Income uncertainty, nonlinear dynamics and consumption: a subjective income expectations framework*

Henrique S. Basso
Banco de España and CEMFI

Olympia Bover
CEMFI

Julio Galvez
CUNEF Universidad

Laura Hospido
Banco de España and CEMFI

This draft: March 25, 2026

Abstract

Household consumption and saving decisions depend on perceived income risk, yet macroeconomic models often discipline income processes using realized income data. We show that these two objects differ in systematic and economically meaningful ways. We develop a framework that combines subjective expectations with realized income data, while addressing non-classical elicitation error and focal-point responses, to estimate flexible income processes that capture nonlinear persistence and distributional asymmetries. Households perceive income to be substantially more persistent and less dispersed than realized data imply. Our findings support incorporating subjective probability data into structural models of income dynamics, consumption, and precautionary behavior.

Keywords: Subjective expectations, panel data, life cycle, earnings dynamics, nonlinear persistence, consumption.

JEL codes: D15, D31, D84, E21.

*A previous version of the paper circulated under the title "Subjective Income Expectations and Consumption Dynamics". We are grateful to Manuel Arellano, Stéphane Bonhomme, Richard Blundell, Christopher Carroll, Jesús Fernández–Villaverde, Pamela Giustinelli, Nezih Guner, Fatih Guvenen, Thierry Magnac, Benjamin Moll, Wilbert van der Klauuw, and Siqi Wei for valuable suggestions. We also thank Jose Maria Casado for his early contributions to the project, and Jiayi Wen for excellent research assistance. We also thank seminar and conference participants at CUNEF Universidad, Erasmus University Rotterdam, Encuentro de Economía Pública (Valencia), EEA-ESEM (Cologne and Barcelona), International Panel Data Conference (Vilnius), 6th Conference on Household Finance and Consumption (Dublin), SEHO meeting (Copenhagen), the Conference on econometric methods and empirical analysis of micro data in honor of Manuel Arellano (Madrid), IAAE (Oslo), the Workshop on Subjective Expectations (Lisbon), and the Econometric Society World Congress (Seoul) for useful feedback. The views expressed here are those of the authors and, therefore, do not necessarily coincide with those of the Banco de España or the Eurosystem.

1 Introduction

Household consumption and saving decisions depend on perceived uncertainty about future income, not on ex-post realizations. Structural models discipline this uncertainty using statistical income processes estimated from realized income data. Implicitly, this assumes that the income risk econometricians infer from past outcomes coincides with the risk households perceive when making decisions. In this paper, we show that this assumption is violated in economically meaningful ways. Using subjective income expectations, we recover income processes that differ substantially from those estimated using realized income data, with implications for persistence, tail risk, and heterogeneity in income dynamics. These differences matter for the understanding of consumption smoothing, precautionary saving, and the distributional consequences of income risk.

Recent work using administrative and survey data has documented pronounced nonlinearities in income dynamics, including state-dependent persistence, asymmetric shocks, and non-Gaussian income changes (Arellano, Blundell, and Bonhomme, 2017; Guvenen, Karahan, Ozkan, and Song, 2021; De Nardi, Fella, and Paz-Pardo, 2020). These features imply that income risk varies sharply across households and over the income distribution, generating heterogeneous consumption-smoothing behavior and precautionary saving motives. For example, large positive shocks to low-income households or large negative shocks to high-income households tend to be less persistent, effectively resetting income histories, while moderate shocks exhibit higher persistence. Such nonlinear dynamics alter wealth accumulation and inequality, underscoring why accurate measurement of income risk is central for macroeconomic analysis (Kaplan and Violante, 2010; De Nardi et al., 2020; Gálvez and Paz-Pardo, 2026). However, existing estimates of these nonlinear income processes are almost exclusively based on realized income data, abstracting from the information households possess about their future income prospects.

A central difficulty in measuring income risk from realized income data is that the econometrician does not observe households' information sets (Jappelli and Pistaferri, 2010), nor the process by which expectations are formed. Households often possess advance knowledge about future income changes—such as promotions, bonuses, contract renewals, or lay-

offs—that is unobservable to the researcher. Households expectations may be shaped by diagnostic biases, with agents overweighting salient states and underreacting to base rates, effectively assigning disproportionate weight to outcomes that appear representative of recent or attention-grabbing information (Bordalo, Gennaioli, and Shleifer, 2013, 2018)). As a result, income innovations identified from realized data mix true uncertainty with forecastable variation under rational expectations. From a macroeconomic perspective, this distinction is crucial: households smooth consumption and save against uncertainty, not against income changes they anticipate. When income processes are estimated solely from realized outcomes, persistence, volatility, and tail risk may therefore be mismeasured from the standpoint relevant for economic decision-making.

Subjective income expectations provide a direct window into households' perceived income risk. An increasing number of large-scale surveys elicit probabilistic expectations about future income, asking respondents to assign probabilities to alternative income growth scenarios over a fixed horizon. These data capture households' beliefs about both the likelihood and the dispersion of future income changes, incorporating private information and perceived tail risks that may not materialize within the observed sample period. From a macroeconomic perspective, subjective expectations are therefore closer to the object that disciplines consumption, saving, and insurance decisions than income processes inferred from realized data alone. Moreover, because each survey response reveals an entire perceived distribution, expectations data can uncover heterogeneity in income risk without requiring long panels of realized outcomes.

We propose an empirical framework to estimate the parameters of a *flexible income process* by combining probabilistic subjective income expectations with realized income data. The framework is designed to accommodate the kind of nonlinear persistence and non-normal shock distributions that recent studies have highlighted, but we approach these issues from the perspective of subjective income risk. Specifically, we leverage newly available data from the Spanish Survey of Household Finances (EFF), which in its 2014, 2017, 2020, and 2022 waves asked households to distribute 10 points across five predefined categories of income growth for the next 12 months (ranging from large drops to large gains). This elicitation yields for each

household an entire probability distribution over next-year income growth. We combine these subjective distributions with each household's current income and observable characteristics to estimate the conditional distribution of future income. An appealing feature of the EFF data is that it provides a panel dimension, and also contains realized income, consumption, and wealth information. This allows us not only to estimate perceived income risk, but also to compare subjective income dynamics to those inferred from actual income realizations for the same households, and to discuss implications for consumption behavior. Our work is, to our knowledge, the first to rigorously estimate an income process with rich nonlinear features using subjective expectations from a representative household sample in a developed economy, and to contrast it with estimates from realized income data. This contributes to bridging two strands of literature: one that estimates income risk from realized data, and another that uses survey expectations to measure risk.

Estimating income risk from probabilistic subjective expectations data poses unique challenges. Unlike realized income, which can be treated as continuous outcomes observed each period, subjective expectations are typically reported in coarse form – e.g., a few probability points allocated to broad bins of income growth. This raises two issues: (i) coarse granularity of data – the survey provides only interval probabilities rather than a precise distribution or moments, and (ii) reporting bias or focal responses – a substantial fraction of respondents may give answers that heap all probability on a single outcome (for example, more than 40% of households in 2014 allocated all points to the *approximately no change* category, essentially stating that they are certain their income will remain roughly stable). Such focal-point responses might reflect genuine near-certainty for some households, or they might indicate cognitive limitations or question framing effects (respondents defaulting to the middle option when unsure). We address these challenges by developing an estimator that combines a probabilistic latent-variable approach with a mixture model for response behavior.

