Income Risk Inequality: Evidence from Spanish Administrative Records

Manuel Arellano[1] Stéphane Bonhomme[2] Micole De Vera[1] Laura Hospido[3]
Siqi Wei[1]

[1]CEMFI

[2]University of Chicago

[3]Banco de España and IZA

16 December 2021
SAEe 2021

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.

Introduction

- **Motivation:** understanding individual income risk has been a central motivation of the income dynamics literature and is a key determinant of economic decisions
- **Goal:** develop a methodology for constructing measures of individual income risk and measuring inequality
- **Methodology:** prediction-based approach, leveraging availability of administrative records of employment histories
- **Alternative approaches:**
  - Statistical models of the dynamics of income with a small state space, often in combinations with models of choice
  - Recent nonparametric approach to income dynamics in the spirit of Guvenen et al. (2021) produces statistics such as the conditional moments of income growth that are related to income risk, yet this approach does not target risk directly
  - Subjective probabilistic expectations from surveys (Dominitz and Manski, 1997)

Our focus: Inequality of individual income risk

- Salient features of the Spanish labor market have been the high level of unemployment and it large cyclical fluctuations
- Inequality in income risk is related to unemployment, but also the large share of short-term temporary employment
- We find that income risk is highly unequal in Spain: more than half of the economy has close to perfect predictability of their income while others face considerable uncertainty

Outline

Outline

Linked social security, tax and census records (MCVL)

- Panel data obtained by matching social security employment histories with income tax and census records for a 4% sample of social security affiliates from 2005–2018
- **Sample selection:** Ages 25–55, no self-employment, no Basque country or Navarre, males (for this presentation)
- **Income concepts:**
  - Labor income: individual income from paid employment in a calendar year, as reported by employers to the tax authority. For statistics on income inequality and dynamics,
    - Residualized earnings net of age dummies by year and gender
    - Trim earnings below a threshold (working part time for one quarter at the minimum wage)
  - Extended measure: more comprehensive and includes
    1. observations below the threshold including zeros
    2. unemployment benefits

Outline

GID: mebdi.org/global-income-dynamics-project

- **Goal:** build open-access, cross-country harmonized database of rich micro statistics on earnings inequality and earnings dynamics from administrative panel data
- Initiative by Fatih Guvenen (Minnesota), Luigi Pistaferri (Stanford), and Gianluca Violante (Princeton)
- Database scheduled for launch in 2022

- **Argentina:** "The Evolution of the Earnings Distribution in a Volatile Economy: Evidence from Argentina," Blanco, Diaz de Astarloa, Drenik, Moser, and Trupkin.
- **Brazil:** "Earnings Inequality and Dynamics in the Presence of Informality: The Case of Brazil" by Engbom, Gonzaga, Moser, and Olivieri.
- **Canada:** "Four Decades of Canadian Earnings Inequality and Dynamics Across Workers and Firms", by Bowlus, Gouin-Bonenfant, Liu,Lochner, and Park.
- **Denmark**: "Trends in Income Risk in Denmark 1987-2016" by Leth-Petersen and Saeverud
- **France:** "Inequality and Earnings Dynamics in France: National Policies and Local Consequences", by Kramarz, Nimier-David, and Delemotte.
- **Germany:** "Inequality and Income Dynamics in Germany", by Drechsel-Grau, Peichl, Schmieder, Schmid, Walz, and Wolter.
- **Italy:** "Labor Reforms and Earnings Dynamics: The Italian Case", by Hoffman, Malacrino, and Pistaferri.
- **Mexico:** "Income Dynamics and Inequality: The Case of Mexico", by Puggioni, Calderón, Cebreros Zurita, Fernandez Bujanda, Gonzalez, and Jaume.
- **Norway:** "Earnings Dynamics and Its Intergenerational Transmission: Evidence from Norway", by Halvorsen, Ozkan, and Salgado.
- **United States:** "U.S. Long-Term Earnings Outcomes by Sex, Race, Ethnicity, and Place of Birth", by Abowd, McKinney, and Janicki.
- **United Kingdom:** "Income Dynamics in the United Kingdom 1975-2020", by Bell, Bloom, and Blundell.
- **Spain:** "Income Risk Inequality: Evidence from Spanish Administrative Record", by Arellano, Bonhomme, De Vera, Hospido, and Wei.
- **Sweden:** "Income Dynamics in Sweden 1985-2016," by Friedrich, Laun, and Meghir.

