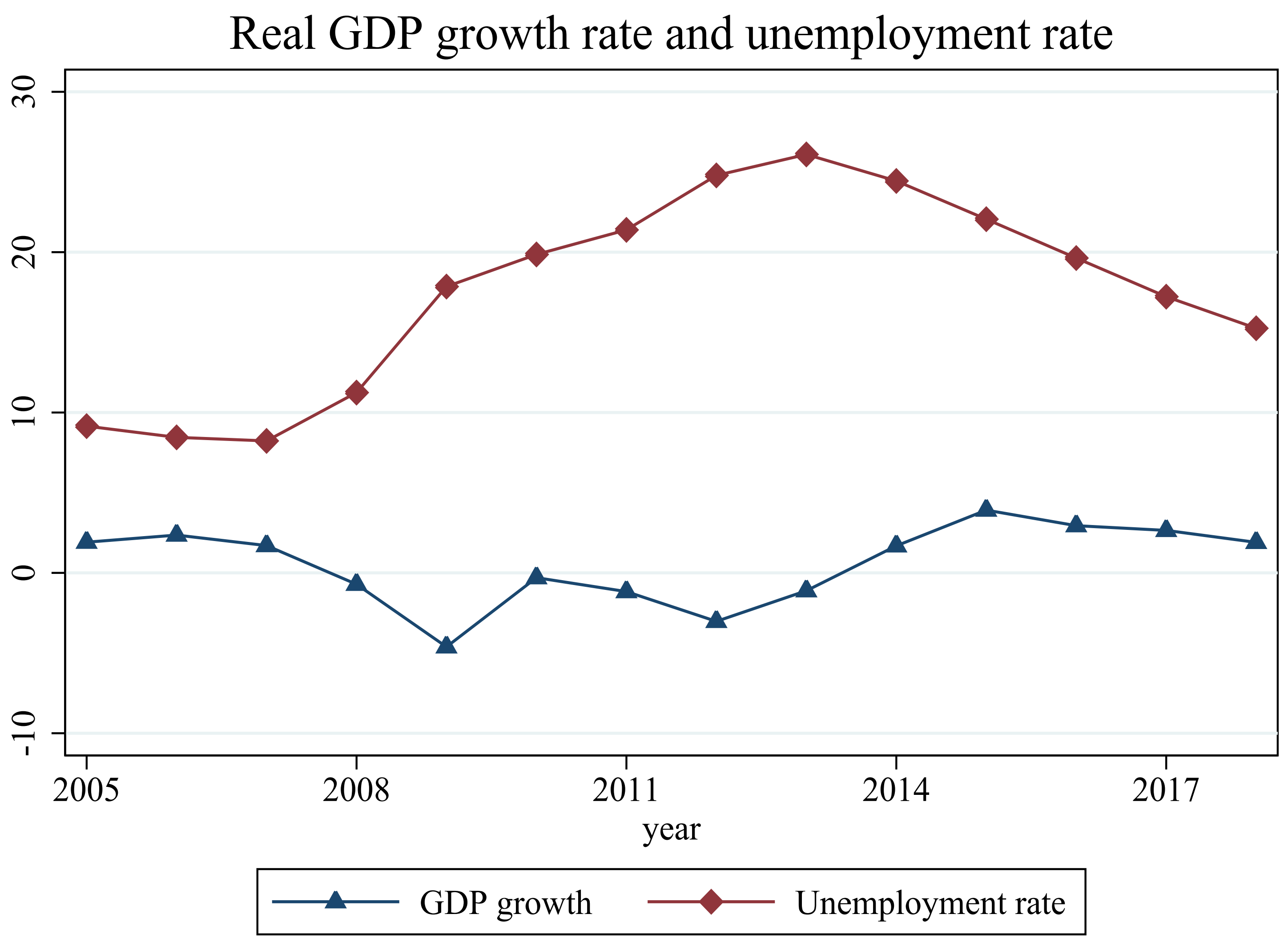
Income Risk Inequality

Evidence from Spanish Administrative Records

Manuel Arellano, Micole De Vera, Siqi Wei (CEMFI),
Laura Hospido (Banco de España), Stéphane Bonhomme (Chicago)

Global Income Dynamics Conference
November 19, 2020

Aggregate conditions in Spain



Our focus: inequality in income risk

- Salient features of the Spanish labor market have been the high level of unemployment rate and its large cyclical fluctuations.

- Also important but less well understood is the large cross-sectional inequality in individual income risk at given age and over the life-cycle.

- 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.

- We measure income risk and document the evolution of income risk inequality.

Income risk measures as predictors

• We construct individual measures of income risk as flexible functions of past employment history, income, demographics, and unobserved heterogeneity.

• Then we can study inequality in income risk, its persistence, and how it changes over the life and business cycles.

• The econometrics of measuring income risk is a prediction problem.

• In the absence of unobserved heterogeneity, standard regression and machine learning prediction techniques can be used.

• In the presence of unobserved heterogeneity we turn to a nonlinear grouped fixed-effects approach.

Part 1: A Refresher

Linked social security, tax, and census records

• We match social security employment histories with income tax and census records for a 4% sample of social security affiliates from 2005 to 2018 (Muestra Continua de Vidas Laborales, MCVL).

• We use individual and firm characteristics from social security records (age, gender, etc), and those matched from tax and census records (country of birth, education).

• Two main limitations:

-Relatively short period of observation.

-Not possible to link individuals into households.

Income

• We use individual income from paid employment accumulated in a calendar year, as reported by employers to the tax authority.

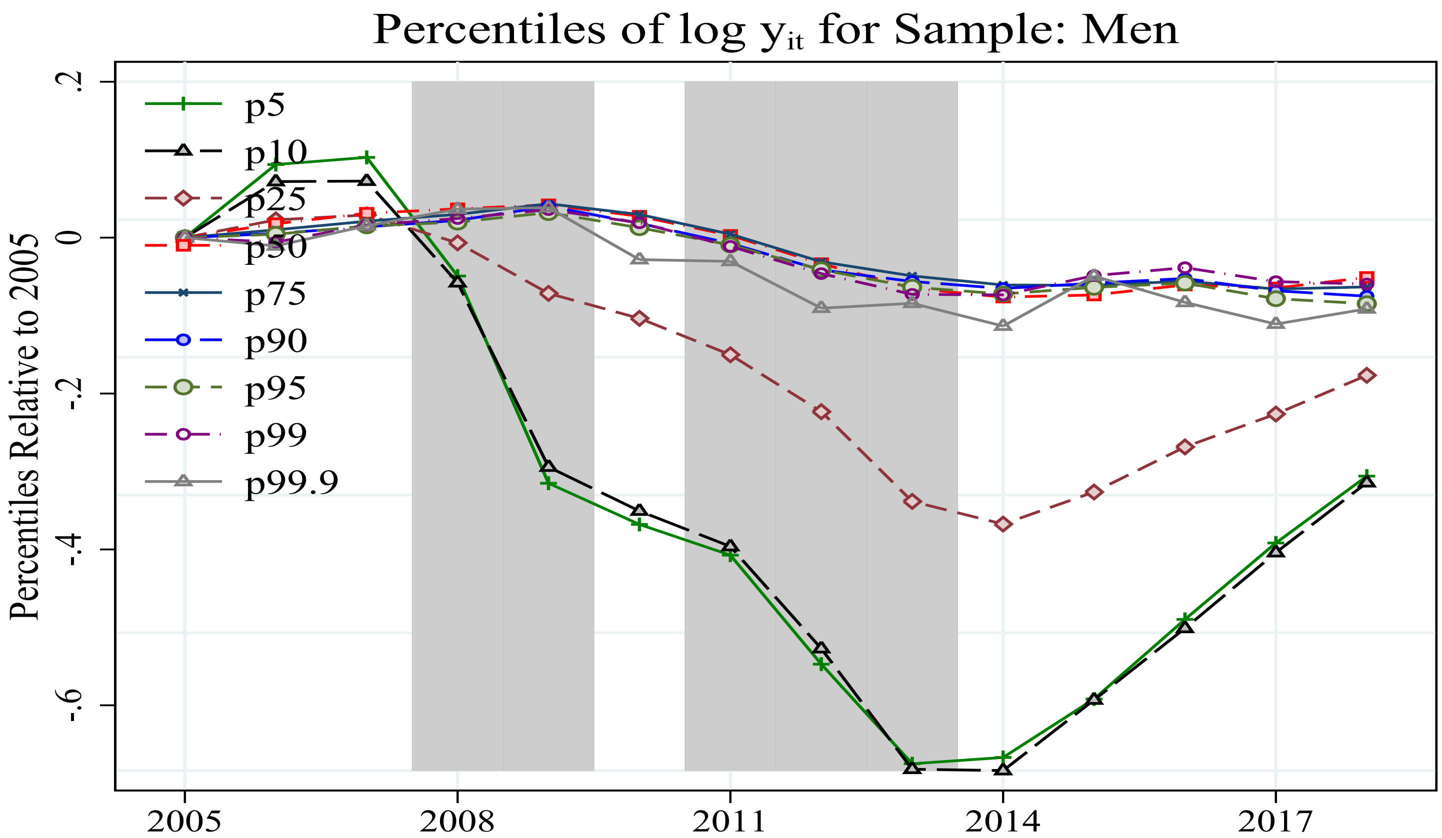
• Age restriction: 25-55, trimming annual earnings below some threshold (working part time for one quarter at the minimum wage).

• In Part 2 we use a broader measure of income without trimming in the calculation of individual income risk.

• In this presentation we focus on males.

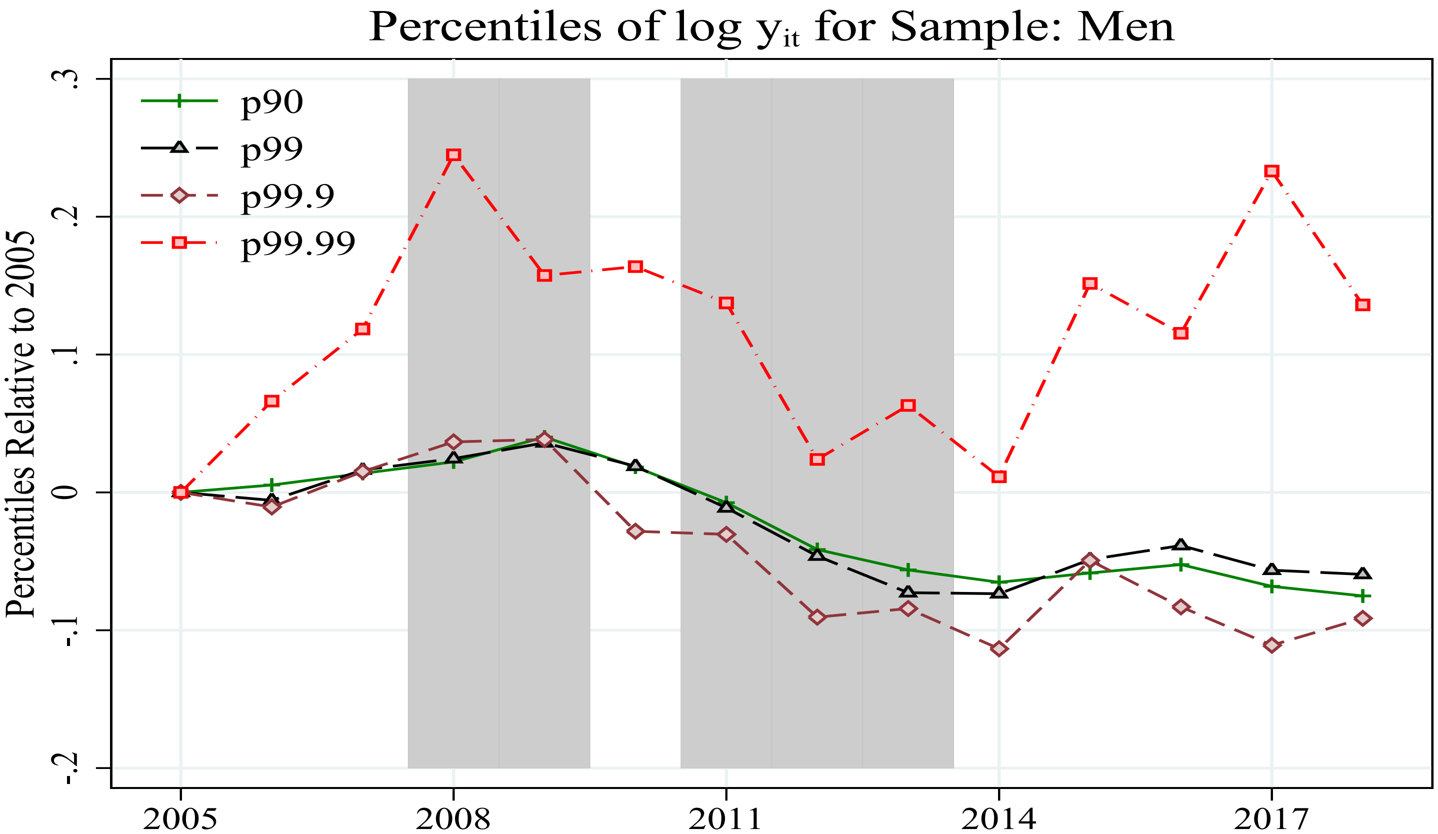
• To keep in mind in Part 1: the percentage of observations below the threshold is substantial and varies over the business cycle.