First, we assume that each household has an underlying latent income process that governs the distribution of future income given current information – either a state-dependent AR(1) process or a fully nonlinear quantile process.[1] The former representation provides a tractable

[1]For example, in one specification we let household i's log income follow $y_{i,t+1} = \rho(X_{it}), y_{it} + X'_{it}\beta + \eta_i + \sigma(X_{it}), u_{i,t+1}$, where X_{it} are observed characteristics (such as age, education, employment type, etc.), η_i is an

yet flexible extension of the standard linear model, while the latter allows for full distributional flexibility. Using both allows us to compare results — such as the distribution of persistence coefficients and the skewness of shocks across income levels — contrasting the implications of using either subjective and realized income data while ensuring conclusions are not driven by restrictive functional form assumptions. Second, given a particular latent income process, we derive the implied probabilities of each bin for a household with a given current income y_{it} and covariates X_{it}.[2] Third, our framework introduces a two-regime *hurdle* model to capture focal-point responses, especially the prevalent case of all probability mass placed on the middle bin (no change). Building on approaches by de Bresser and van Soest (2013), Kleinjans and van Soest (2014), and Heiss, Hurd, van Rooij, Rossmann, and Winter (2022), we allow each household $\times$ wave response to be generated by one of two regimes: (a) a deterministic reporting regime, in which the household ignores the underlying uncertainty and reports a sure outcome (by putting all 10 points in one bin, typically the middle), or (b) a probabilistic regime, in which the household allocates points according to the perceived distribution.[3]

Our approach thus blends insights from the literature on survey response behavior (Manski and Molinari, 2010; Giustinelli, Manski, and Molinari, 2022a,b) with an economic model of income dynamics. Notably, we maintain a parametric and structural interpretation of the data – in contrast to purely nonparametric bounds or partial identification approaches to dealing

unobserved individual effect, and $u_{i,t+1}$ is an innovation (Gaussian shock). The persistence parameter $\rho(X_{it})$ and shock volatility $\sigma(X_{it})$ are allowed to depend on the state X_{it}, which can include the level of y_{it} itself – capturing the idea that income may be more or less persistent or variable at different points of the income distribution (Chamberlain and Hirano, 1999; Meghir and Pistaferri, 2004; Browning, Ejrnaes, and Alvarez, 2010; Hospido, 2012; Gu and Koenker, 2017). The canonical random-walk or AR(1) income process used in many models is a special case of this specification with constant ρ and σ. In a more general formulation inspired by Arellano et al. (2017), we model the conditional distribution of $y_{i,t+1}$ given y_{it} as a quantile function $Q_y(\tau \mid y_{it}, X_{it})$ that is flexible in both τ (quantile rank of the shock) and the state variables. This quantile-based approach can capture complex patterns like varying skewness or kurtosis of income shocks; for instance, it lets us compute how the persistence of an income shock might differ between a bottom-decile shock and a top-decile shock. It also yields measures of conditional heteroskedasticity (e.g., the interquartile range of future income) and conditional skewness as functions of current income, which are objects of interest for evaluating income risk.

[2]For the linear Gaussian specification, this is done by integrating the Normal predictive distribution between the cutoffs that define the five bins (e.g., "income drops more than 10%", "drops 2–10%", "roughly unchanged (±2%)", etc.). For the quantile specification, we can similarly obtain the probability of each bin by evaluating the model's conditional CDF at the bin boundaries; our piecewise-linear quantile function (with enforced monotonicity and proper tail behavior) makes this calculation analytical. Thus, from the perspective of the true (unobserved) income process parameters, each household's survey response can be seen as a draw from a multinomial distribution with cell probabilities determined by the household's perceived income distribution.

[3]In particular, we model the probability π_{it} that household i at time t falls into the deterministic regime, putting all 10 points in one bin, as a logistic function of some observables Z_{it} (which may include demographics or indicators of financial sophistication, etc.) and an individual-specific random effect b_i capturing inherent willingness to give focal answers.

with reporting errors – because our goal is to recover income process parameters that can eventually be used in economic models of consumption and saving. This structural approach lets us perform likelihood-based comparisons and simulate implications in a way that a less parametric approach would not easily allow.

This paper makes four contributions. First, we show that income processes estimated from subjective expectations differ systematically from those inferred from realized income data, even for the same households. In particular, perceived income risk exhibits different persistence, dispersion, and tail behavior, implying that standard income processes calibrated on realizations may mischaracterize the uncertainty relevant for consumption and saving decisions. Second, we provide new evidence on nonlinear income dynamics from the perspective of households' beliefs, documenting state-dependent persistence and asymmetric perceived risks across the income distribution. These patterns mirror, but do not coincide with, nonlinearities estimated from realized income data, highlighting the importance of distinguishing perceived from objective income risk. Third, we develop an empirical framework that allows researchers to recover flexible income processes from probabilistic subjective expectations data, accounting for coarse probability reporting and focal responses. Forth, we integrate these processes into a life-cycle model of consumption and savings in the presence of income risk, and show the uncovered differences have first-order implications for precautionary saving, wealth accumulation, and the degree of consumption insurance.

A number of prior studies have explored using subjective expectations data to infer income risk or to test consumption models, but our approach differs in scope and methodology. Early contributions like Guiso, Jappelli, and Pistaferri (2002) demonstrated the value of subjective earnings expectations in explaining consumption and portfolio choices, while more recent works (Attanasio and Augsburg, 2016; Attanasio, Kovacs, and Molnar, 2020; Kovacs, Rondinelli, and Trucchi, 2021) have also incorporated income expectations to estimate earnings dynamics or to distinguish permanent vs. transitory shocks. More recent work has tried to incorporate information from subjective income expectations into macroeconomic models (Rozsypal and Schlafmann, 2023; Wang, 2023; Caplin, Gregory, Lee, Leth-Petersen, and Sæverud, 2023) to study the implications of perceived income risks on consumption and sav-

ings decisions. However, most of these studies assume relatively simple (linear or log-normal) income processes and often focus on first and second moments (e.g., variance of shocks) rather than the full distribution. In contrast, our framework flexibly allows for state-dependent and nonlinear dynamics – including heterogeneity in persistence of shocks and in conditional skewness or higher moments – which have been emphasized by the nonlinear panel data literature (Arellano et al., 2017; Guvenen et al., 2021). Thus, our work connects the expectations-based approach to the advanced income-process models.[4]

The rest of the paper is organized as follows. Section 2 describes the data we use and the probabilistic income expectations measure. In Section 3 we present the income process specifications (state-dependent AR(1) and nonlinear quantile models) and defines the main risk measures we focus on. Section 4 then details our estimation strategy for the subjective income process, introducing the multinomial likelihood framework and the hurdle model for reporting behavior, as well as the extension to nonlinear processes. In Section 5 we discuss the comparative results of estimating income risk from subjective expectations versus realized income. Section 6 considers the implications of our findings for household consumption and wealth accumulation and Section 7 assesses the information content in subjective expectations. Finally, Section 8 concludes.