Brief summary of selected results

- **Income inequality**
  - Inequality in earnings increased in the recession, driven mostly by inequality in the bottom part of the earnings distribution (Bonhomme and Hospido, 2017)
  - Cohorts starting during the recession have an initial lower earnings distribution but a subsequent steeper profile
- **Income changes**
  - Dispersion in income changes increases during the recession (Storesletten et al., 2004; Guvenen et al., 2014)
  - Skewness in income changes becomes more negative in the recession consistent with procyclicality of skewness of income growth well-documented in several countries (Busch et al., forthcoming; Hoffmann and Malacrino, 2019; Pora and Wilner, 2020)
  - Dispersion and skewness of future income depends on past income suggesting important nonlinearities in earnings dynamics (Arellano et al., 2017; Guvenen et al., 2021)

Outline

Outline

Income Prediction and Income Risk

- **Main idea**:
  - We mimic the agent's prediction problem as closely as we can (in the absence of expectations data)
  - Income risk faced an agent refers to features of this predictive distribution of income (in this paper, we mainly focus on the dispersion)
- Target the distribution of income level Y_{it} given predictors X_{it}:
  - Micro predictors: past income, past employment, labor contract types, demographics
  - Macro predictors: GDP growth rate and unemployment rate (national and provincial)

A coefficient of variation (CV) measure of income risk

- **Main measure:** Coefficient of variation, computed as the ratio of the MAD of income divided by the mean income, conditional on X_{it}

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

- An individual with an expected income of 20,000 euros and a CV of 10% expects a deviation of her next year's income from its mean of $\pm$ 2,000 euros
- Notes:
  - The choice of MAD instead of the SD in the numerator is to minimize sensitivity to extreme observations
  - When CV is small, it can be approximated by the standard deviation of log income, $CV(X_{it}) \approx \sqrt{\frac{2}{\pi}}\text{Std}(\log Y_{it} \mid X_{it})$. However, the CV remains well-defined when $Y_{it} = 0$.

Econometric approach

- Estimating the numerator and the denominator of the CV are two prediction tasks
- Since $Y_{it} \geq 0$ and $|Y_{it} - \mathbb{E}(Y_{it} \mid X_{it})| \geq 0$, a natural parametric estimator is based on **exponential specifications**:

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

- We estimate these two quantities using two **Poisson regressions** and report the ratio

Discussion

Welfare interpretation to CV

- Following Lucas (1987), an approximation for the welfare gain from eliminating consumption risk is given by

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

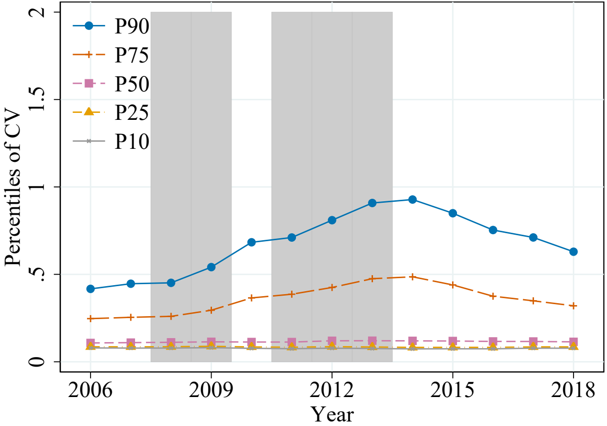
where θ_i is individual coefficient of risk aversion

Considerations

- Log-normality of income may not be a good approximation $\Rightarrow$ higher-order moments of the predictive income distribution such as skewness or kurtosis will matter
- Welfare interpretation depends on specified utility function: under Stone-Geary utility where there is a subsistence consumption level, inequality in CV will underestimate the degree of inequality in the economic costs of risk