Income percentiles

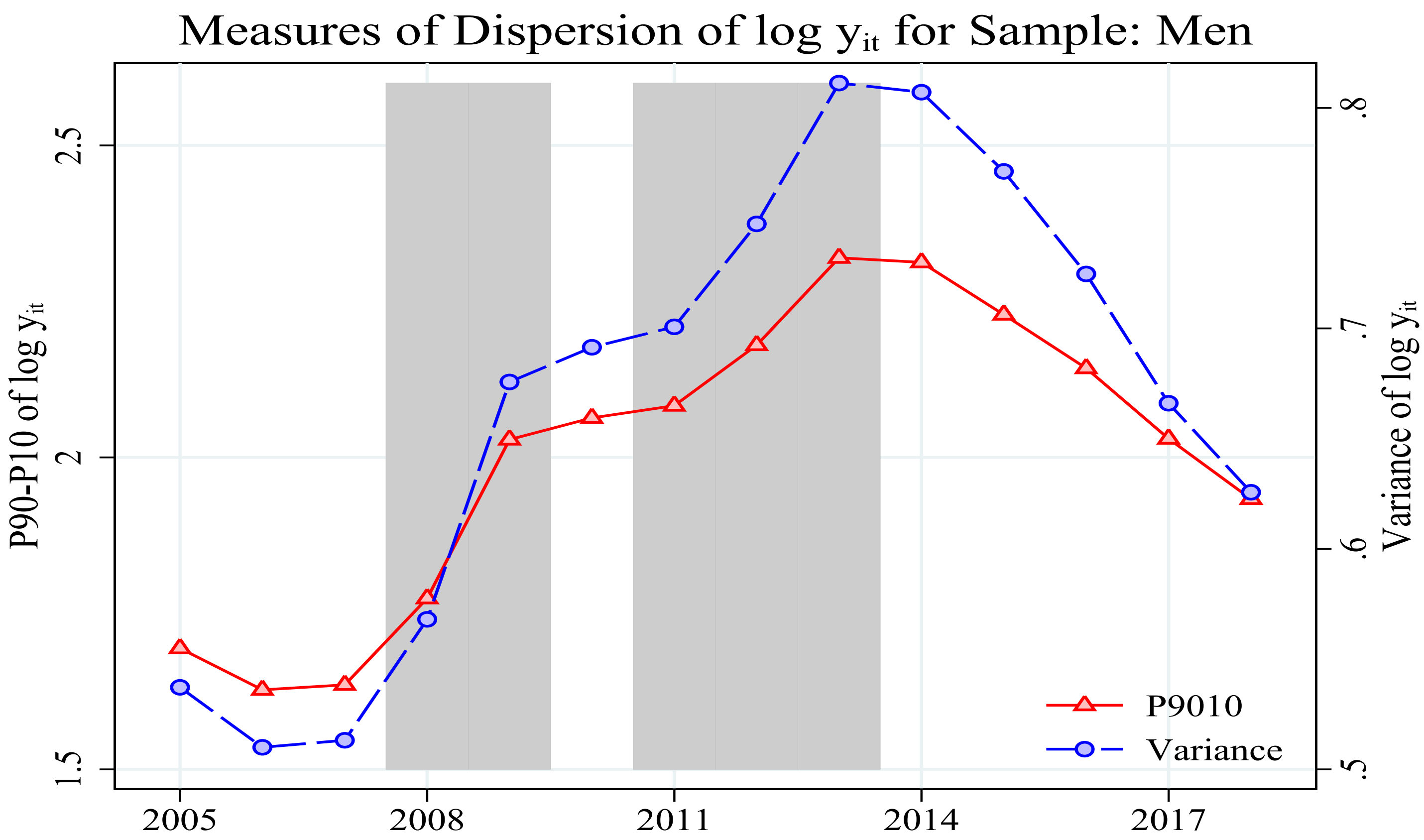


Notes: Recession months are shaded.