2 Data

We use data from the Spanish Survey of Household Finances (EFF). It is a representative survey of the Spanish population with the goal of understanding household income and wealth. It contains detailed information on household assets, debts, income, consumption, and demographic characteristics. As it was especially designed for the study of household wealth, the

[4] A contemporaneous study by Arellano, Attanasio, Augsburg, and Crossman (2022) also develops methods to estimate nonlinear income processes using subjective expectations data, but there are two key differences. First, their analysis focuses on rural households in developing countries (Colombia and India) with survey questions that elicit expected income levels under hypothetical best- and worst-case scenarios (anchoring expectations at extremes), whereas we use data on expected growth rates of income with an anchoring at *approximately no change* as the central scenario. Second, our representative Spanish sample and richer dataset (which include panel data on income, consumption, and wealth) enable a comprehensive comparison between perceived income risk and actual income dynamics in a developed economy setting, as well as an exploration of how perceived risk correlates with household financial decisions. In this sense, our contribution is complementary: we extend the study of subjective income processes to a different context and provide methodological innovations to handle particular features of the data.

survey oversamples wealthy households based on individual wealth tax files. This oversampling permits a deeper exploration of fluctuations in the right tail of the income distribution that would not be possible with a conventional survey design. Moreover, there is an important rotating panel component, with a refreshment at each wave to ensure that the sample is representative of the population. These panel households can observed from at least two up to at most four waves. Currently, there are seven waves available for empirical analysis; the first six waves are triennial, from 2002-2020. Starting from the 2020 wave, however, the survey design changed from a triennial to a biennial frequency. As the question on household income expectations is only available from 2014, we focus on the 2014, 2017, 2020, and 2022 waves. We restrict our analysis to working-age households (heads aged 25–65) who have at least one working adult, to focus on active income processes and to avoid complications from pensions or retirement. We provide a detailed description of the variables, the selection criterion, and the summary statistics in Appendix A.

2.1 Eliciting subjective income expectations

Starting in 2014, the EFF introduced a question to elicit household income expectations. Household heads were asked to distribute ten points among five potential scenarios on total household income growth over the next twelve months.

The exact wording of the question 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. Share out 10 points among the five options given below, assigning more points to the options you think are more likely (assign 0 points 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%*
> - *Do not know*
> - *No answer*

Some important remarks about the particular formulation of this expectation question are in order. First, households are offered to distribute the probability mass across five distinct intervals. Second, respondents are asked to distribute 10 points instead of 100 points, making answering cognitively less demanding for respondents. Third, the question elicits the distribution using a probability density approach instead of a cumulative distribution one[5]. Despite some potential added difficulty in answering questions in a probabilistic format, most of the empirical evidence shows that respondents are willing to answer probabilistic questions.[6] Indeed the literature has emphasized that, if the questions concern well defined events that relate to respondents' lives, the responses are generally sensible and internally consistent, as individuals know a lot about themselves but much less about aggregate circumstances (Van der Klaauw, Bruine de Bruin, Topa, Potter, and Bryan (2008)).

2.2 Response behavior of households

Before we discuss how to estimate income processes, we first investigate household response behavior to the subjective income expectations question. In particular, following Bover (2015), we define four indicator variables: first, we define the variable *drop* as an indicator that takes on the value of one for households who placed all 10 points on the first two bins of the income expectations question. The variable *increase*, meanwhile, is an indicator that takes on the value of one for households who placed all 10 points in the last two bins of the income expectations question. The variable *bunching* is an indicator that takes on the value one for households who bunch all ten points in the middle bin. Finally, the variable *uncertainty* is an indicator that takes on the value one if the household is not in one of the three categories earlier described.

Table 1 provides summary statistics of household response behavior. The table shows that around 41.7% of households bunch in the middle bin across the four waves of the EFF. Bunching behavior, though, decreases over time. In particular, from 57.6% of households exhibiting

[5]The survey is administered by a computer assisted face to face interview. Explanations are provided by the interviewer when needed. Respondents are also handed out a sheet of paper containing the question and the response options on which they could draft their answers. An automatic prompt would appear on the screen whenever the answers entered in the computer by the interviewer do not add up to 10. In such cases, the household and the interviewer are asked to revise the answer. Finally, all these characteristics are also present in a house price expectation question that is formulated before the question on probabilistic expectations about future total household income.

[6]Approximately 2.5% of households do not respond to the subjective income expectations question across all waves.

bunching behavior in the EFF in 2014, this proportion has decreased to approximately 31.8% in the 2022 wave. This is followed by households who express some uncertainty in their answers to the income expectations question, who comprise 37.7% of households across waves.

Table 1: Household response behavior across time, income expectations question in the EFF

Wave	Drop	Increase	Uncertain	Bunching
2014	.0924	.0899	.2416	.5762
2017	.0794	.1236	.3176	.4793
2020	.1124	.0869	.4562	.3445
2022	.0888	.1438	.4490	.3184
ACROSS WAVES	.0942	.1107	.3772	.4179

Note: Proportion of households that exhibit the following reporting behaviors to the subjective income expectations question. *Drop* is a dummy variable that takes on one when households placed their ten points to the left bins. *Increase* is a dummy variable that takes on the value one when households placed their ten points to the right bins. *Uncertain* is a dummy variable that takes on one for households who spread their points across the bins. *Bunching* is a dummy variable that takes on the value one for households who spread all their points on the middle bin.

We then look at whether these characteristics are correlated with some demographics related to the household. To do so, we estimate we estimate linear probability models, where the dependent variables are each of the four indicators, respectively. The regression results are in Table 2. Households who have completed secondary education are less likely to bunch, and are more likely to express some uncertainty in their responses to the income expectations question. Female heads are likely to bunch. We also find that bunching in the middle becomes more likely as the household head becomes older. Self-employed household heads tend to bunch less, while those who have a permanent contract tend to bunch.

We also check the average probability distributions across all waves, and for each wave of the survey. These are displayed in Figure 1. As can be observed, across all waves, respondents place a majority of the probability mass in the stable region. On average, around 60 percent of the probability mass is in the stable region. There is evidence, though, of households placing less mass in the stable region when we compare the probability distributions by wave. In particular, From 68% of the mass placed in 2014, we find that 53% of the mass in 2022 is in the stable region. While most of the mass is in the middle, there is a tendency of households placing more probability mass in the right part of the distribution.

