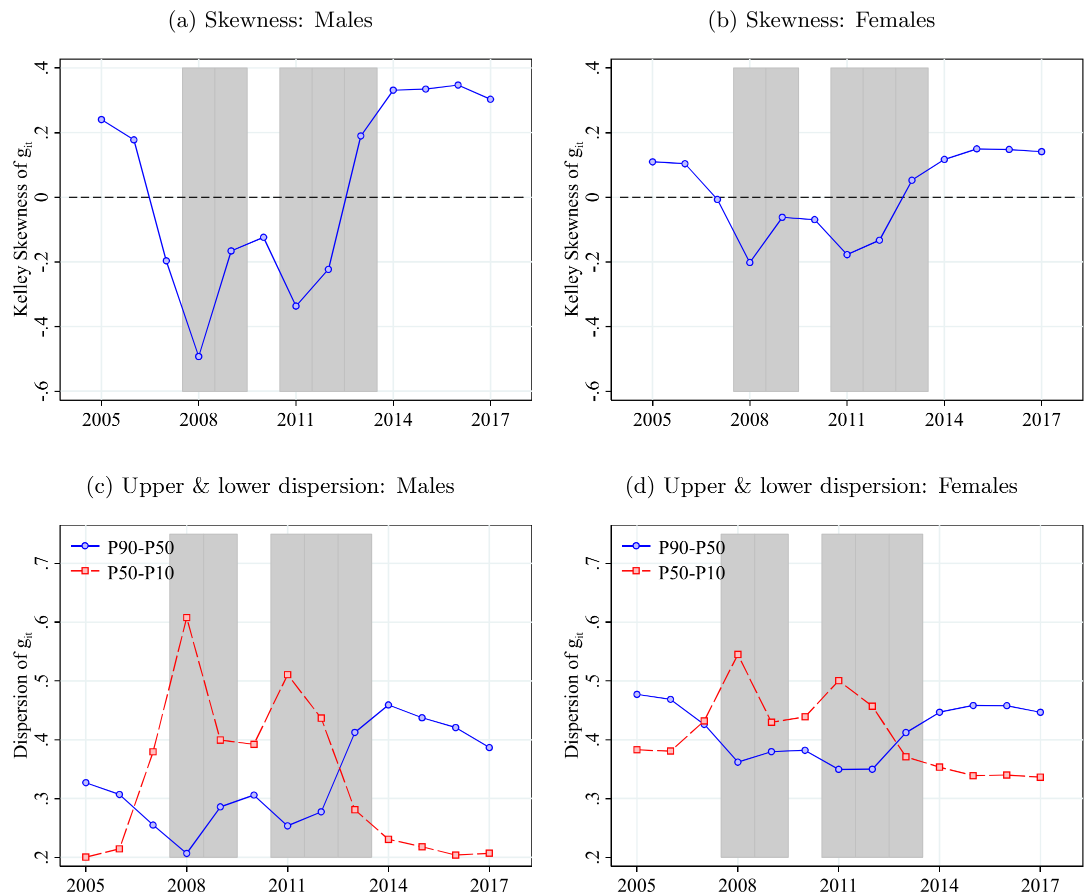
[12] In Appendix Figure A11 we also show rescaled standard deviation.

[13] In Appendix Figure A10 we document the dispersion of five-year log earnings changes.

[14] In Online Appendix Figure S-A5 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.

[15] In Appendix Figure A11 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. Skewness and excess kurtosis of five-year income changes are reported in Appendix Figure A12.