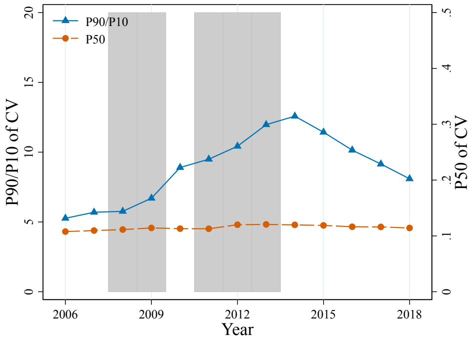
Outline

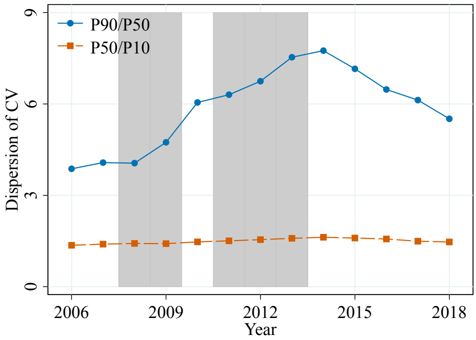
Distribution of CV over time



- At least half of the Spanish economy at any time faces little uncertainty in future income
- Distribution of CV has a long right tail which lengthens in the recession

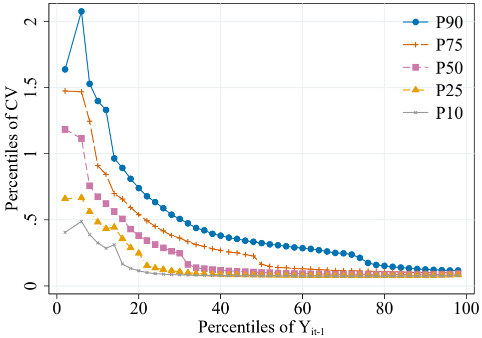
Inequality of income risk



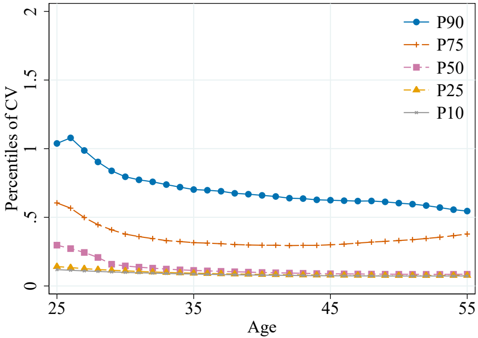


- Inequality in income risk increased in the recession
- Changes driven by the top of income risk distribution

Correlates of income risk



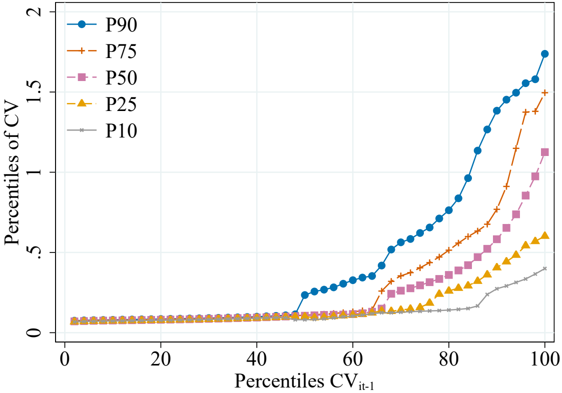
(a) By lagged income



(b) By age

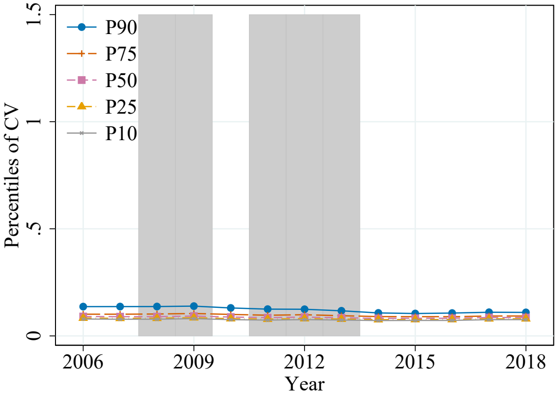
- High income individuals face low levels and small dispersion of income risk in contrast to individuals at the bottom of the income distribution
- Younger individuals (less than 30) tend to face higher levels of income risk and larger income risk dispersion

Persistence of income risk



- Income risk is highly persistent over time, especially for low income risk levels
- More than half of the Spanish economy is effectively shielded from income risk: bottom half of the risk distribution today face virtually no risk next year