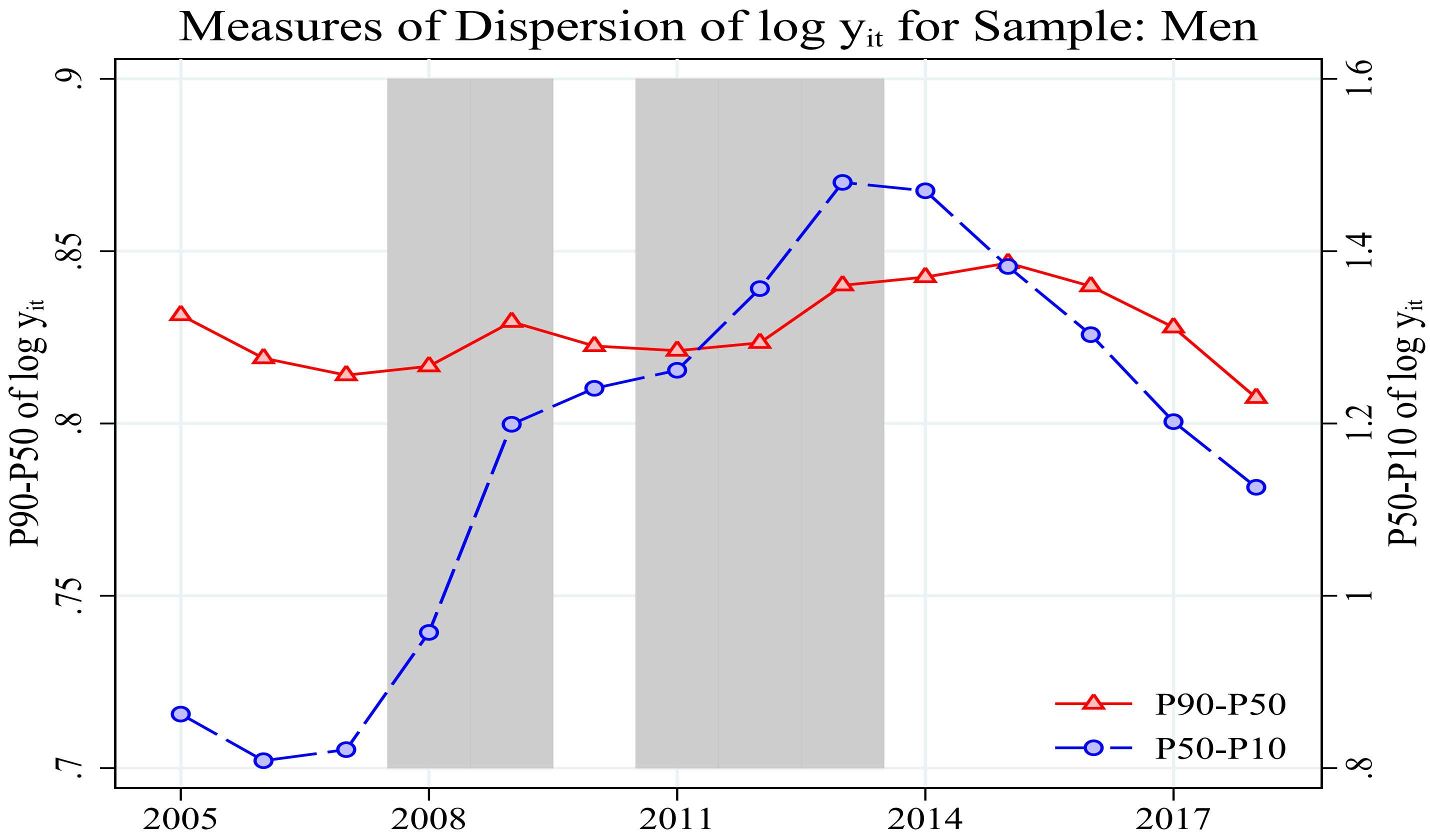
Top income percentiles



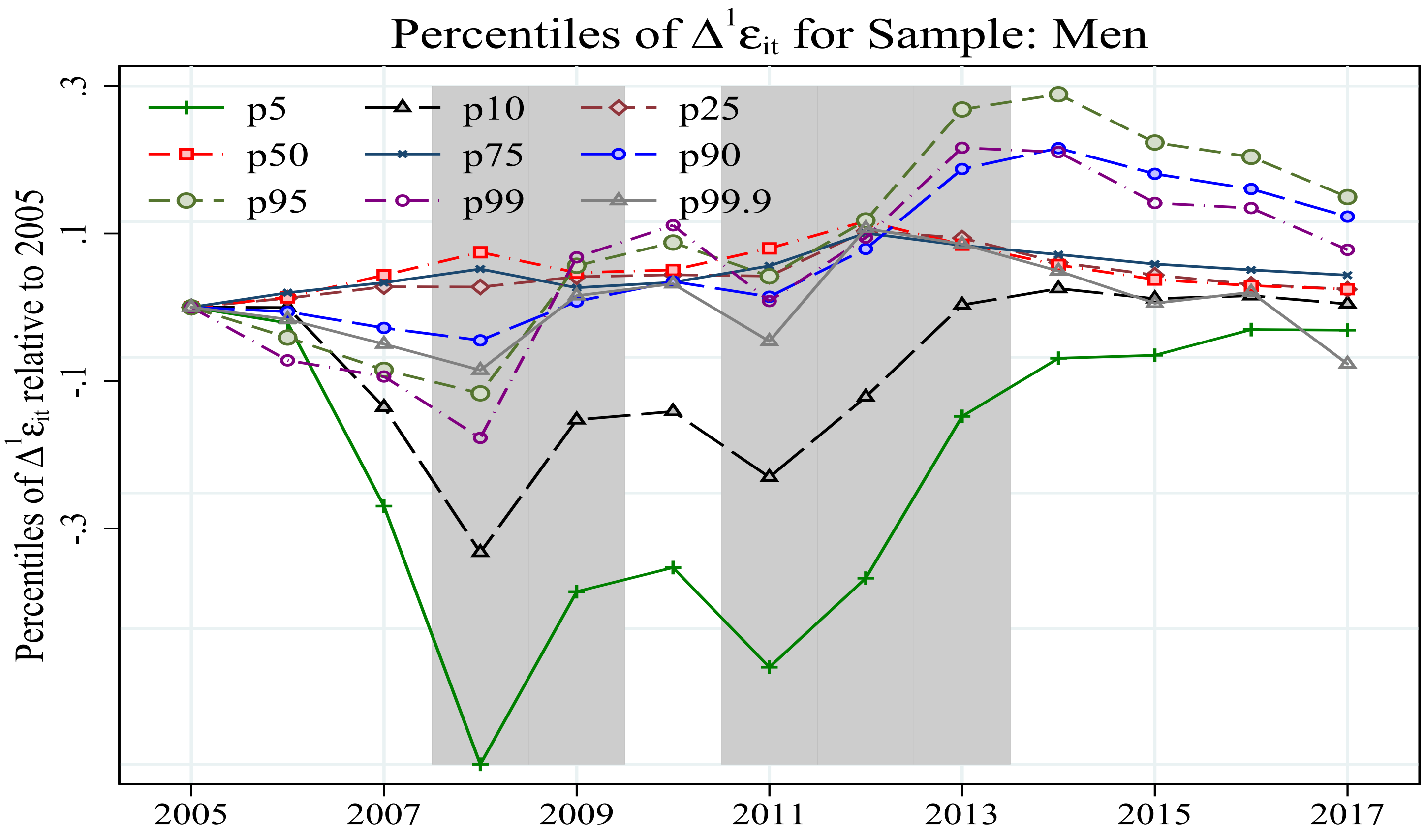
Income inequality



Upper and lower income inequality

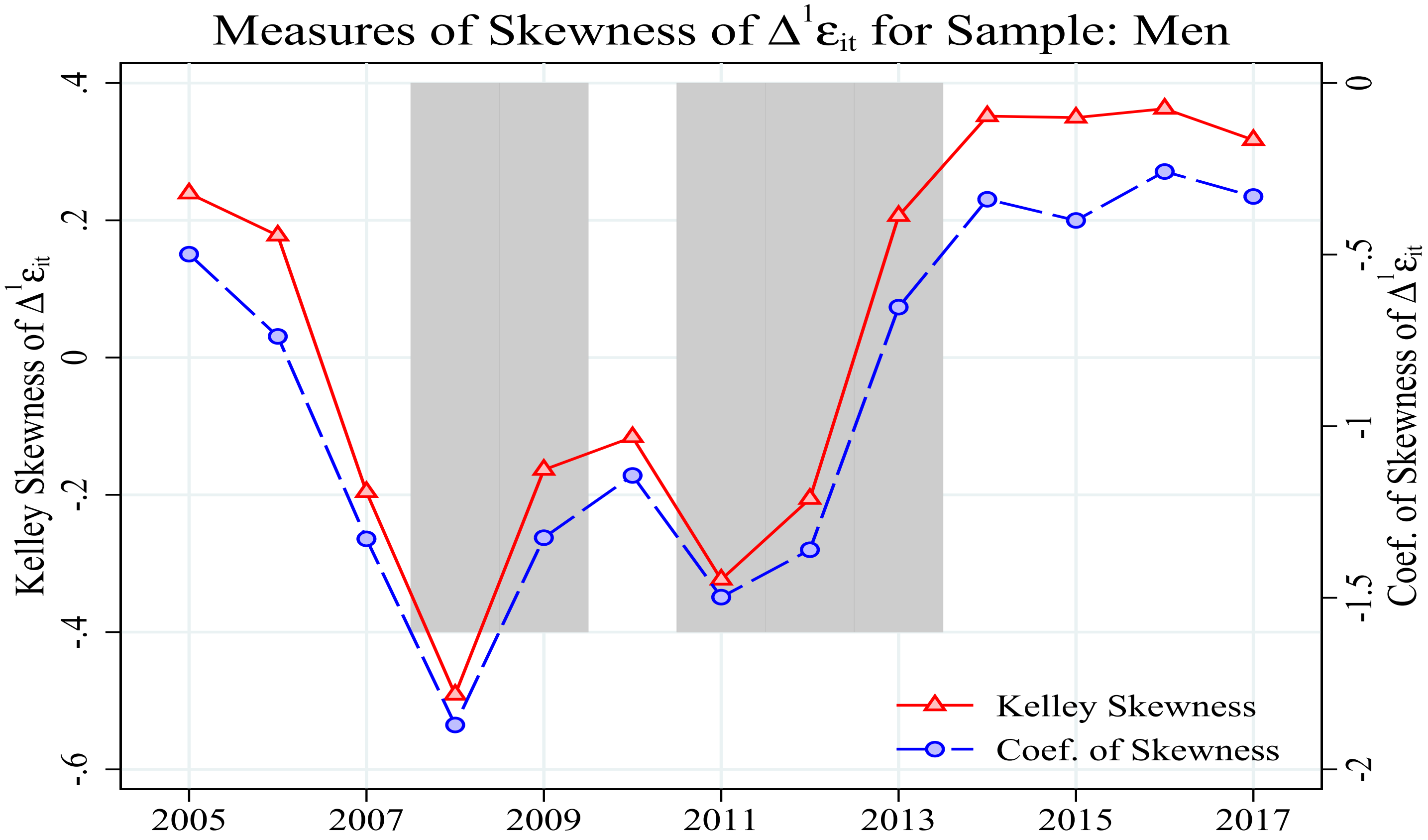


Percentiles of income growth residuals

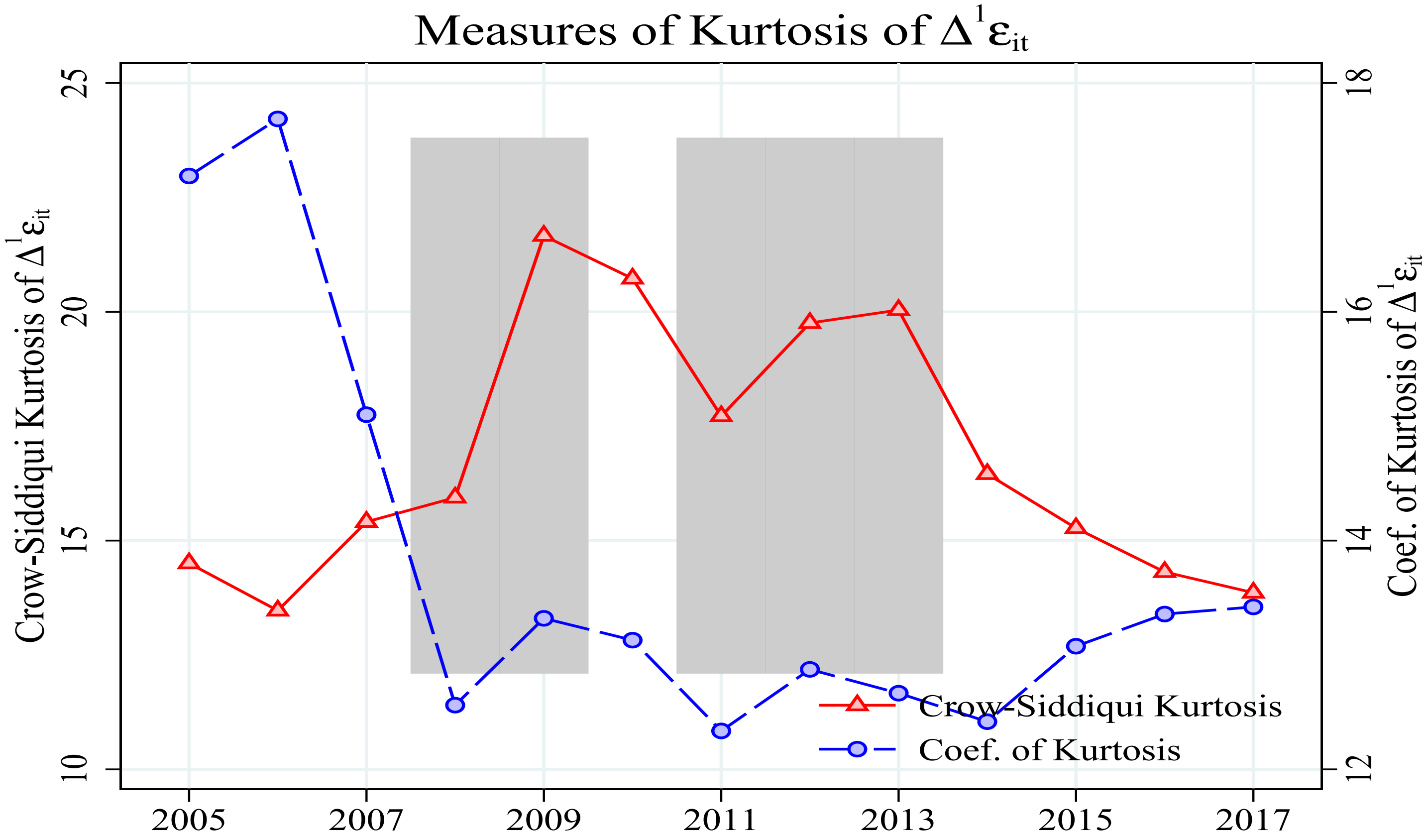


Notes: Residuals are computed by regressing log-income on age dummies, separately by year and gender.