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

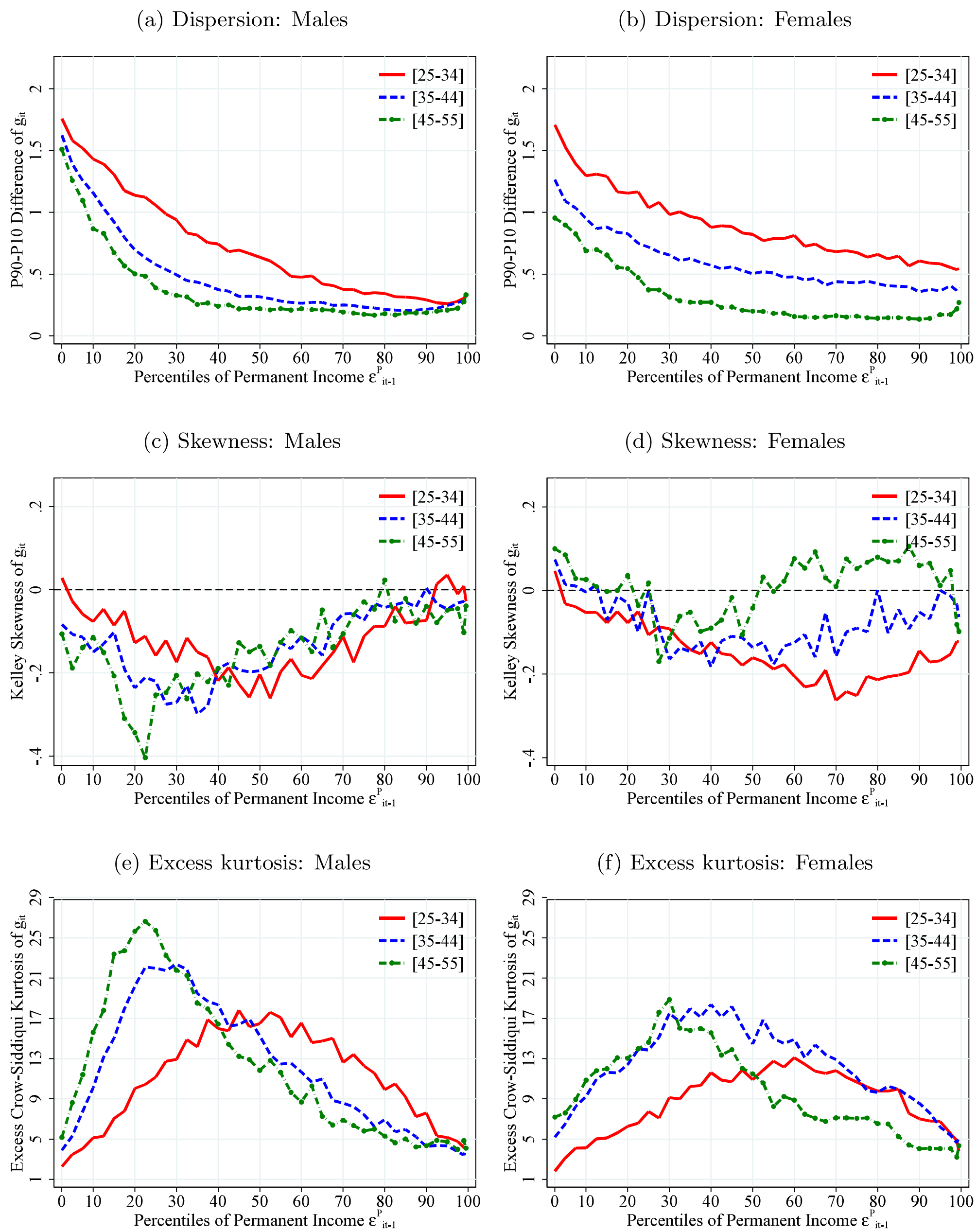


Notes: LS sample, one-year changes in residualized log earnings. Kelley skewness is $\frac{P90-2P50+P10}{P90-P10}$. *The shaded areas indicate recession years.*

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. (2021) 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. (2021), 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

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



Notes: H sample, one-year changes in residualized log earnings, data pooling 2008–2013. 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.