Income risk of civil servants



- Civil servants are known to enjoy high levels of job and income security
- CV levels are low compared to rest of economy: 90th percentile comparable to median of overall distribution
- CV distribution virtually unaffected by the recession

Outline

Robustness checks

Main Concerns

- Chosen predictors do not correspond to the agent's information set
- Prediction model is misspecified

Robustness

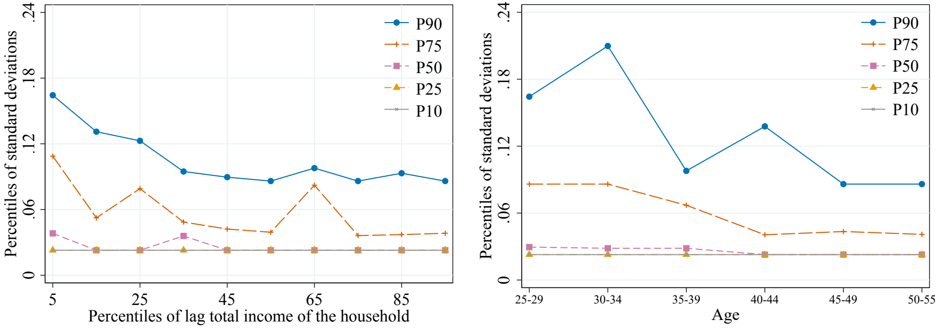
- Fixed discrete unobserved heterogeneity
- Flexible predictors using neural nets
- CV robust to outliers
- Quantile-based measure of income risk

Outline

Income risk: What do subjective expectations say?

- We compare our CV measure of income risk to a measure based on subjective expectations data
- Broad agreement between our prediction-based measure and the subjective expectations-based measure (in spite of measurement differences) increases our confidence in both measures
- We rely on subjective income expectations questions from the 2014 Spanish Survey of Household Finances (EFF)
- Assuming a household-specific log normal random walk predictive income process, we estimate **subjective standard deviations of income growth**

Subjective standard deviations: Main results



(a) By income

(b) By age

- Many households face low levels of risk but there is substantial dispersion between households
- Subjective standard deviations tend to be higher for households with low income and the young

Outline

Conclusions

- We have documented a number of new empirical facts on income inequality and income dynamics in Spain
- We developed a methodology for constructing measures of individual income risk and measuring income risk inequality:
  - Evidence of high inequality of income security in Spain
  - A large mass of workers with negligible income risk coexists with many who anticipate large fluctuations in future earnings
- Key patterns of individual income risk:
  - Income risk is more unequal and higher on average for the young
  - Inequality of income risk increased during the Great Recession
  - Income risk is more unequal and higher for those with lower income
  - Low levels of risk are more persistent than higher levels of risk

Other concluding remarks

- Some of the underlying causes of the inequality of income risk that we have documented are familiar to labor economists that have studied the Spanish labor market
- Our perspective abstracts from shorter term labor market transitions and puts the focus on the unequal security that individuals face at a relevant time horizon
- Open questions:
  - How are individual risks mitigated at the household level?
  - How much of income risk is exogenous to the agent and how much is a result of choice?
  - What are the macroeconomic consequences of the heterogeneous patterns of individual income risk?

THANK YOU!

Feel free to email me for any questions, comments or suggestions:

micole.devera@cemfi.edu.es

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, 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.

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

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

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*.

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.

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.

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.

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.

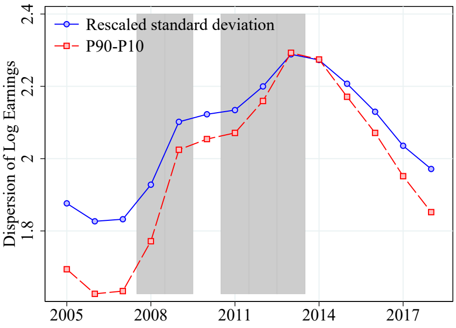
Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron (2004) "Cyclical Dynamics in Idiosyncratic Labor Market Risk," *Journal of Political Economy*, 112 (3), 695–717, http://www.jstor.org/stable/10.1086/383105.