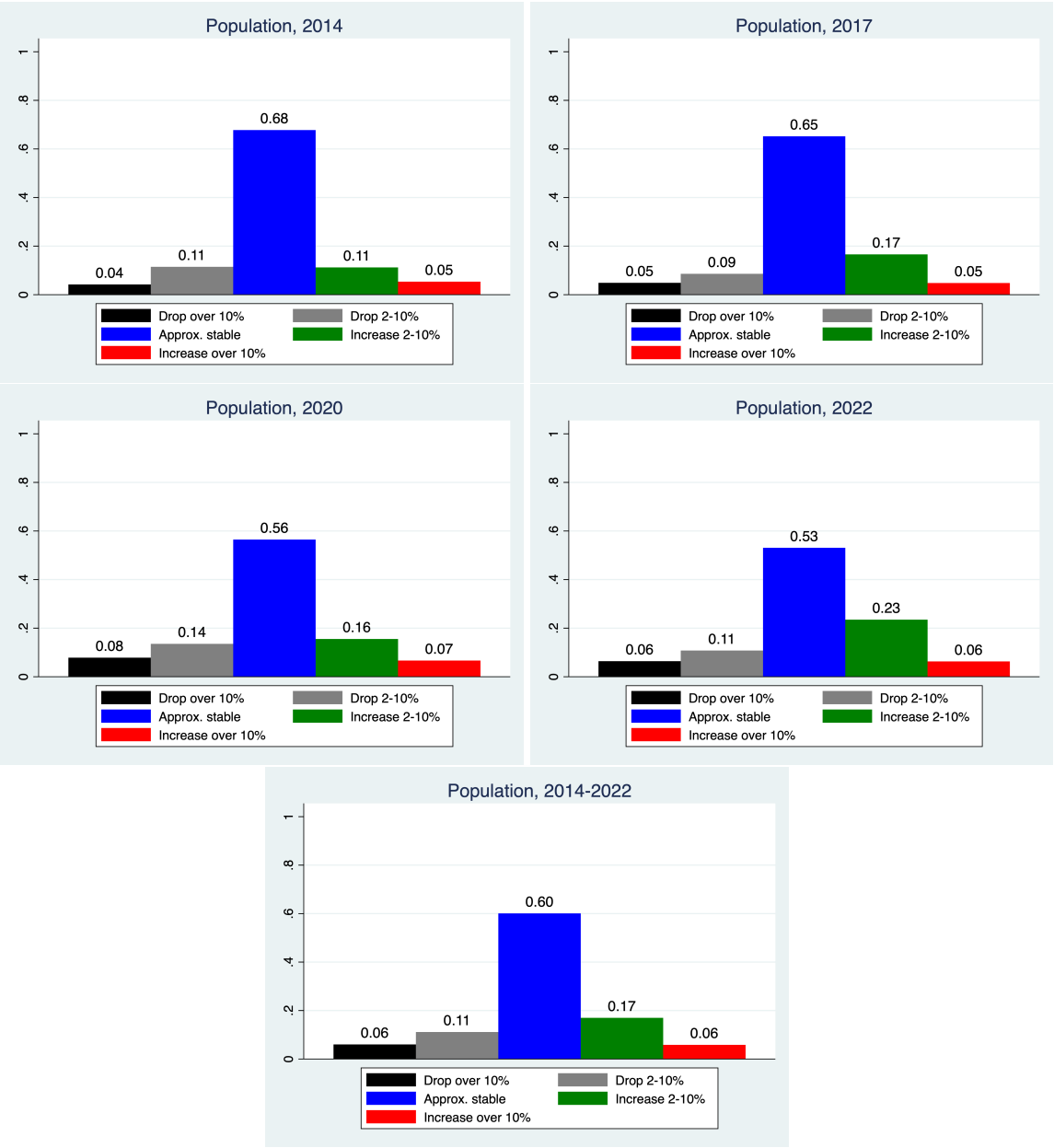
To summarize, the evidence we present in this section shows that bunching is prevalent in

Table 2: Correlates of household characteristics and response behavior

VARIABLES	(1) Bunching	(2) Drop in income	(3) Increase in income	(4) Uncertainty in income
Secondary education	-0.031*	-0.030**	0.027**	0.034*
	(0.018)	(0.012)	(0.014)	(0.019)
Higher education	-0.027	-0.045***	0.008	0.064***
	(0.022)	(0.012)	(0.015)	(0.022)
Female head	0.066***	-0.003	-0.018*	-0.046***
	(0.016)	(0.010)	(0.010)	(0.016)
Reference person is between 35 and 45	0.125***	0.005	-0.024	-0.105*
	(0.045)	(0.022)	(0.038)	(0.058)
Reference person is between 45 and 55	0.147***	0.039*	-0.034	-0.152***
	(0.043)	(0.022)	(0.038)	(0.057)
Reference person is between 55 and 65	0.232***	0.063**	-0.059	-0.237***
	(0.047)	(0.025)	(0.040)	(0.059)
Household head has a permanent contract	0.060***	-0.029**	-0.053***	0.021
	(0.023)	(0.013)	(0.016)	(0.022)
Household head is self-employed	-0.072***	0.034**	0.009	0.029
	(0.022)	(0.016)	(0.019)	(0.025)
Household head is risk averse	0.124***	-0.000	-0.009	-0.115**
	(0.043)	(0.023)	(0.029)	(0.056)
Household head exhibits medium risk	0.072	-0.008	-0.020	-0.044
	(0.045)	(0.025)	(0.031)	(0.059)
Household bunches in the house price question	0.249***	0.006	0.020	-0.275***
	(0.019)	(0.011)	(0.012)	(0.016)
Constant	0.132**	0.106***	0.145***	0.617***
	(0.065)	(0.034)	(0.047)	(0.083)
Observations	8,822	8,822	8,822	8,822
R-squared	0.110	0.017	0.011	0.121

Notes: These are the results of linear probability models that relate household response behaviors to the subjective income expectations question. *Drop* is a dummy variable that takes on one when households placed their ten points to the left bins. *Increase* is a dummy variable that takes on the value one when households placed their ten points to the right bins. *Uncertain* is a dummy variable that takes on one for households who spread their points across the bins. *Bunching* is a dummy variable that takes on the value one for households who spread all their points on the middle bin. Wave dummies are included in the estimation. Robust standard errors in parentheses. Significance levels *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure 1: Average probability distributions of future household income



Notes: These are the histograms of household responses to the income expectation questions, for each of the four waves, and for the entire dataset. Each bin corresponds to one of the five intervals of the subjective income expectations question. Drop over 10%: one-year future income growth drops below 10%. Drop 2-10%: one-year future income growth drops to around 2-10%. Approximately stable: income growth falls between -2% to 2%. Increase 2-10%: one-year future income growth increases from 2 to 10%. Increase over 10%: one-year future income growth rises above 10%.

the EFF. One interpretation of this result is that households who bunch might have cognitive difficulties in answering the subjective expectations question. Alternatively, it could also be the case that households are truly certain that their income will be stable year-on-year. Because both possibilities are plausible, we develop a model of the *subjective* income process that takes these features into account.

3 Income dynamics and objects of interest

Income processes are key to understand household consumption and savings decisions, and are key ingredients in the calibration and estimation of life-cycle models. These models are a standard tool to quantify, for example, the extent of consumption insurance, portfolio choice decisions, and household welfare (see, for example, Kaplan and Violante (2010), De Nardi et al. (2020) and Gálvez and Paz-Pardo (2026) for recent work). In this context, income processes are a representation of the uncertainty that households face when deciding how much to spend and how much to save.

In this paper, we consider two representations of income dynamics that have been used in quantitative macroeconomic models: a state-dependent autoregressive income process, and a nonlinear income process. We discuss the specification of each process, and the objects of interest that can be recovered from each representation. These objects of interest are what we will use to compare the implications of the *realized* and the *subjective* income processes.