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.[16] 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 permanent 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 dispersion 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 such a 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 ε_{it}^{P}. 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

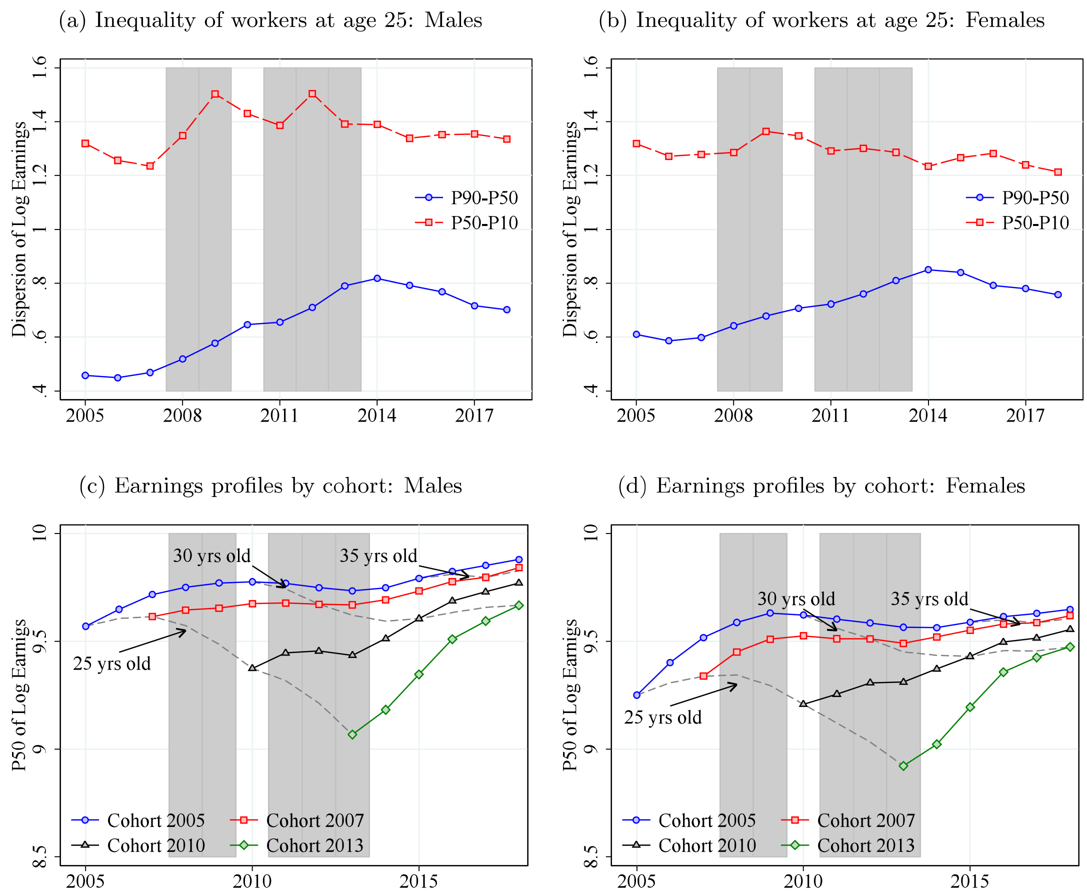
We next focus on inequality by cohort and age groups, and on earnings persistence and mobility. In the upper panel of Figure 7 we report the P90-P10, P90-P50, and P50-P10 percentile differences at age 25, by gender, from 2005 to 2018.[17] The results show that, for both genders, inequality in the upper 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

[16]Results for the overall population, pooling both genders, are reported in Appendix Figure A13. In Appendix Figure A14 we report the corresponding moment-based measures of dispersion, skewness, and kurtosis. In Appendix Figures A15 and A16, we report results on the conditional dispersion, skewness and kurtosis of five-year log earnings changes.

[17]In Appendix Figure A17 we show percentiles of log annual earnings at age 25. Additionally, Appendix Figure A18 report results pooling both genders.