More details on extended measure

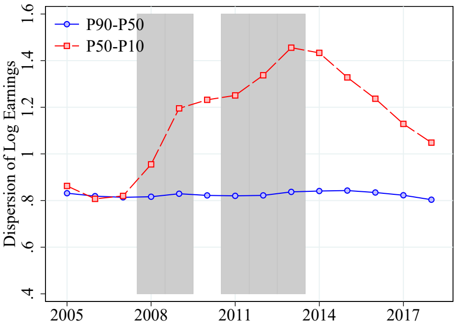
- The fraction of (male) observations below the threshold fluctuates between 3% in 2005 and 22% in 2013
- If no labor contract or UI benefits are observed in a full year, annual income is set to zero
- We exclude segments of more than two consecutive years without a contract

Back

Earnings inequality



(a)

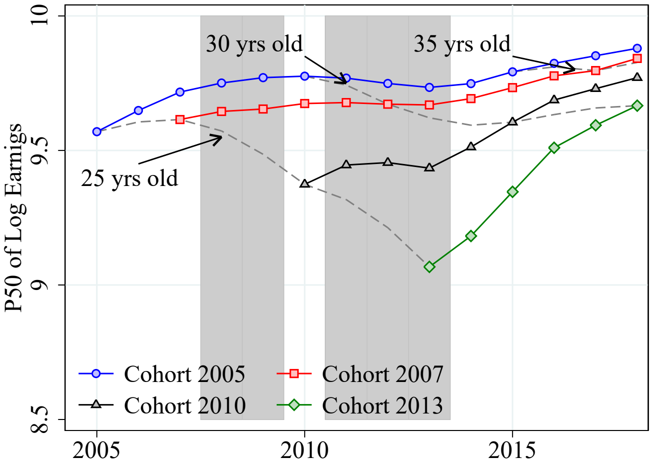


(b)

- Both P90-P10 and σ inequality increased during the recession
- Lower inequality increases around the recession while upper inequality stays flat

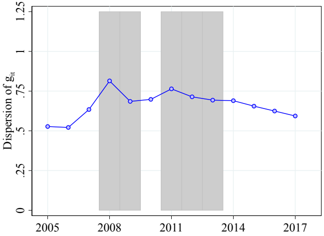
Back

Median earnings profiles for young workers

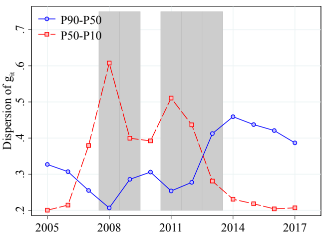


Back

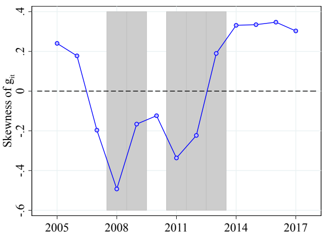
Dispersion and skewness of one-year log earnings changes



(a) Dispersion



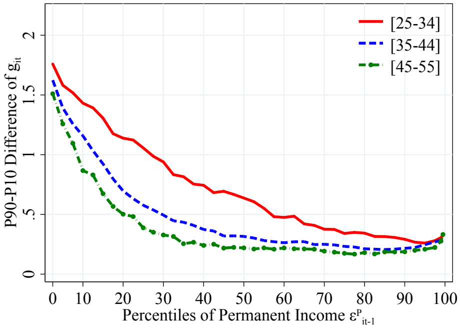
(b) Upper & Lower Dispersion



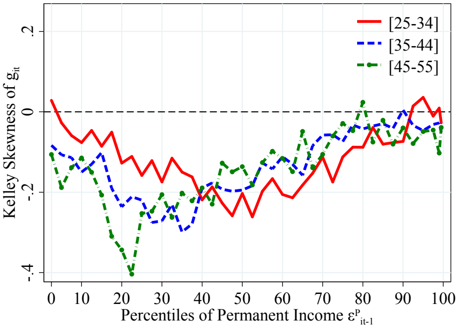
(c) Skewness

Back

Conditional dispersion and skewness of one-year log earnings changes



(a) Dispersion



(b) Skewness

Back

Welfare interpretation to CV: Stone-Geary preferences (1/2)

- Consider utility

$$U_i(C_{it}) = \frac{(C_{it} - C_m)^{1-\theta_i} - 1}{1 - \theta_i}$$

with $C_{it} = \lambda(X_{it})Y_{it}$ for proportionality factor $\lambda(X_{it})$