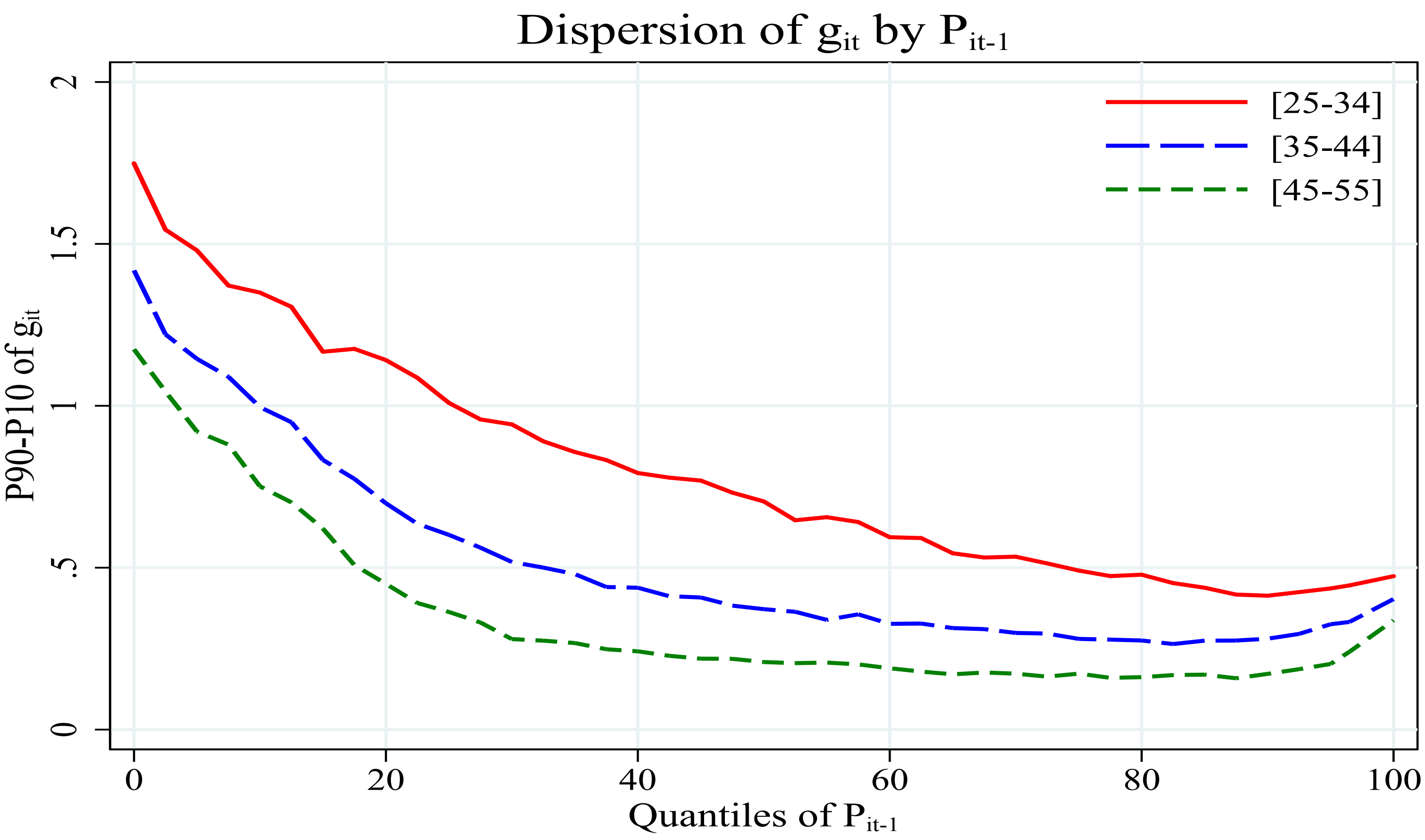
Skewness of income growth residuals



Kurtosis of income growth residuals

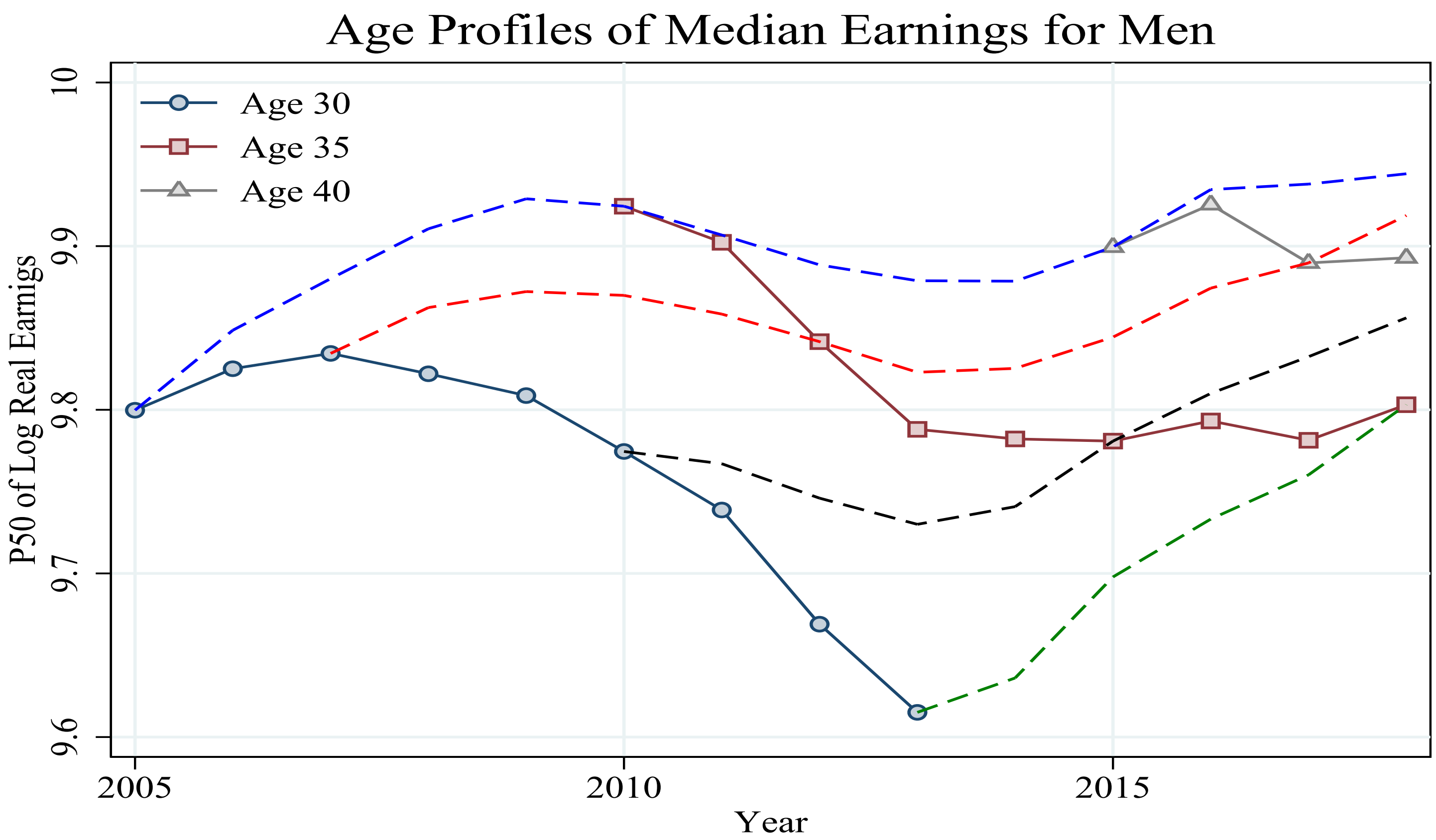


Conditional dispersion of income growth residuals



Notes: g_{it} *are log-income growth residuals,* P_{it} *is permanent log-income computed as an average over three years.*

Cohort and age profiles



Notes: Solid lines correspond to different cohorts, dashed lines to different ages within a cohort.

Part 2: Income Risk Inequality

Part 2a: Quantifying Income Risk

Predicting income

• Our goal is to mimic the agent's prediction problem (as closely as we can, in the absence of expectations data).

• We target the distribution of income levels Y_{it} given predictors X_{it}.

• Micro predictors: past income, past employment, labor contract, demographics. + unobserved predictors ("types").

• Macro predictors: GDP growth and unemployment rate, national and province.

• Here income is a comprehensive measure including (1) observations below Part 1's threshold and zeros, (2) unemployment benefits.

Measuring risk using CV

• We compute the coefficient of variation:

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

• We use the MAD instead of the standard deviation in the numerator to minimize sensitivity to extreme observations. A rescaled version — by $\approx .7$ — is directly comparable to the standard CV.

• When (standard) CV is small it can be approximated by the standard deviation of the log: Std(ln$(Y_{it})|X_{it}$). However, CV remains well-defined when $Y_{it} = 0$.

Welfare interpretation

• In the spirit of Lucas (1987), one can approximate the welfare gain to an individual associated with eliminating income risk.

• For an individual with CRRA indirect utility $U_i(Y_{it}) = \frac{Y_{it}^{1-\theta_i}-1}{1-\theta_i}$, the gain can be approximated in %income as

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

• That is, alternatively,

$$\text{Welfare gain} \approx \frac{1}{4} \times \theta_i \times CV(X_{it})^2.$$

Part 2b: Econometric Methods

The basic approach

- Since $Y_{it} \geq 0$, a natural parametric estimator is based on the two exponential specifications

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

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

- We estimate these two quantities using exponential regressions, and report the ratio.

- In the next slides we explain: (1) how we make the specification more flexible, and (2) how we incorporate unobserved heterogeneity.

Estimating CV using neural networks

• Consider the numerator of CV (we proceed similarly for the denominator).

• A one-layer neural network specification is

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

• For this presentation we take $\tau(u) = \mathrm{max}(u, 0)$ ("rectified linear unit" — ReLU), $M = 25$, and use the Poisson loss function.

• Many variations are available: different $\tau(\cdot)$, multiple layers, penalization and tuning using cross-validation..., and they are on our to-do list.

Accounting for unobserved individual heterogeneity

• To mimic the individual's prediction problem it is important to account for predictors that we as researchers do not observe.

• Our goal is to augment the predictors set as (X_{it}, ξ_i), where ξ_i is a latent component. There are 2 standard approaches:

– Modeling ξ_i using random-effects is not practical, since that would require modeling the joint distribution of $(Y_{i1}, X_{i1}, ..., Y_{iT}, X_{iT}, \xi_i)$.

– Estimating ξ_i parameters using fixed-effects in $\mathbb{E}(Y_{it}|X_{it}, \xi_i)$ and $\mathbb{E}\left(|Y_{it} - \mathbb{E}(Y_{it} \,|\, X_{it})| \,|\, X_{it}, \xi_i\right)$ is challenging due to the nonlinearity.

Unobserved individual heterogeneity: grouped fixed-effects

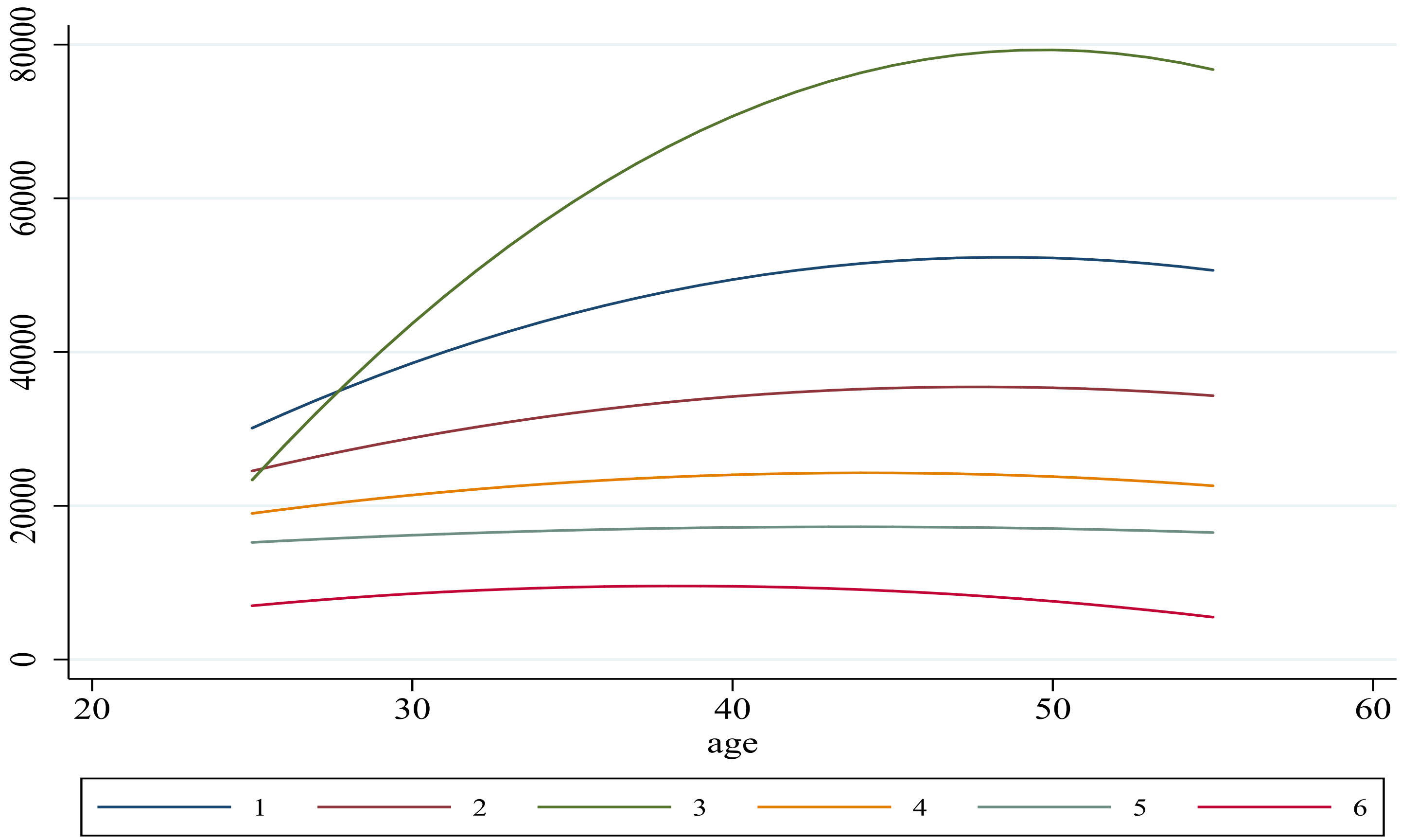
• Following Bonhomme, Lamadon and Manresa (2017) we first group individuals into K categories, and then include the group indicators as predictors to estimate CV.