Figure 7: Inequality and age profiles for young workers



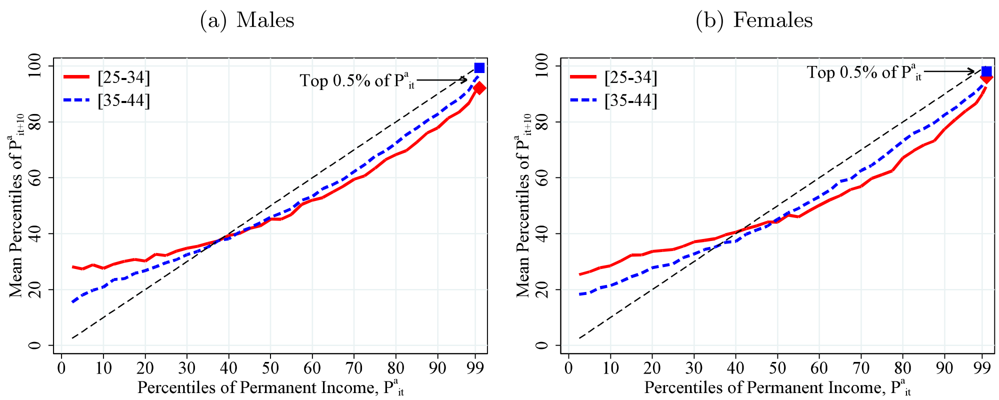
Notes: CS sample, log annual earnings. In the top panel the sample is restricted to age-25 workers only. In the bottom panel the different curves correspond to different cohorts of workers. The shaded areas indicate recession years.

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.[18]

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

[18] In Appendix Figure A19 we show earnings inequality.

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.

survey data. Here we report simple measures of earnings mobility based on our administrative 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.[19]

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

[19]In Appendix Figures A20 and A21, we show additional results on mobility over the life cycle, over time, and at a 10-year horizon. In Appendix Figure A22, we report 10-year average rank-rank mobility for the two age groups pooling both genders.

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):

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

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 $\pm 2,000$ 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 Online Appendix S-B 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

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

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

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

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.[20] 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.

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

In the first exercise, we quantify the 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.

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 risk measure, using regressions. Specifically, we regress $\widehat{\mathrm{CV}}_{it}$ 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, all the other predictors (i.e., the macro indicators, the features of the labor contract, and the income level) all have relatively low explanatory power. The

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.

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.

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

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 $\widehat{\mathrm{CV}}_{it}$ 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.

of income risk inequality.[21] 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.

Table 3: Income risk over the period, in numbers

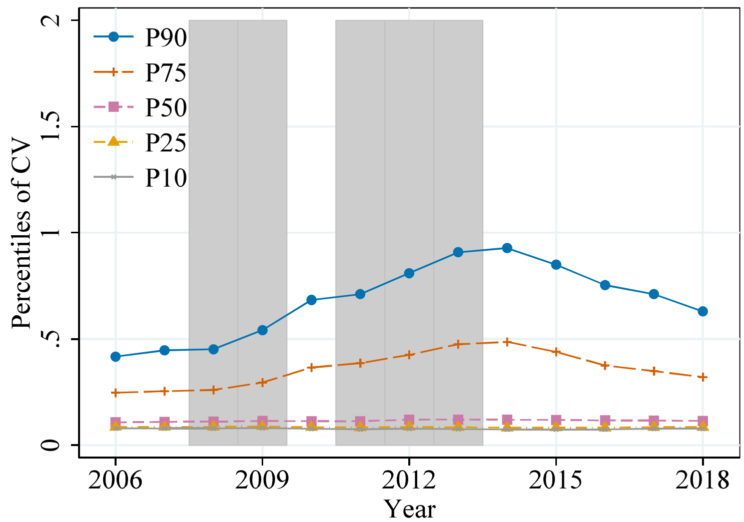
	All	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
P90/P10	8.94	5.26	5.69	5.76	6.70	8.90	9.50	10.42	11.97	12.58	11.43	10.15	9.15	8.09
P90/P50	6.03	3.87	4.07	4.06	4.74	6.05	6.30	6.75	7.53	7.75	7.15	6.48	6.13	5.51
P50/P10	1.48	1.36	1.40	1.42	1.41	1.47	1.51	1.55	1.59	1.62	1.60	1.57	1.49	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.

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 risk inequality — as measured by the P90/P10 ratio — increases substantially during the

[21] In Online Appendix Table S-C1 we report the coefficient estimates that we use to construct our CV.