3.1 State-dependent AR(1) income process

In the state-dependent AR(1) process, log income of household i at time t, y_{it}, takes the following form:

$$y_{it+1} = \rho(\mathbf{X}_{it})\, y_{it} + \mathbf{X}'_{it}\beta + \eta_i + \sigma(\mathbf{X}_{it})\, u_{it+1}, \qquad u_{it+1} \sim N(0,1), \tag{1}$$

where $\mathbf{X}_{it}$ are demographic characteristics that determine household income, its persistence and volatility, u_{it+1} are shocks to household income, and η_i is time-invariant unobserved heterogeneity. The state-dependent AR(1) process permits the possibility that the persistence and

volatility of household income are heterogeneous across households and across time[7]. To do so, we specify persistence ($\rho(\mathbf{X}_{it})$) and volatility ($\sigma(\mathbf{X}_{it})$) as functions of household characteristics. In this sense, the income process that we specify here shares the same spirit as those considered by Chamberlain and Hirano (1999), Meghir and Pistaferri (2004), Browning et al. (2010), Hospido (2012), Gu and Koenker (2017) and Almuzara (2020), who all consider different dimensions of heterogeneity in income dynamics. Notice that the canonical autoregressive process is a special case of equation (1); that is, $\rho(\mathbf{X}_{it}) = \rho$ and $\sigma(\mathbf{X}_{it}) = \sigma$.

The estimation of the state-dependent AR(1) income process allows us to recover the distributions of persistence and volatility. In particular, we can compute the mean and the variance of household income persistence

$$\mu_\rho = \mathbb{E}_{it}[\rho(\mathbf{X}_{it})] \text{ and } \sigma^2_\rho = \mathbb{V}_{it}[\rho(\mathbf{X}_{it})],$$

and the mean and variance of household income volatility

$$\mu_s = \mathbb{E}_{it}[\sigma(\mathbf{X}_{it})] \text{ and } \sigma^2_s = \mathbb{V}_{it}[\sigma(\mathbf{X}_{it})].$$

In both cases, the expectations are taken across individuals and over time.

Estimating the parameters of this income process using *realized* income data can be done via standard methods. In particular, we adopt a random effects approach and maximize the average log-likelihood, following Alvarez and Arellano (2022). More details on the estimation of this income process are detailed in Appendix B.

Finally, we consider an alternative earnings process where household persistence and volatility are potentially household specific but time-invariant; that is,

$$y_{it+1} = \rho_i y_{it} + \mathbf{X}'_{it}\beta + \sigma_i u_{it+1}, \quad u_{it+1} \sim N(0,1). \tag{2}$$

We detail the approach that we take to estimate the process, and the empirical results in Appendix D.

[7]The more general specification of this model is

$$y_{it+1} = \rho_{it} y_{it} + \mathbf{X}'_{it}\beta + \sigma_{it} u_{it+1}, u_{it+1} \sim N(0,1).$$

Estimating this model with realized income data can potentially be done via factor methods if the data contain a sufficiently long panel dimension (see Bonhomme and Denis (2025) for applications to static panel data models). We take a more parsimonious approach and specify them directly as functions of household characteristics.

3.2 Nonlinear income process

Recent work in the income dynamics literature has called into question the long-standing view that household income processes can be represented by linear models. In particular, Arellano et al. (2017) and Guvenen et al. (2021) present evidence, that contrary to the implications of the standard AR(1) income process, pre-tax household income exhibit skewness and volatility that change depending on the households' position in the income distribution and household age.

In light of this evidence, Arellano et al. (2017) propose a flexible representation of household earnings dynamics that allows for non-linearity, non-normality and age-dependence in the moments of the income distribution. We consider a simplified version of their model, where we model next-period income y_{it+1} via the following quantile model:

$$y_{it+1} = Q_y(u_{it+1} \mid y_{it}, \mathbf{X}_{it}; \theta), \quad u_{i,t+1} \sim U(0,1), \quad t \geqslant 1. \tag{3}$$

Intuitively, the quantile function maps random draws from the uniform distribution in (0,1) (i.e., cumulative probabilities) into draws of household income (i.e., quantiles). The standard and the state-dependent AR(1) processes can be seen as special cases of this generalized income process.

One way to understand the role of non-linearities within this framework is to consider a generalized notion of persistence, a measure of the *persistence of earnings histories*. Specifically, we can write the persistence of y_{it}, given y_{it-1}, when hit by a shock of rank τ as:

$$\rho(y_{it-1}, \tau) = \frac{\partial Q_y(\tau | y_{it-1})}{\partial y_{it-1}}, \tag{4}$$

which measures the persistence of lagged income y_{it-1} when households are hit by an income shock that has rank τ. The general model allows persistence to be a function of the past realization y_{it-1} and the sign and magnitude of the shock τ. Hence, this modeling strategy allows us to capture shocks that may be associated with different persistence, including the heterogeneous impact of unusual shocks like job losses or changes of career.

Also, as model (11) does not restrict the form of the conditional distribution of y_{it} given y_{it-1}, it allows for general forms of conditional heteroskedasticity. Indeed, we can consider

a measure of period-t uncertainty generated by the presence of shocks to earnings, for some $\tau \in (0.5, 1)$, as

$$\sigma(y_{it-1}, \tau) = Q_y(\tau|y_{it-1}) - Q_y(1-\tau|y_{it-1}). \tag{5}$$

Similarly, we can compute a measure for conditional skewness, for some $\tau \in (0.5, 1)$, as

$$sk(y_{it-1}, \tau) = \frac{Q_y(\tau|y_{it-1}) - Q_y(1-\tau|y_{it-1}) - 2Q_y(0.5|y_{it-1})}{Q_y(\tau|y_{it-1}) - Q_y(1-\tau|y_{it-1})}. \tag{6}$$

Finally, we can also generate simulated data from the estimated process, define a set of n grid points $\Xi = \{y_{1,t+1}, \ldots, y_{n,t+1}\}$, and use the simulated data to calculate the probabilities $\Pr(y_{i,t+1} \mid y_{j,t})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, n$ that form the discrete probability transition matrix $\mathbb{P}$ for all grid points.

Estimating this model using realized income data can be done via quantile regressions. We outline the approach that we take in Appendix B.

4 Estimating subjective income processes

One of the main goals of this paper is to compare the implications of *subjective* and *realized* income processes. In this section, we describe how to estimate *subjective* household income processes using subjective income expectations together with observed current income and covariates. The data we use do *not* include realized next-period income y_{it+1} at the time of estimation, as households do not observe y_{it+1} when they answer the subjective income expectations question; instead, we observe current income y_{it}, characteristics in the information set, and respondents' reported probability mass over coarse bins of *future* income growth.

A key challenge is that subjective expectations data are elicited as probabilities over pre-specified intervals (bins), which provides only coarse information about the distribution of y_{it+1} conditional on $(y_{it}, \mathbf{X}_{it})$. Moreover, reported probabilities may feature elicitation noise and focal-point responses (e.g., placing all points on a single bin), potentially reflecting cognitive constraints. Hence, one cannot rely on standard methods developed for realized income data to estimate the parameters of *subjective* income processes.