• A simple approach is to group individuals based on their mean income $\frac{1}{T}\sum_{t=1}^{T} Y_{it}$. However, in an unbalanced panel this naive approach tends to conflate individual heterogeneity with age.

• To account for age we minimize $\sum_{i,t}(Y_{it} - \widetilde{X}_{it}'\beta(k_i))^2$ with respect to parameters $\beta(k)$ and group indicators k_i, where $\widetilde{X}_{it} = (1, age_{it}, age_{it}^2)'$.

• We implement this idea using a variation of Lloyd's algorithm for kmeans clustering.

Age income profiles for the estimated groups ($K = 6$)



Part 2c: Results

Prediction performance, mean (CV denominator)

	In-Sample			Out-Of-Sample		
	Homog.	Heterog.	Neural Net	Homog.	Heterog.	Neural Net
p1	1.0000	0.5638	0.2531	0.9238	0.6037	0.2913
p5	1.0000	0.5743	0.2581	0.9688	0.6093	0.2899
p10	1.0000	0.5788	0.2615	0.9619	0.6097	0.2880
p25	1.0000	0.6236	0.3028	1.0457	0.6610	0.3148
p50	1.0000	0.5777	0.3109	1.0162	0.6445	0.3279
p75	1.0000	0.5910	0.3761	1.0288	0.6699	0.4014
p90	1.0000	0.6059	0.4338	1.0429	0.6981	0.4736
p95	1.0000	0.6522	0.4648	1.0372	0.7308	0.5093
p99	1.0000	0.8001	0.5262	1.1029	0.8133	0.5499

Notes: All numbers are in percentage of the in-sample error for the homogeneous exponential specification. OOS is for 2018.

Prediction performance, mean absolute deviation (CV numerator)

	In-Sample			Out-Of-Sample		
	Homog.	Heterog.	Neural Net	Homog.	Heterog.	Neural Net
p1	1.0000	0.7211	0.4829	1.0959	0.6743	0.4268
p5	1.0000	0.7278	0.4910	1.1153	0.6970	0.4281
p10	1.0000	0.7548	0.5110	1.1161	0.7368	0.4508
p25	1.0000	0.6621	0.4539	1.0827	0.6872	0.4285
p50	1.0000	0.6339	0.4040	1.0521	0.6471	0.3972
p75	1.0000	0.5860	0.4060	1.0549	0.6499	0.4105
p90	1.0000	0.6225	0.4530	1.0396	0.7797	0.5083
p95	1.0000	0.7574	0.5032	1.0409	0.8753	0.5761
p99	1.0000	0.7779	0.5646	1.1728	0.8391	0.6649

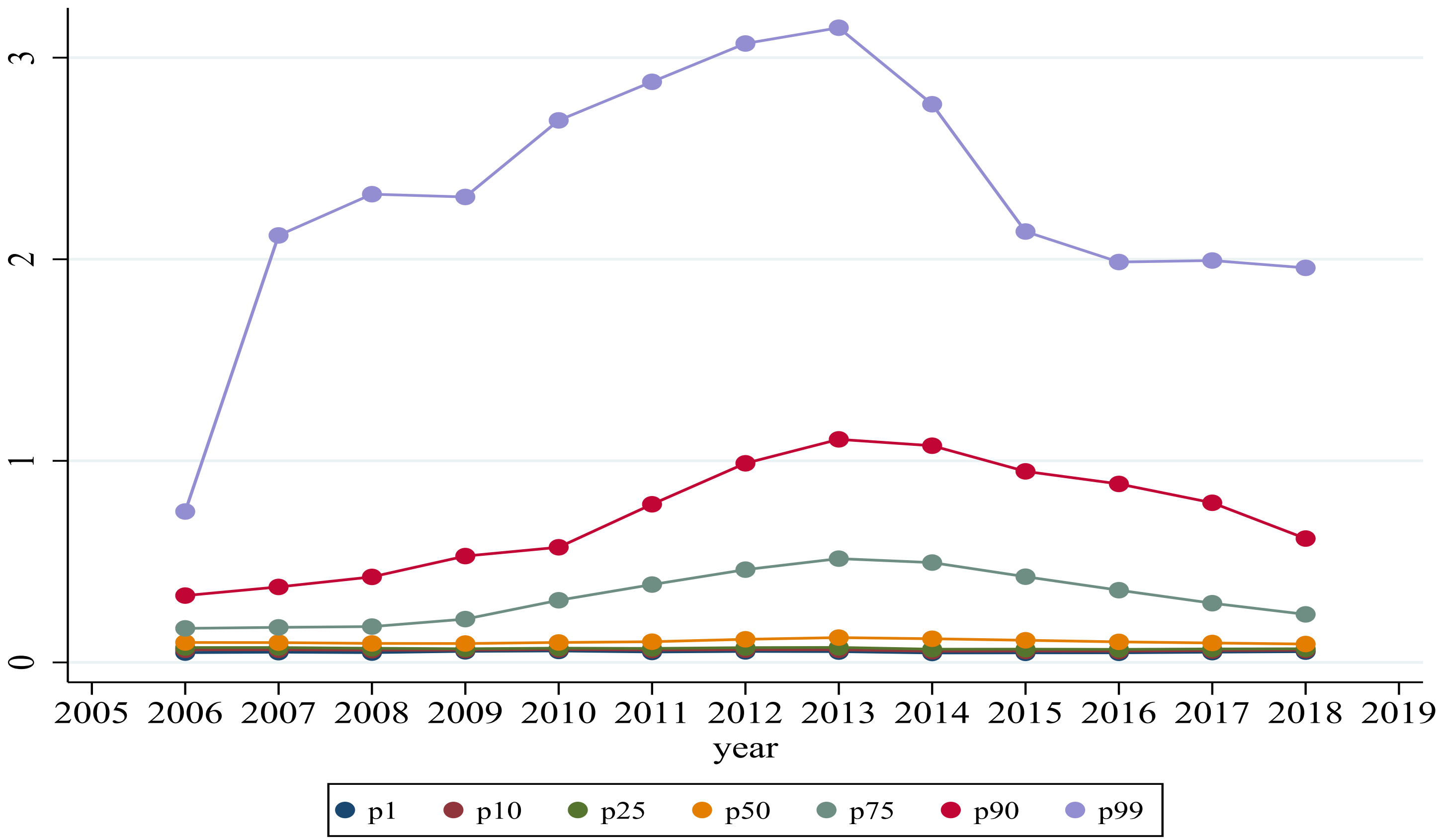
Notes: All numbers are in percentage of the in-sample error for the homogeneous exponential specification. OOS is for 2018.

Explaining variation in CV

Age Range	26-30	36-40	46-50
Business Cycle	0.0183	0.0057	0.0047
Permanent (t-1)	0.0011	0.0016	0.0015
Fulltime (t-1)	0.0404	0.0331	0.0290
Days Worked (t-1)	0.3808	0.3248	0.2376
Income (t-1)	0.0041	0.0137	0.0031
Unobs. Het.	0.0792	0.0535	0.0268

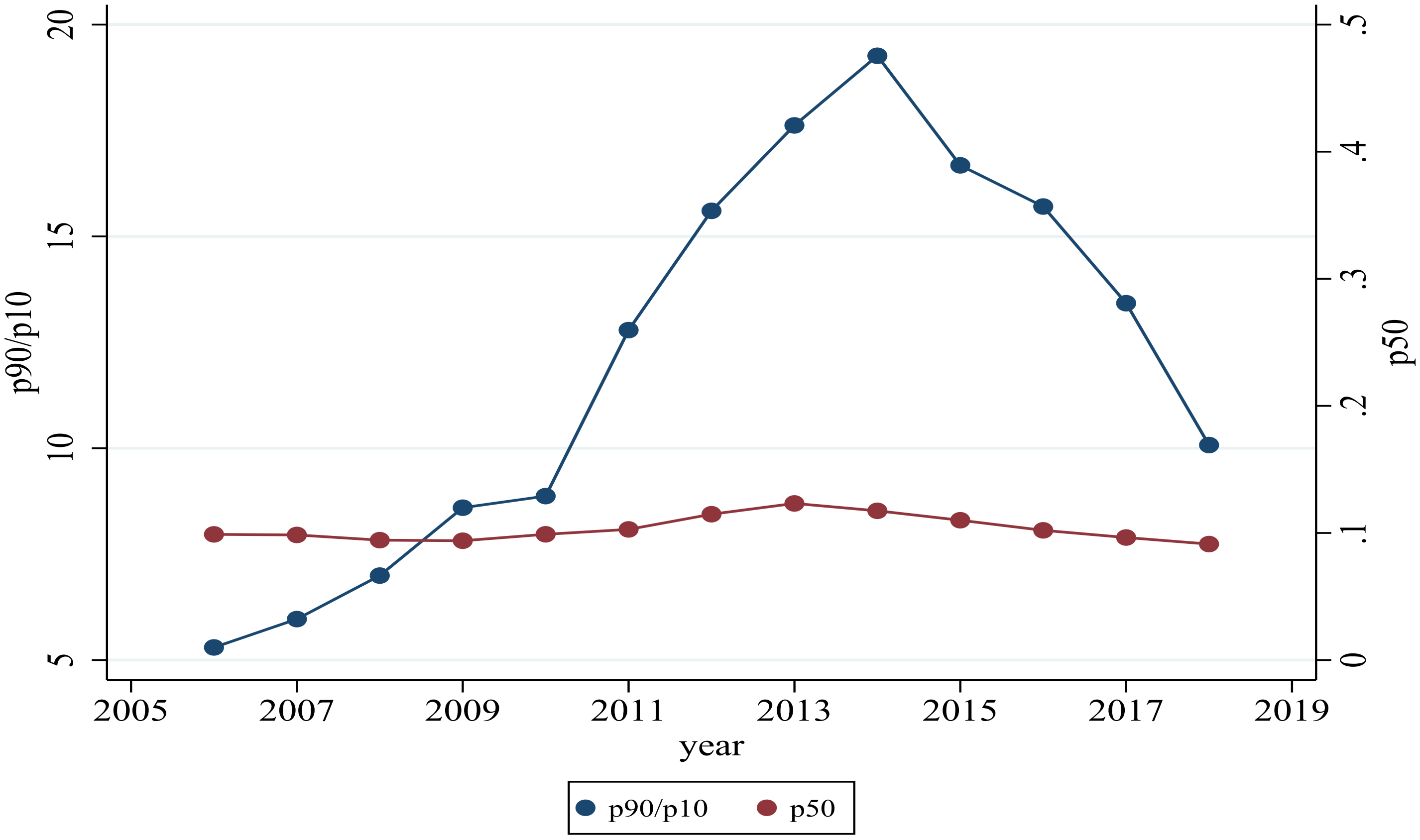
Notes: Partial R^2 in CV regressions. Neural network specification with unobserved heterogeneity groups.