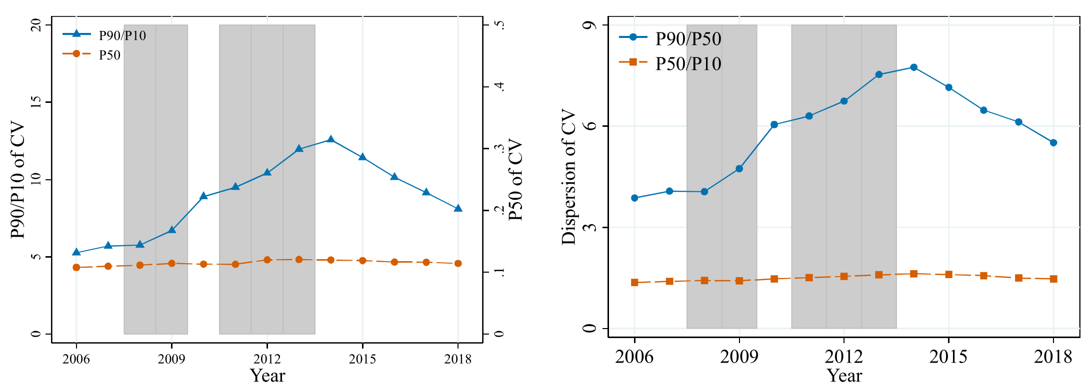
Figure 9: Income risk over the period, percentiles of CV



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

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.

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.

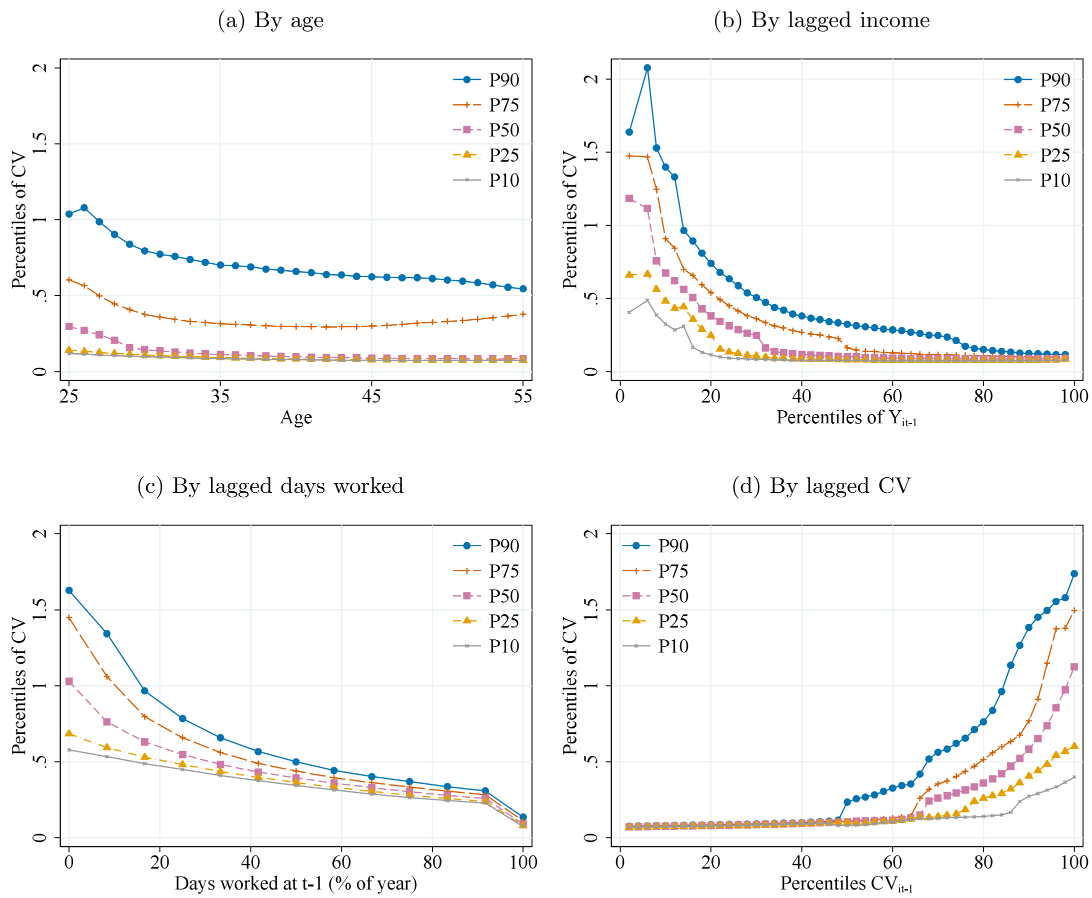
In order to illustrate the magnitude of the life-cycle variation in income risk, in Table 4 we report the ratio of age-specific percentiles of CV to the unconditional percentiles, by age. For example, the third row shows that, at the median, 25-year-olds experience almost three times as much risk — as 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.[22] Compared to the