We propose a framework that addresses both issues. In the first subsection, we discuss how to estimate the state-dependent and nonlinear income processes using subjective income

expectations data without taking focal-point answers into account. We then consider the estimation of a model that takes into account focal-point responses in the second subsection. Throughout, we interpret y_{it} as (log) household income so that income growth is

$$\Delta y_{it} \equiv y_{i,t+1} - y_{it}.$$

4.1 Coarse information

Consider respondent i at time t. The survey asks her to allocate M points across $J = 5$ pre-specified intervals for future income growth. Let the growth cutoffs satisfy

$$-\infty = r_1 < \cdots < r_{J-1} < r_J = +\infty,$$

so that bin $j \in \{1, \ldots, J\}$ corresponds to $(r_{j-1}, r_j]$. The respondent observes current income y_{it} and covariates $\mathbf{X}_{it}$ that are part of her information set $\mathcal{I}_{it}$. To answer the subjective income expectations question, we assume that households have a perceived income process in mind. We discuss the estimation strategy when the households' income process is the state-dependent AR(1) process and the nonlinear income process, respectively.

4.1.1 State-dependent AR(1) process

Suppose the respondent's perceived income process is the state-dependent AR(1) process:

$$y_{it+1} = \rho(\mathbf{X}_{it})\, y_{it} + \mathbf{X}_{it}'\beta + \eta_i + \sigma(\mathbf{X}_{it})\, u_{it+1}, \qquad u_{it+1} \sim N(0,1).$$

To answer the subjective income expectations question, she computes the probabilities of future income falling in one of the pre-specified bins via model (1)[8].

Given $(y_{it}, \mathbf{X}_{it})$ and a value of η_i, the model implies a conditional Normal distribution for y_{it+1}. Define the corresponding level cutoffs

$$s_{j,it} \equiv y_{it} + r_j,$$

so that the event $\Delta y_{it} \in (r_{j-1}, r_j]$ is equivalent to $y_{i,t+1} \in (s_{j-1,it}, s_{j,it}]$. Then, the probability

[8]One can also do similar calculations assuming that the distribution of shocks is logistic, following Arellano et al. (2022). We use the Normal distribution as it is more widely assumed in income processes that are used in calibrated quantitative macroeconomic models.

that future income falls in bin j is

$$\begin{aligned} p^*_{j,it}(\eta_i) &\equiv \Pr\big(r_{j-1} < \Delta y_{it} \leqslant r_j \,\big|\, \mathcal{I}_{it}, \eta_i\big) \\ &= \Pr(y_{i,t+1} \leqslant s_{j,it} \mid \mathcal{I}_{it}, \eta_i) - \Pr(y_{i,t+1} \leqslant s_{j-1,it} \mid \mathcal{I}_{it}, \eta_i) \\ &= \Phi\left(\frac{s_{j,it} - \rho(\mathbf{X}_{it})y_{it} - \mathbf{X}'_{it}\beta - \eta_i}{\sigma(\mathbf{X}_{it})}\right) - \Phi\left(\frac{s_{j-1,it} - \rho(\mathbf{X}_{it})y_{it} - \mathbf{X}'_{it}\beta - \eta_i}{\sigma(\mathbf{X}_{it})}\right), \end{aligned} \tag{7}$$

where $\Phi(\cdot)$ is the standard Normal CDF.

Elicitation as multinomial sampling. Let $d_{jit} \in \{0, 1, \ldots, M\}$ denote the number of points assigned to bin j by respondent i at time t, with $\sum_{j=1}^{J} d_{jit} = M$. We interpret the survey mechanism as forcing the respondent to express her beliefs using M discrete points, and we model

$$\mathbf{d}_{it} \equiv (d_{1it}, \ldots, d_{Jit})' \sim \text{Multinomial}(M, \mathbf{p}^*_{it}(\eta_i)) \,, \mathbf{p}^*_{it}(\eta_i) \equiv \big(p^*_{1,it}(\eta_i), \ldots, p^*_{J,it}(\eta_i)\big)'. \tag{8}$$

This specification captures non-classical elicitation noise arising from discretization (and, if desired, can be augmented with additional reporting noise or rounding features).[9]

Random effects and integrated likelihood. Since η_i is unobserved, we complete the model by assuming

$$\eta_i \sim N(0, \sigma^2_\eta),$$

independent across i. Conditional on η_i, the per-period contribution is the multinomial probability mass. The individual integrated likelihood contribution is

$$L_i(\theta) = \int \prod_{t=1}^{T} \left[\frac{M!}{\prod_{j=1}^{J} d_{jit}!} \prod_{j=1}^{J} \big(p^*_{j,it}(\eta_i)\big)^{d_{jit}} \right] \frac{1}{\sigma_\eta} \phi\left(\frac{\eta_i}{\sigma_\eta}\right) d\eta_i, \tag{9}$$

where $\phi(\cdot)$ is the standard Normal pdf and

$$\theta \equiv \big(\text{parameters in } \rho(\cdot),\ \text{parameters in } \sigma(\cdot),\ \beta,\ \sigma_\eta\big).$$

The sample log-likelihood is

$$\ell_N(\theta) = \sum_{i=1}^{N} \log L_i(\theta), \qquad \widehat{\theta} = \arg\max_{\theta}\ \ell_N(\theta). \tag{10}$$

[9] One can argue that when households answer the subjective income expectations question, they discretize the probabilities with points that are different from the M in the survey. To take this into account, one can allow for an additional reporting layer (e.g., latent M_i). We take M as fixed and known from the survey design. We experimented with specifications where we take M as latent and estimated from the data. Our experiments resulted with M being close to ten.

Because $L_i(\theta)$ involves an integral over η_i with nonlinear dependence through (7), we approximate it numerically by Gauss–Hermite quadrature with 12 points. We compute standard errors via non-parametric bootstrap with 1,000 replications.

Implementation. A key modeling choice is the specification of $\rho(\mathbf{X}_{it})$ and $\sigma(\mathbf{X}_{it})$. In our empirical implementation, we use a logistic specification for persistence and an exponential specification for dispersion:

$$\rho(\mathbf{X}_{it}) = \Lambda(\rho_0 + \mathbf{X}_{it}'\rho_1)\,, \qquad \sigma(\mathbf{X}_{it}) = \exp(\sigma_0 + \mathbf{X}_{it}'\sigma_1)\,,$$

where $\Lambda(z) = \exp(z)/(1+\exp(z))$. Following the literature, $\mathbf{X}_{it}$ includes education dummies, indicators for the presence of other income earners, contract type, and a self-employment indicator. We may also include (log) income y_{it} as an additional state variable in the dispersion specification, in which case $\sigma(\cdot)$ is understood as $\sigma(\mathbf{X}_{it}, y_{it})$.

4.1.2 Non-linear income process

Motivated by recent work in income dynamics, we also consider a nonlinear income process in the spirit of Arellano et al. (2017). In this specification, households' perceived next-period income is modelled by a conditional quantile function:

$$y_{it+1} = Q_y(u_{i,t+1} \mid y_{it}, \mathbf{X}_{it}; \theta), \quad u_{i,t+1} \sim U(0,1), \quad t \geqslant 1. \tag{11}$$

Following Arellano et al. (2017), we parameterize the quantile model as:

$$Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) = \sum_{k=0}^{K} a_k(\tau)\, \psi_k(y_{it}) + \mathbf{X}_{it}'\beta(\tau), \tag{12}$$

where ψ_k are low-order Hermite polynomials, and the coefficients $\{a_k(\tau)\}_{k=0}^{K}$ and $\beta(\tau)$ are modeled as piecewise-linear splines in τ over a grid $\{\tau_1, \ldots, \tau_L\} \in (0,1)$. To discipline behavior near $\tau = 0$ and $\tau = 1$, we impose exponential-tail restrictions at the boundary knots with parameters λ_1 and λ_L. This permits the implied conditional cdf

$$F_{it}(y;\theta) \equiv \Pr(y_{i,t+1} \leqslant y \mid y_{it}, \mathbf{X}_{it}; \theta)$$

to have closed-form expressions in the tails. A particular advantage of this specification is that we can write an approximation to the conditional CDF, even if the model is quite flexible. We write the complete form of the conditional CDF in Appenidix C.1.