Income risk inequality over the business cycle: quantiles



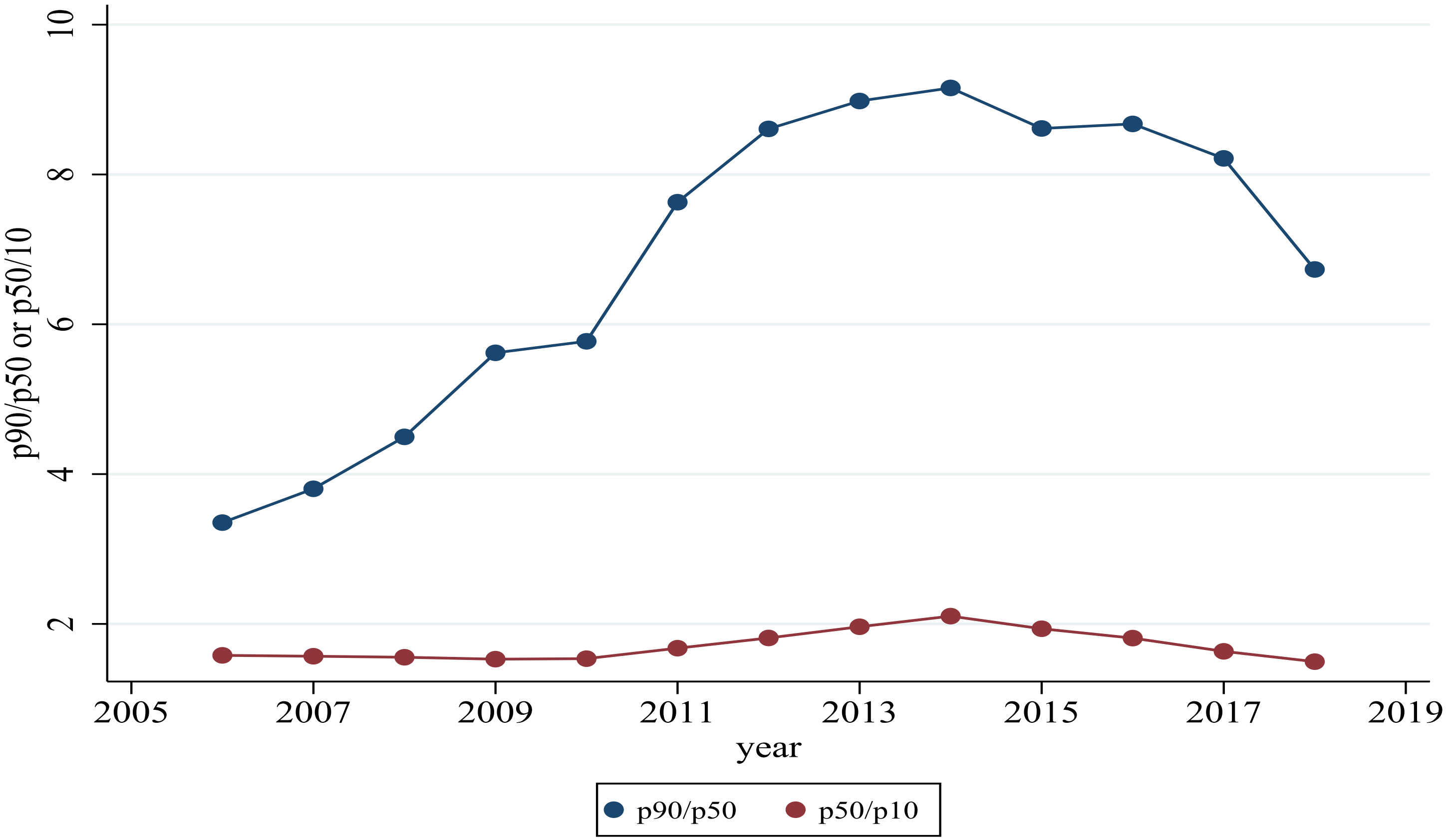
Notes: Quantiles of CV. Neural network specification with unobserved heterogeneity groups.

Income risk inequality over the business cycle (cont.)



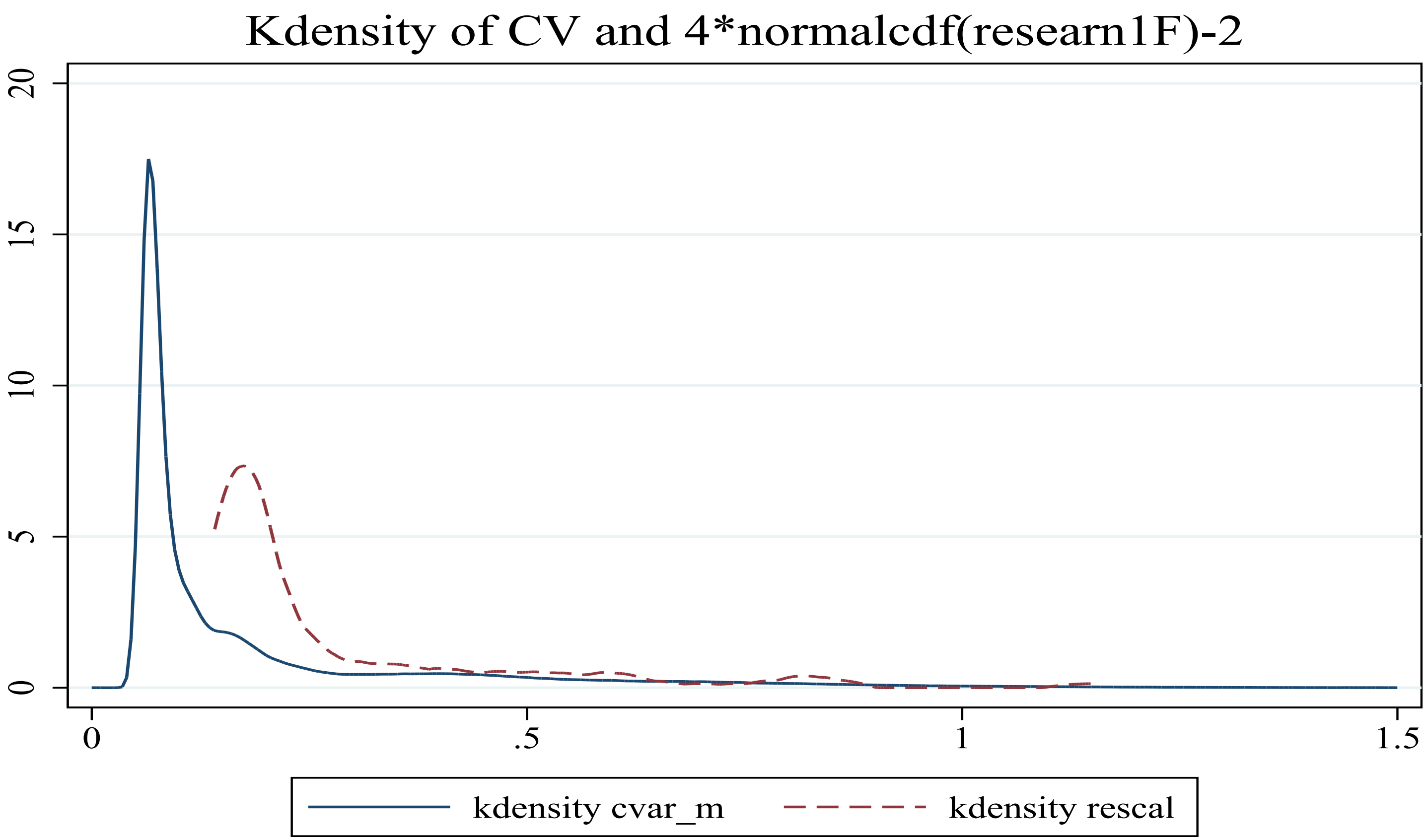
Notes: Neural network specification with unobserved heterogeneity groups.

Income risk inequality over the business cycle (cont.)



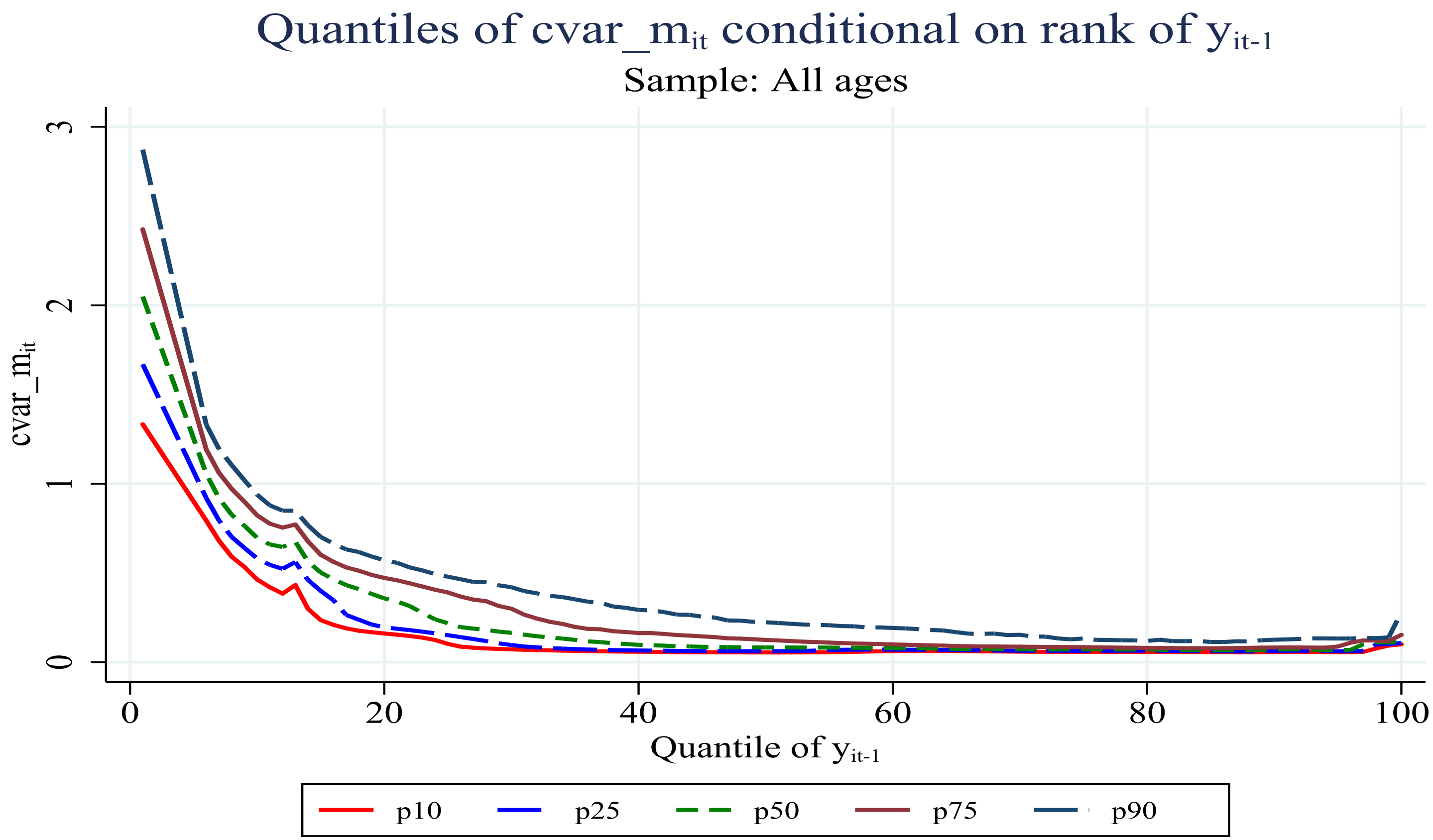
Notes: Neural network specification with unobserved heterogeneity groups.