[22] In Online Appendix Tables S-A5 and S-A6, 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.

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 standard deviation, the density of CV is more skewed to the right and has a larger mass of observations near zero.[23]

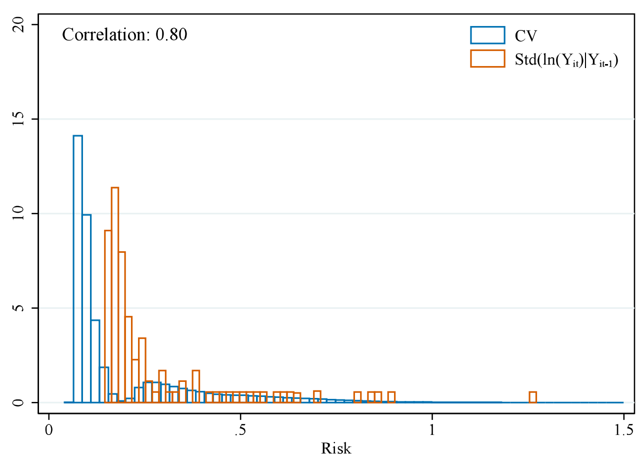
[23] In Appendix Figure C1 we compare various conditional percentiles of CV with the conditional standard deviation $\mathrm{Std}(\log(Y_{it})|Y_{it-1})$, as a function of lagged income percentiles. The conditional standard deviation of log income suggests a level of risk that is close to the 90th percentile 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.

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 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.[24] This suggests that more than half of 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. (2022). 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.[25]

5.4 The income risk for civil servants

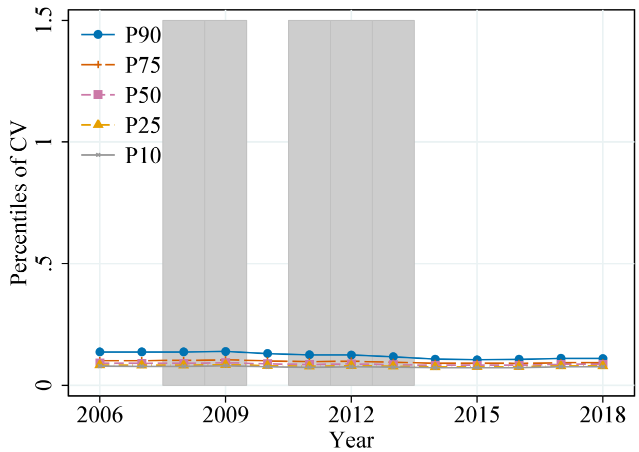
In this subsection we consider a particular category of workers, civil servants (*funcionarios*), as a convenient test case of the ability of our administrative records-based CV measure to correctly represent the risk individuals face. In Spain, as in other countries, civil servants are known to enjoy 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.[26] The income risk levels we find are low

[24] In addition, risk persistence tends to increase with age, as we show in Appendix Figure C2.

[25] We use this approach to document the level and evolution of the skewness of the predictive income distribution in Appendix Figure B7.