- Suppose $\ln(C_{it} - C_m) \mid X_{it} = x \sim \mathcal{N}(\mu(x), \sigma(x)^2)$
- Agent is willing to give up a% of consumption each period to eliminate income risk such that

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

- Then,

$$a \approx \frac{\pi}{4}\theta_i \frac{\mathbb{E}(C_{it}|X_{it})}{\mathbb{E}(C_{it}|X_{it}) - C_m} CV^2$$

Back

Welfare interpretation to CV: Stone-Geary preferences (2/2)

$$a \approx \frac{\pi}{4}\theta_i \frac{\mathbb{E}(C_{it}|X_{it})}{\mathbb{E}(C_{it}|X_{it}) - C_m} CV^2$$

- With $C_m = 0$ this simplifies to the approximate welfare gains under CRRA utility
- For individuals whose consumption is close to the subsistence level (i.e., $C_m/\mathbb{E}(C_{it}|X_{it})$ is non-negligible), the squared CV underestimates the welfare cost of income risk
- We find risk and income are inversely correlated so a CV-related measure will tend to underestimate the degree of inequality in the welfare cost of risk

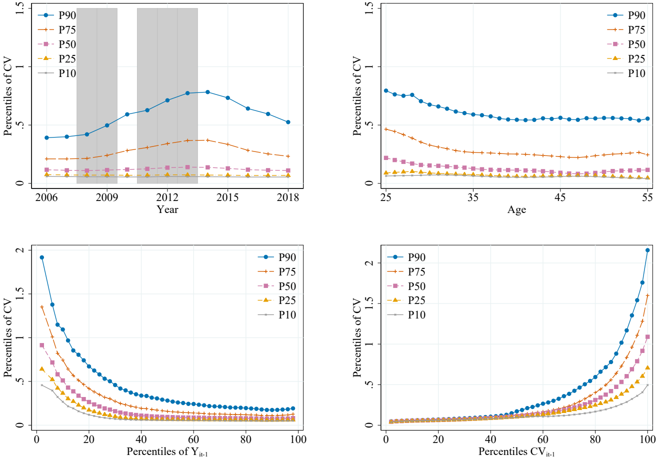
Back

CV with unobserved heterogeneity (1/2)

- Augment the predictor set as (X_{it}, ξ_i) where ξ_i is a latent component
- Standard RE or FE approaches are intractable
- **Proposal:** Grouped fixed effects (GFE) estimator as in Bonhomme et al. (2021)
  - Two-step: (1) classify individuals by k-means, and (2) estimating model allowing for group-specific heterogeneity
  - Can be extended to embed reclassification
  - Benefit of this approach is that we can handle incomplete models where we do not need to specify feedback processes or initial conditions
  - Option to allow specific unobserved heterogeneity components: one for conditional mean, one for conditional MAD

Back

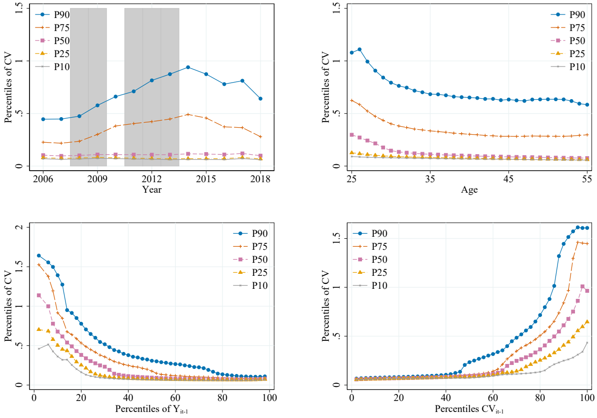
CV with unobserved heterogeneity (2/2)