Given the income process, the household calculates the model-implied bin probabilities:

$$p_{j,it}(\theta) = \Pr(y_{i,t+1} \in (s_{j-1,it}, s_{j,it}] \mid y_{it}, \mathbf{X}_{it}; \theta) = F_{it}(s_{j,it}; \theta) - F_{it}(s_{j-1,it}; \theta). \tag{13}$$

Interpreting the M reported points as M exchangeable draws from this subjective distribution, the observed-data log-likelihood is

$$\ell_N(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\log \frac{M!}{\prod_{j=1}^{J} d_{jit}!} + \sum_{j=1}^{J} d_{jit} \log \left(p_{j,it}(\theta) \right) \right]. \tag{14}$$

As directly maximizing (14) can be computationally demanding under spline parameterizations and/or potential monotonicity constraints, we implement a stochastic EM algorithm as a practical estimation device. We describe the algorithm in the main text, and relegate the full details in Appendix C.2.

Stochastic EM algorithm. For a given parameter iterate $\theta^{(s)}$, define the bin interval in income levels

$$I_{j,it} \equiv (s_{j-1,it}, s_{j,it}].$$

Interpreting the M reported points as M latent draws from the subjective distribution, we treat the unobserved latent outcomes $\{y_{i,t+1}^{(m)}\}_{m=1}^{M}$ as missing data, where we only observe their bin membership through $\mathbf{d}_{it}$.

Given $\theta^{(s)}$, the conditional density of a latent draw restricted to bin j is the truncated version of the model-implied density:

$$f\left(y \mid y_{it}, \mathbf{X}_{it},\, y \in I_{j,it}; \theta^{(s)}\right) \propto f\left(y \mid y_{it}, \mathbf{X}_{it}; \theta^{(s)}\right)\, \mathbf{1}\{y \in I_{j,it}\}. \tag{15}$$

Starting from an initial guess $\theta^{(0)}$, we iterate on the following two steps from $s = 1, 2, \ldots$ until convergence:

1. **Stochastic E-step.** For each (i, t) and each bin j, draw d_{jit} latent values from the truncated distribution (15). Collect the pooled augmented sample $\{y_{i,t+1}^{(s,m)}, y_{it}, \mathbf{X}_{it}\}_{i,t,m}$ with exactly M draws per (i, t).

2. **M-step.** Update the spline parameters by estimating quantile regressions at the spline knots $\{\tau_\ell\}_{\ell=1}^{L}$ on the augmented sample. For each knot τ_ℓ, estimate

$$(\widehat{a}(\tau_\ell), \widehat{\beta}(\tau_\ell)) = \arg\min_{a(\tau_\ell), \beta(\tau_\ell)} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \rho_{\tau_\ell} \left(y_{i,t+1}^{(s,m)} - \sum_{k=0}^{K} a_k(\tau_\ell) \psi_k(y_{it}) - \mathbf{X}_{it}' \beta(\tau_\ell) \right),$$

where $\rho_\tau(\cdot)$ is the standard check function.

We monitor convergence using the observed-data log-likelihood (14), evaluated at each iterate $\theta^{(s)}$, and stop when

$$\frac{\left|\ell_N(\theta^{(s+1)}) - \ell_N(\theta^{(s)})\right|}{1 + \left|\ell_N(\theta^{(s)})\right|} < \varepsilon$$

for a tolerance ε. As in the state-dependent process, we compute standard errors via non-parametric bootstrap with 1,000 replications.

4.2 A hurdle model for bunching behavior

A large fraction of respondents place all M points in the middle bin. To capture this focal-point behavior, we augment the model with a hurdle (two-regime) reporting structure by adapting the approach in de Bresser and van Soest (2013), Kleinjans and van Soest (2014), and Heiss et al. (2022), who model rounding and expectation types in household stock market expectations questions via a finite mixture model, into our context[10]. We illustrate the hurdle model and its estimation in the context of the state-dependent AR(1) model in the main text, and leave the estimation when the underlying income process is non-linear in Appendix C.3.

Let $j^\star$ denote the index of the middle bin, and define the observed bunching indicator

$$S_{it} \equiv \mathbf{1}\{d_{j^\star,it} = M\}.$$

We assume that, when answering, respondents face a sequential decision:

1. They decide whether to report deterministically by placing all M points in the middle bin.

2. If not, they allocate points according to the multinomial mechanism implied by the income process.

Let $R_{it} \in \{0, 1\}$ indicate whether the respondent is in the deterministic focal-point regime. We model

$$\pi_{it}(b_i) \equiv \Pr(R_{it} = 1 \mid \mathbf{Z}_{it}, b_i) = \Lambda(\gamma_0 + \mathbf{Z}_{it}'\gamma_1 + b_i)\,, \tag{16}$$

[10] A different, nonparametric approach to dealing with rounding behavior is the inferential method in Manski and Molinari (2010) and Giustinelli et al. (2022b), who transform individual point expectations into intervals based on their answers. As the goal here is to estimate an income process, we decided on a more parametric approach.

where b_i is a time-invariant respondent-specific random effect governing focal-point propensity, and $\mathbf{Z}_{it}$ collects covariates that may overlap with $\mathbf{X}_{it}$ but can also include exclusion restrictions (e.g., $\mathbf{Z}_{it} = (\mathbf{X}_{it}, \mathbf{W}_{it})$ where $\mathbf{W}_{it}$ affects reporting behavior but not the income process).

In the baseline, we assume independent random effects[11]:

$$\eta_i \sim N(0, \sigma_\eta^2), \qquad b_i \sim N(0, \sigma_b^2), \qquad \eta_i \perp b_i.$$

Likelihood function. Conditional on (η_i, b_i), define the probability of placing all points in the middle bin under the multinomial mechanism as

$$A_{it}(\eta_i) \equiv \Pr\left(d_{j^\star,it} = M \mid \eta_i, \mathcal{I}_{it}\right) = \left(p^*_{j^\star,it}(\eta_i)\right)^M,$$

where the $p^*_{j^\star,it}(\eta_i)$ is the true probability calculated from the state-dependent AR(1) process (1).

Then, the per-period likelihood contribution is

$$L_{it}(\eta_i, b_i) = \begin{cases} \pi_{it}(b_i) + \left(1 - \pi_{it}(b_i)\right) A_{it}(\eta_i), & \text{if } S_{it} = 1, \\ \left(1 - \pi_{it}(b_i)\right) \dfrac{M!}{\prod_{j=1}^J d_{jit}!} \displaystyle\prod_{j=1}^J \left(p^*_{j,it}(\eta_i)\right)^{d_{jit}}, & \text{if } S_{it} = 0. \end{cases} \tag{17}$$

The two cases have a clear interpretation:

- If $S_{it} = 1$, bunching can arise either because the respondent is in the deterministic regime (with probability $\pi_{it}(b_i)$) or because she is not deterministic but the multinomial mechanism assigns all M points to the middle bin (with probability $(1-\pi_{it}(b_i))A_{it}(\eta_i)$).
- If $S_{it} = 0$, the deterministic regime has zero probability mass, so the likelihood is the multinomial likelihood scaled by $(1 - \pi_{it}(b_i))$.