Income risk inequality: $CV(X_{it})$ **versus** $\mathrm{Std}(Y_{it}\,|\,Y_{i,t-1})$



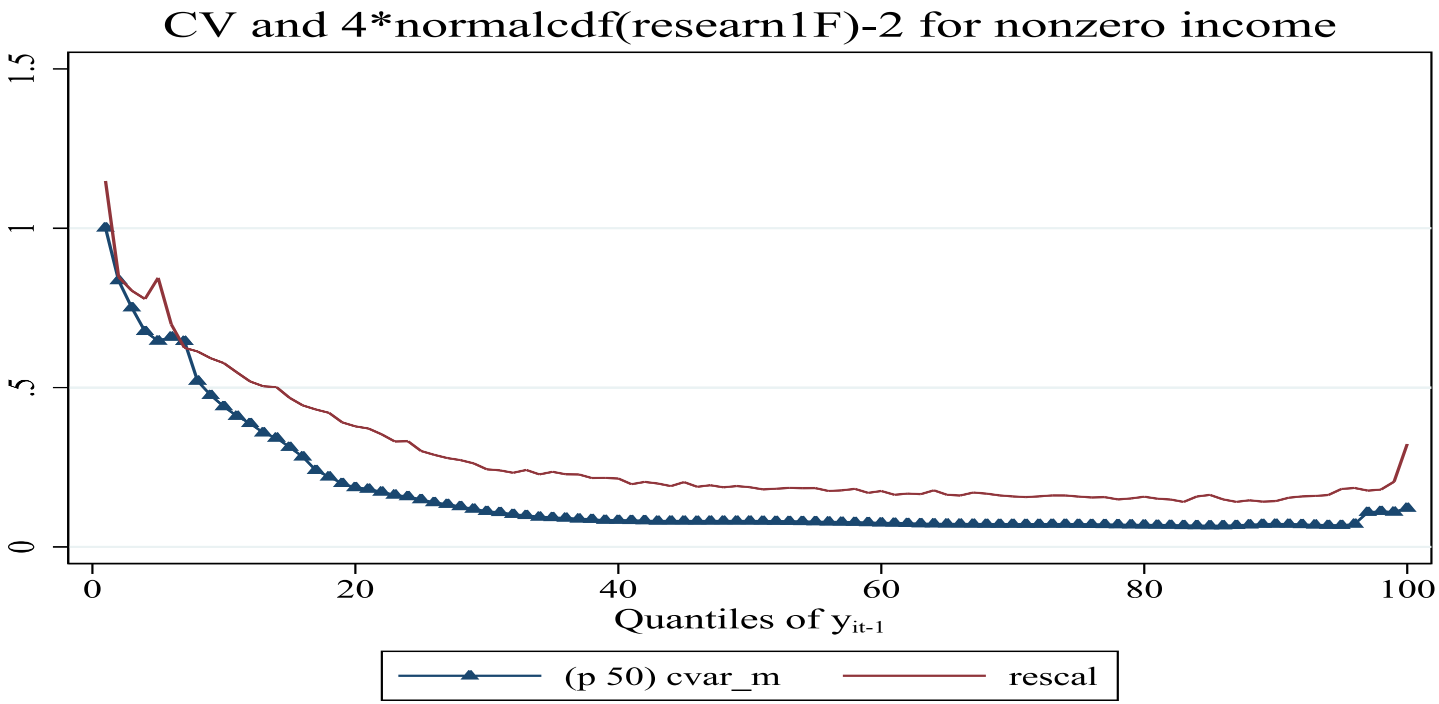
Notes: Kernel density estimates. Neural network specification with unobserved heterogeneity groups.

Income risk and income



Notes: Conditional quantiles of $CV(X_{it})$ given income $Y_{i,t-1}$. Neural network specification with unobserved heterogeneity groups.

Income risk and income: $CV(X_{it})$ versus $\mathrm{Std}(Y_{it} \,|\, Y_{i,t-1})$



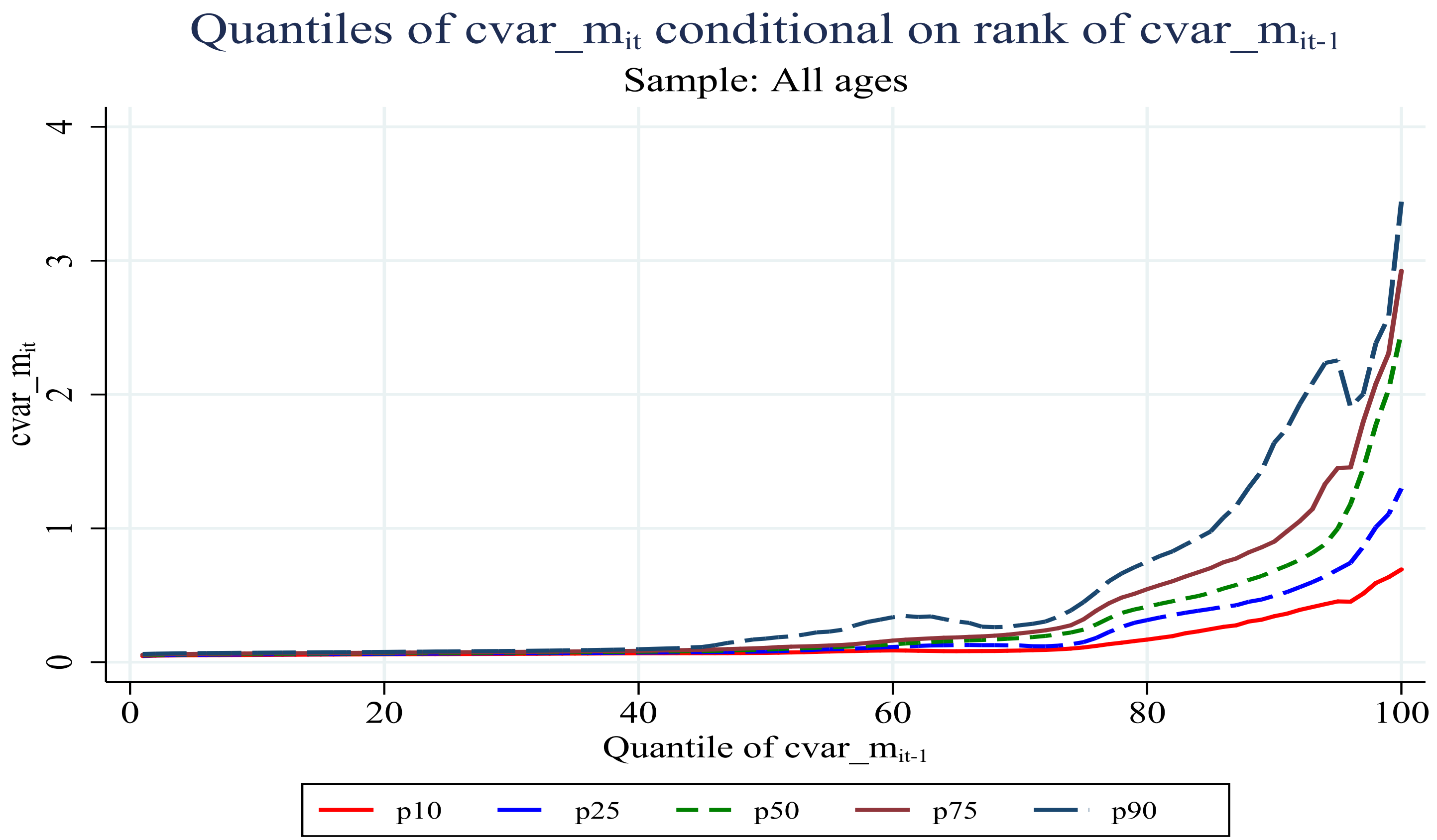
Notes: Conditional mean of $CV(X_{it})$ given income $Y_{i,t-1}$ (blue) and binned estimate of $\mathrm{Std}(Y_{it} \,|\, Y_{i,t-1})$*, rescaled (red). Neural network specification with unobserved heterogeneity groups. Sample with non-zero income.*

Income risk over the life cycle

	Risk_30	Risk_35	Risk_40	Risk_45	Risk_50	Risk_55
P10	1.13	1.04	0.99	0.97	0.96	0.99
P25	1.19	1.04	0.97	0.95	0.94	0.95
P50	1.47	1.01	0.90	0.87	0.84	0.84
P75	1.35	0.95	0.80	0.79	0.82	0.89
P90	1.16	0.93	0.87	0.92	0.97	1.10

Notes: τ-th percentile of $CV(X_{it})$ in a given age bin divided by τ-th percentile of $CV(X_{it})$. Neural network specification with unobserved heterogeneity groups.

Persistence in income risk



Notes: Conditional mean of $CV(X_{it})$ given lagged $CV(X_{i,t-1})$. Neural network specification with unobserved heterogeneity groups.

Summary and future steps

• We measure income risk using prediction methods and a set of observed and latent predictors.

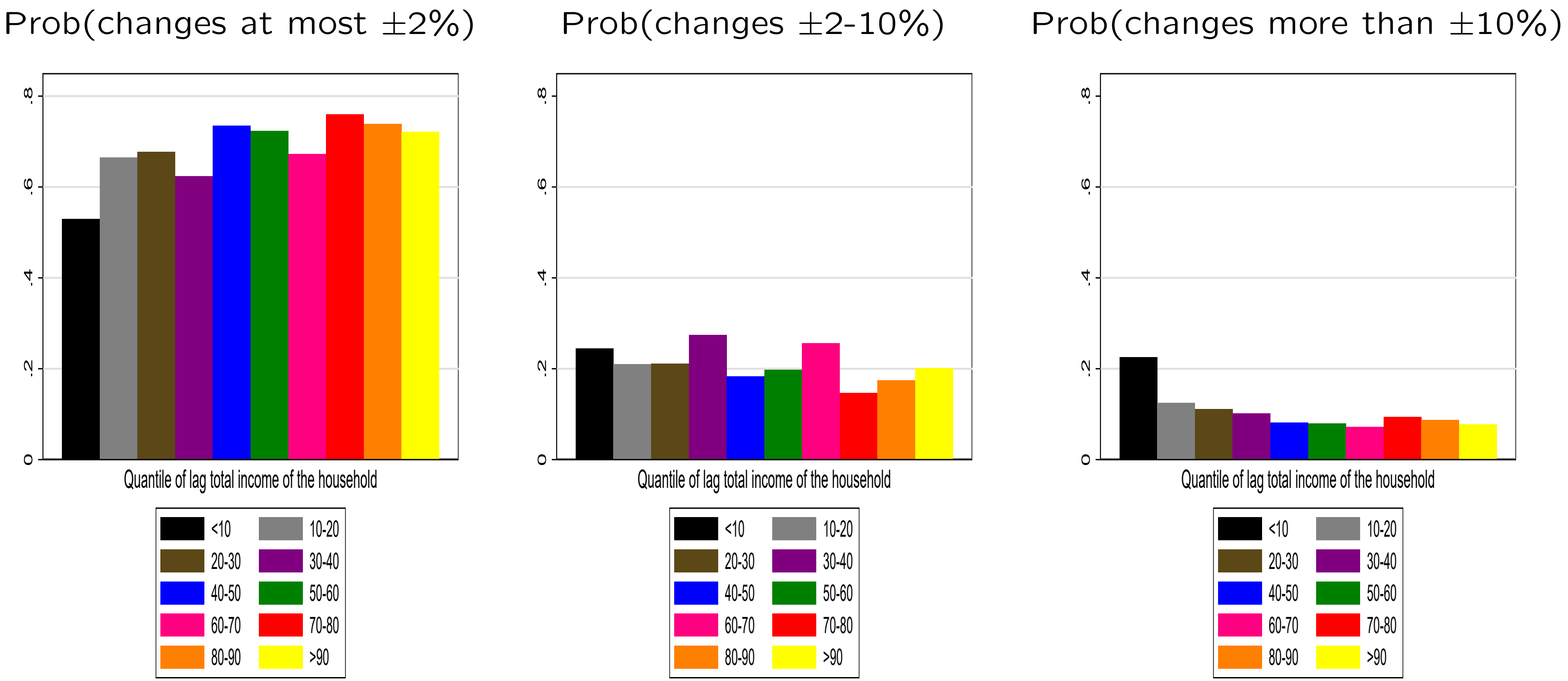
• Risk is highly unequal in Spain: more than half of the economy has close to perfect predictability of their income, while some face considerable uncertainty.

• Many additional robustness checks are needed: neural network, grouping, choice of predictors, robust CV measures...

• Question 1: How to incorporate our measure in life-cycle models? Not today.