- $K = 4$ for both numerator and denominator

Back

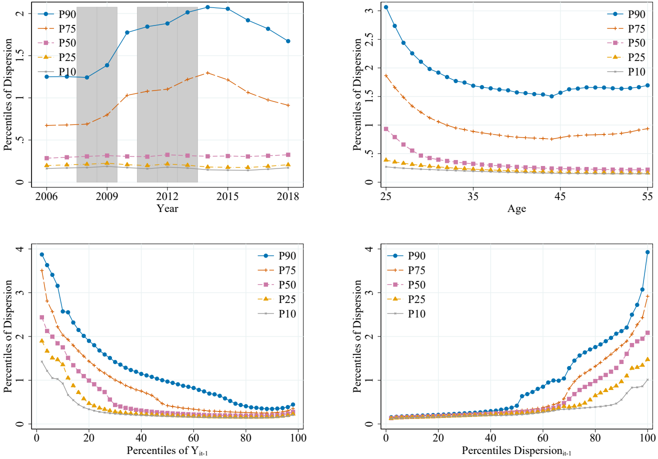
Estimating CV with neural nets



- Focus on 1-layer NN with ReLU link estimated with a Poisson loss function
- Number of nodes chosen with single-fold cross-validation ⇒ 8 nodes for denominator and 7 nodes for numerator

Back

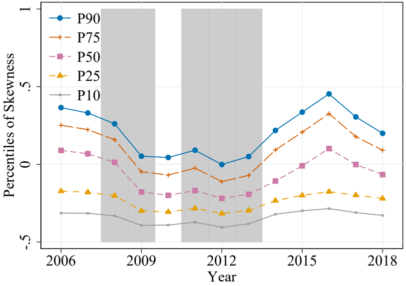
Quantile-based risk measures (1/2)



- Model conditional distribution of log-income using quantiles
- Report dispersion of the predictive distribution measures by $P90(X_{it}) - P10(X_{it})$

Back

Quantile-based risk measures (2/2)

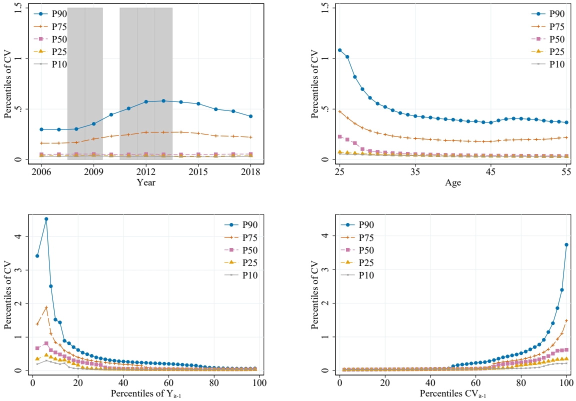


- By modeling the entire predictive distribution, we are able to document more than dispersion and location
- Kelley skewness:
$$\frac{P90(X_{it}) - 2P50(X_{it}) + P10(X_{it})}{P90(X_{it}) - P10(X_{it})}$$
- Skewness decreases in recessions, driven by less individuals with higher skewness

Back

Robust CV (Arachchige et al., 2020)

$$\widetilde{CV}(X_{it}) = \frac{\text{median}\left(|Y_{it} - \text{median}(Y_{it} \mid X_{it})| \mid X_{it}\right)}{\text{median}(Y_{it} \mid X_{it})}$$



Back

Subjective probabilistic income expectation question

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 below, assigning more points to the options you think are more likely:

- 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%

Back

Estimating subjective standard deviations (1/2)

- For every person, we observe the number of points allocated to each of the five events, $\widehat{p}_j$: $\Delta y < -10\%, -10\% < \Delta y < -2\%, -2\% < \Delta y < 2\%, 2\% < \Delta y < 10\%, \Delta y > 10\%$
- Assume log income follows a RW with (i) HH-specific drift μ and (ii) Gaussian shocks with HH-specific SD σ
- Interpret elicited probabilities $\widehat{p}_j$ as noisy measurements of p_j due to rounding and inherent randomness in elicitation
- If $\widehat{p}_j$ are regarded as sample frequencies from a hypothetical random sample of size m, they are unrestricted MLE of p_j
- We use an adjusted (posterior mean) estimator:

$$\widetilde{p}_j = \frac{\widehat{p}_j + (1/2m)}{1 + (J/2m)}$$

with the advantage $\widetilde{p}_j \in (0, 1)$ so the inverse normal cdf transformation is defined with $\widehat{p}_j = 0$ or 1

Back

Estimating subjective standard deviations (2/2)

- We implement a Berkson estimator that enforces the Gaussian restrictions on the posterior means $\widetilde{p}_j$
- This estimator is based on the inverse normal probabilities:

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

where:

- $\alpha = \mu/\sigma$
- $\beta = 1/\sigma$
- c_j are cumulative probabilities $c_j = \sum_{k=1}^{j} p_k$

Back