[26] In Appendix Table C1 we report the corresponding numbers. Note that we do not include the civil servant indicator as a predictor.

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.

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 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:

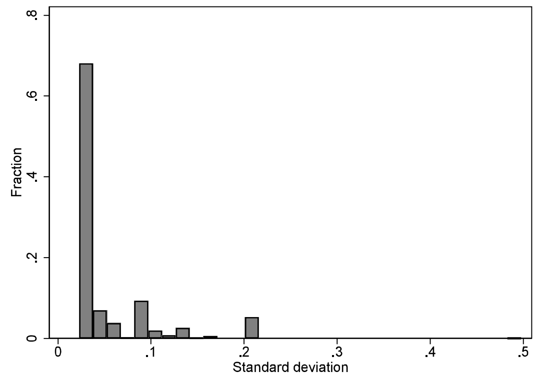
We are interested in knowing how you think the total annual income of your household will change in the next 12 months. Divide 10 points among the five options given below, assigning more points to the options you think are more likely (assign 0 point to options you think are impossible):

- *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%*

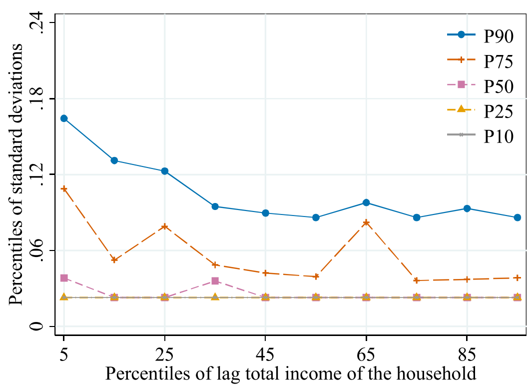
Thus, for every person who answered this question, we observe the fraction of points $\widehat{p}_j$ allocated to each event $j = 1, \dots, J$ (adding up to 1), where $J = 5$. From that information, we calculate summary measures of dispersion under the assumption that the underlying probabilities are normally distributed. We provide the details of the method in Online Appendix S-F. Let us define $\tilde{c}_j = \frac{\sum_{k=1}^{j} \widehat{p}_k + \frac{j}{2m}}{1+\frac{J}{2m}}$ 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

Figure 14: Estimated subjective standard deviations

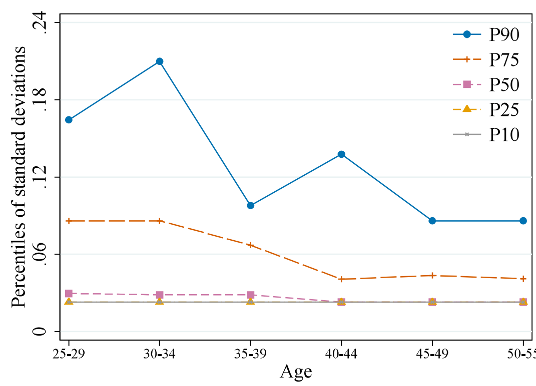
(a) Histogram



(b) By income



(c) By age



Notes: Estimated subjective standard deviations from the EFF 2014.

standard deviation estimates

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

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 et al. (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.

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 (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.

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.

Atkinson, Anthony 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 (2022) "Discretizing Unobserved Heterogeneity," *Econometrica*, 90 (2), 625–643, 10.3982/ECTA15238.

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.

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.

Busch, Christopher, David Domeij, Fatih Guvenen, and Rocio Madera (2022) "Skewed Idiosyncratic Income Risk over the Business Cycle: Sources and Insurance," *American Economic Journal: Macroeconomics*, 14 (2), 207–42, 10.1257/mac.20190019.

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.

García-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.

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, `https://mitpress.mit.edu/books/deep-learning`.

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 (2021) "What Do Data on Millions of U.S. Workers Reveal About Lifecycle Earnings Dynamics?" *Econometrica*, 89 (5), 2303–2339, https://doi.org/10.3982/ECTA14603.

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`.

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), 387–392, 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, Jonathan and Rachel 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 June 2010.