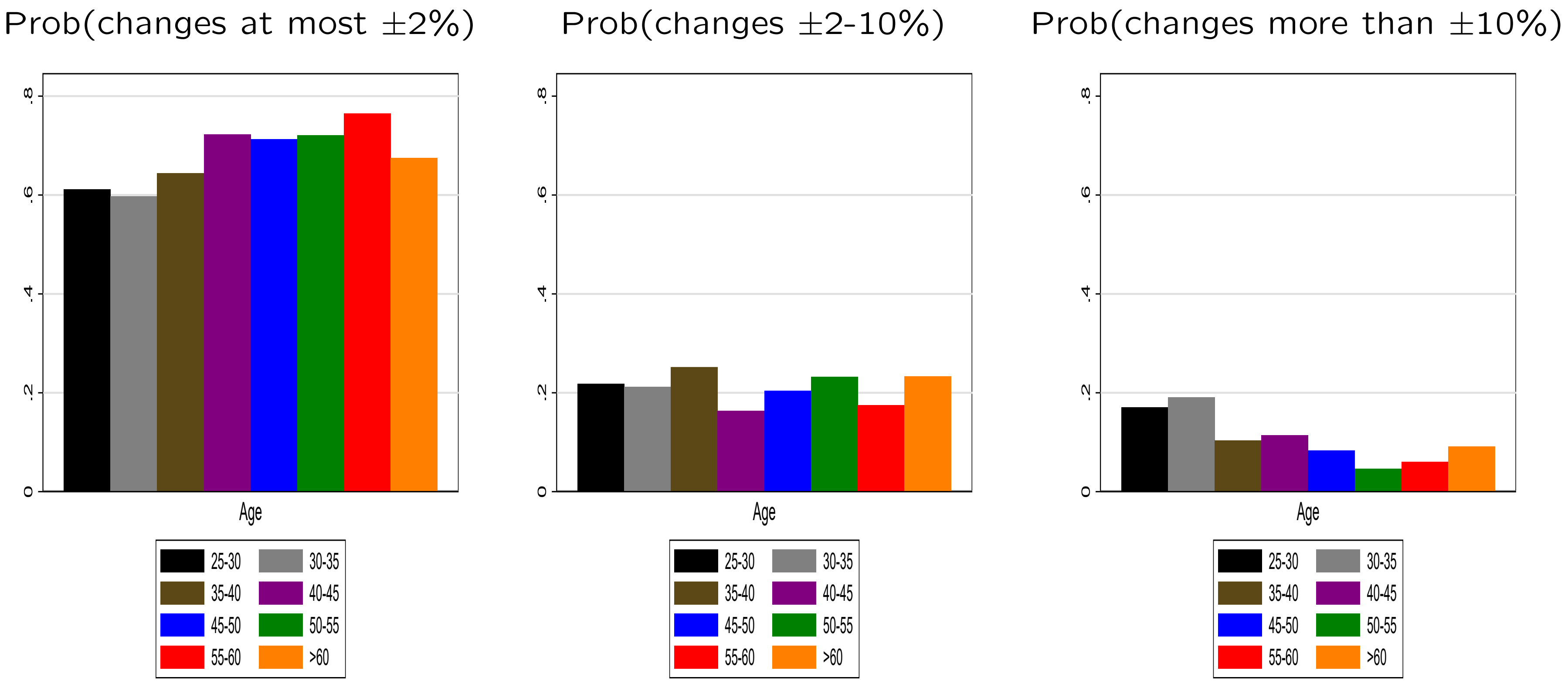
• Question 2: How does our risk measure compare with subjective expectations data? A first look in the next two slides.

Subjective income expectations, by lagged income



Notes: Subjective income expectations from the Encuesta Financiera de las Familias (EFF), 2014.

Subjective income expectations, by age



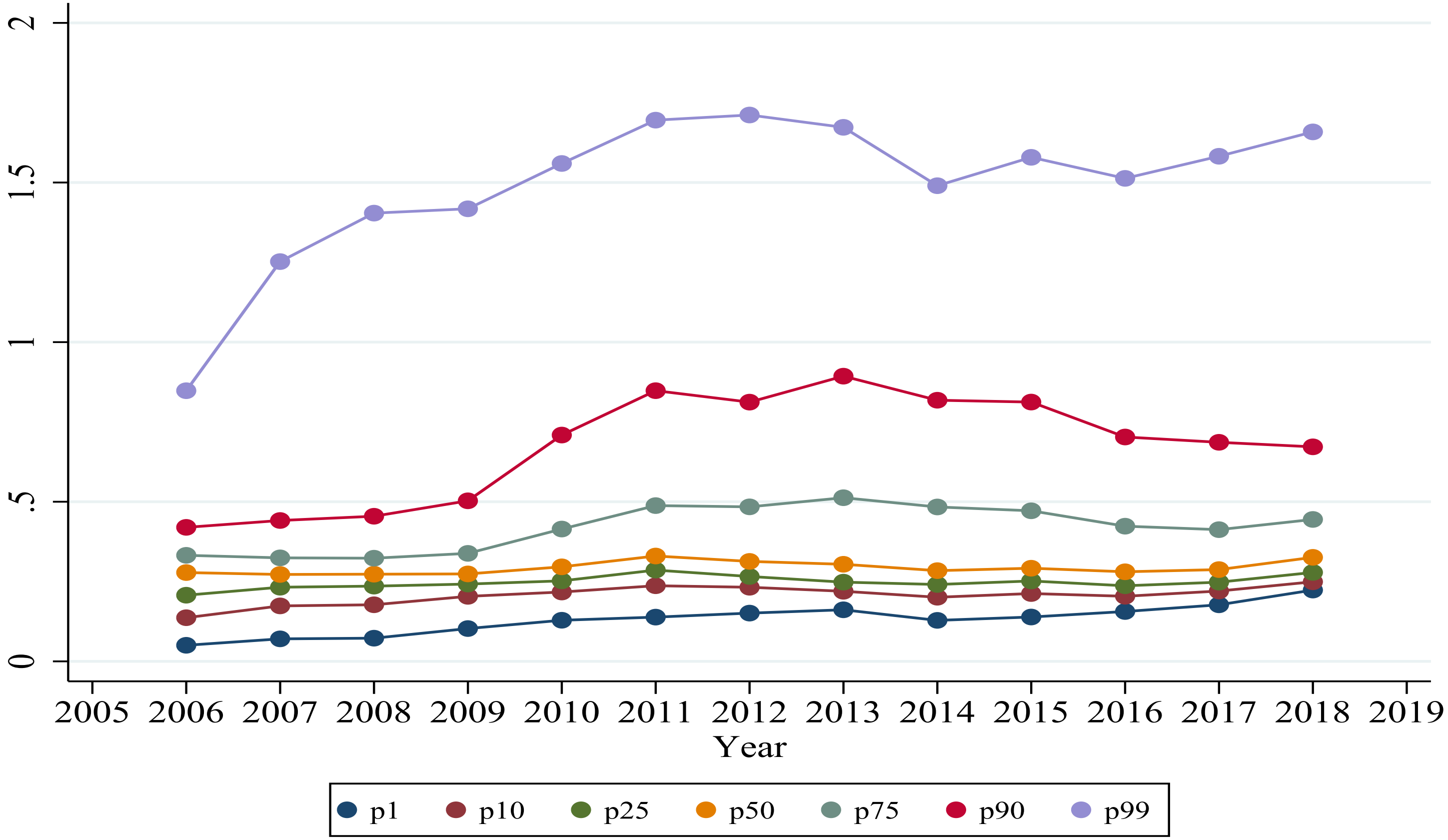
Notes: Subjective income expectations from the Encuesta Financiera de las Familias (EFF), 2014.

Back-up Slides

%Observations below the income threshold (for Part 1)

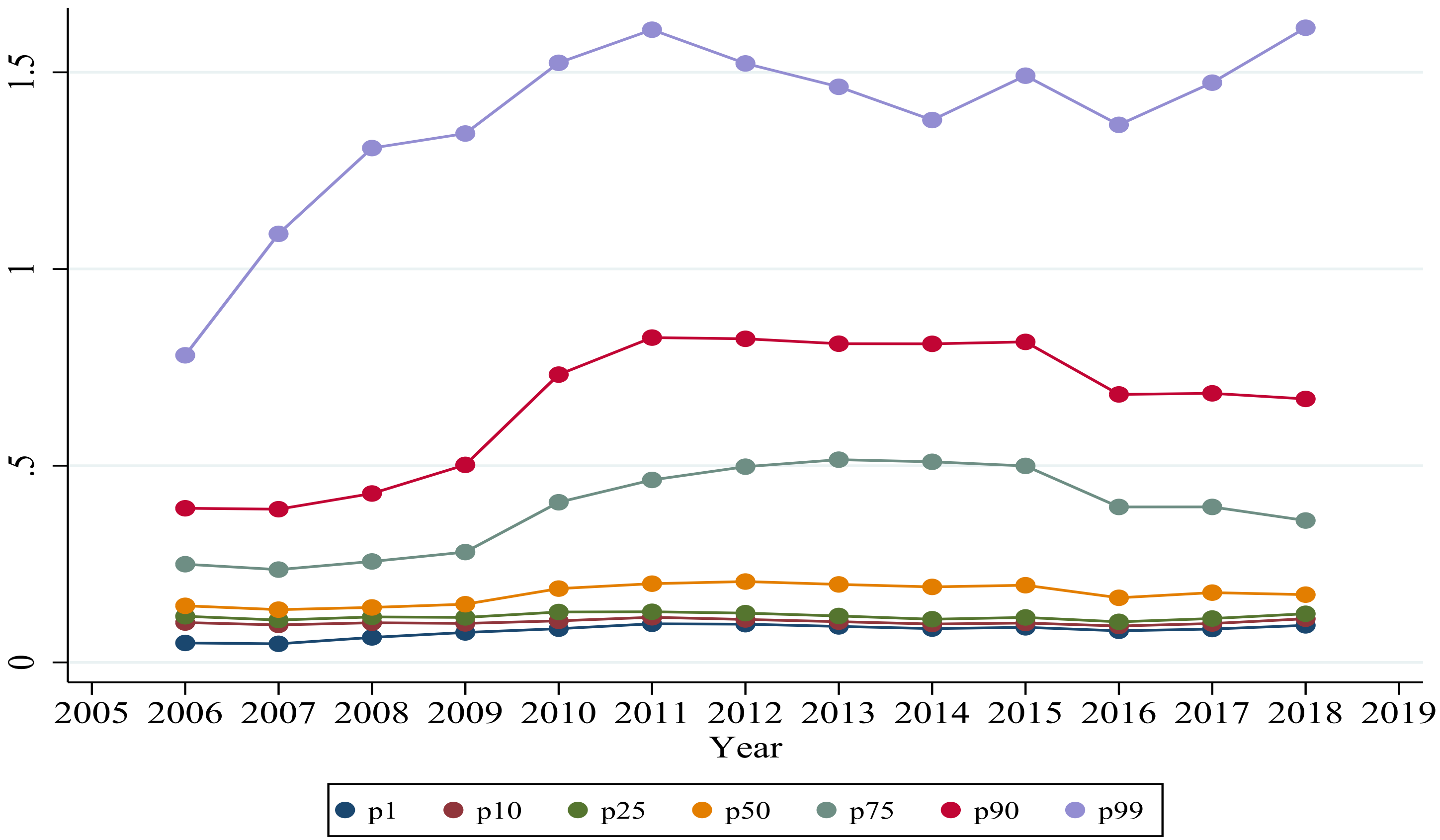
	# Observations	Proportion
2005	243666	.0412532
2006	256059	.0533822
2007	267149	.0589072
2008	273619	.0781378
2009	276898	.1376933
2010	278232	.1685392
2011	275551	.1826196
2012	275937	.2198871
2013	274062	.2304296
2014	270424	.2085577
2015	266706	.1740418
2016	263216	.1450862
2017	259894	.1120611
2018	253637	.0734633

Income risk inequality over the business cycle: quantiles



Notes: Quantiles of CV. Exponential specification without unobserved heterogeneity.

Income risk inequality over the business cycle: quantiles



Notes: Quantiles of CV. Exponential specification with unobserved heterogeneity groups.