The individual integrated likelihood is

$$L_i^f(\theta) = \iint \prod_{t=1}^T L_{it}(\eta_i, b_i)\, f(\eta_i)\, f(b_i)\, d\eta_i\, db_i, \tag{18}$$

[11] Allowing correlation is straightforward by replacing the diagonal covariance matrix with a full bivariate Normal covariance matrix.

and the sample log-likelihood is

$$\ell^f(\theta) = \sum_{i=1}^{N} \log L_i^f(\theta), \qquad \widehat{\theta} = \arg\max_{\theta} \ell^f(\theta).$$

We estimate the model via maximum likelihood, and approximate the two-dimensional integrals in (18) using Gauss–Hermite quadrature. We calculate standard errors via a non-parametric bootstrap with 1,000 replications.

5 Results

This section compares the income processes recovered from subjective expectations with those obtained from realized income data only, focusing on the state-dependent AR(1) model and the nonlinear model. A consistent pattern emerges across specifications: households substantially overstate persistence and understate dispersion of future income relative to the data-generating process inferred from realized outcomes.

5.1 Coarse information

We first discuss the empirical results from the baseline models.

5.1.1 State-dependent AR(1) process

Table 3 presents the mean and the standard deviation of income persistence and volatility according to the realized and perceived income processes. Our estimates indicate that there are substantial differences between perceived and realized income dynamics. First, persistence implied by *realized* income data is on average 0.6773, with a standard deviation of 0.0514. In contrast, *subjective* expectations imply average persistence of 0.9972, with a standard deviation of 0.0012. Second, the dispersion of shocks is markedly smaller in the subjective income process: realized income data imply variability of 0.4651, with a standard deviation of 0.0808, while subjective income data imply a variability of 0.0429, with a standard deviation of 0.0121. Figure 2 further underscores the differences between perceived and realized income dynamics, as the shape of the distributions are markedly different. Thus, our results underscore two stylized facts: households *overestimate* the persistence of future income, which is consistent

with the evidence in Rozsypal and Schlafmann (2023), and *underestimate* the dispersion of the distribution of income.

Table 3: Mean and standard deviation of ρ_{it} and σ_{it}, state-dependent AR(1)

	Mean	Standard Deviation
Panel A: Realized income process		
Persistence	0.6773	0.0514
	[0.6505, 0.7036]	[0.0316, 0.0750]
Volatility	0.4651	0.0808
	[0.4474, 0.4802]	[0.0669, 0.0976]
Panel B: Subjective income process		
Persistence	0.9972	0.0012
Volatility	0.0429	0.0121

Notes: This table shows estimates of the average and the standard deviation of persistence, $\mathbb{E}[\rho_{it}]$ and $\varsigma[\rho_{it}]$, and the average and the standard deviation of income volatility, $\mathbb{E}[\sigma_{it}]$ and $\varsigma[\sigma_{it}]$, for the realized and the subjective income processes. We include wave dummies and control for education and employment characteristics. We report in brackets the 95% block bootstrap confidence intervals, computed with 1,000 bootstrap replications.

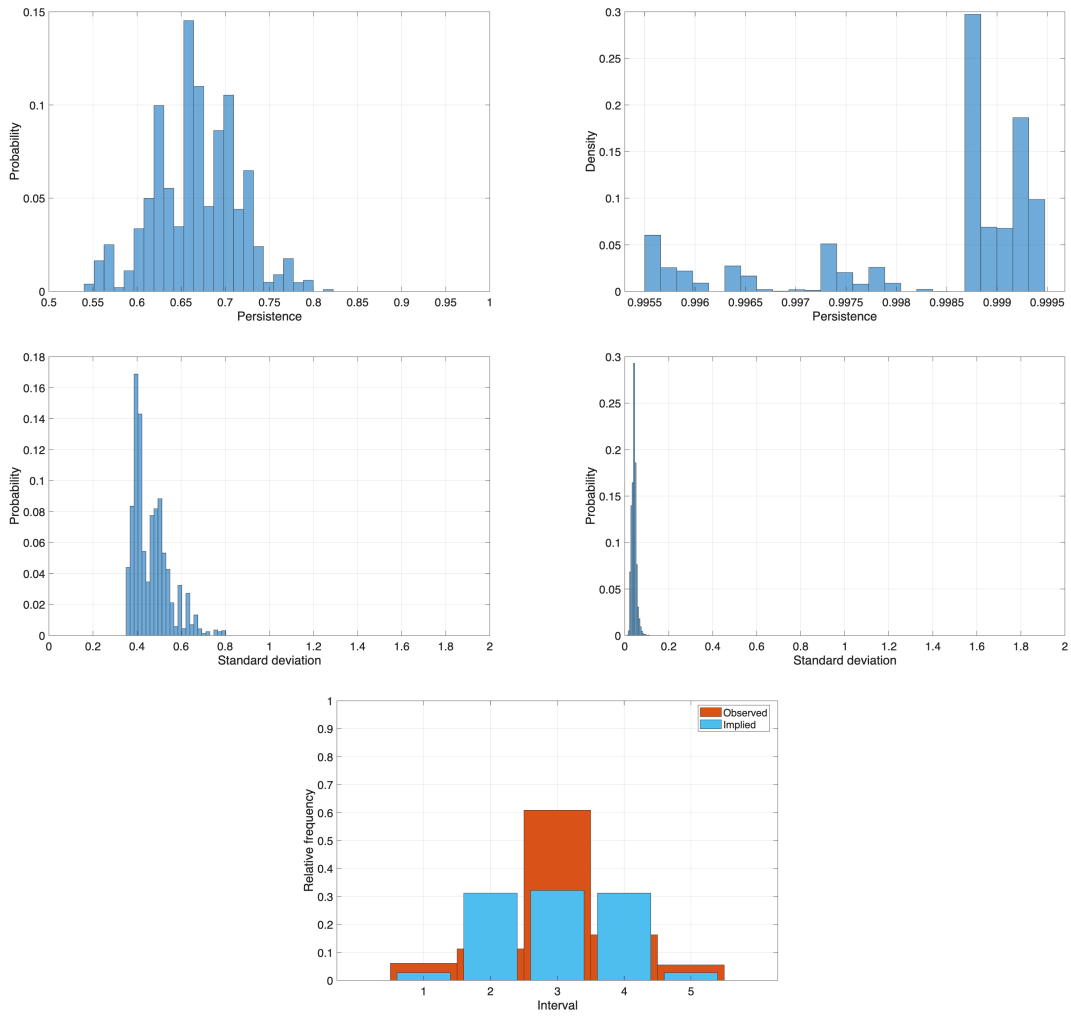
The model fit echoes this divergence. As the bottom panel of Figure 2 shows, the AR(1) model distributes mass too heavily around the central bins and under-predicts the likelihood of extreme outcomes, producing a root-mean-squared error (RMSE) of 0.25. This highlights the difficulty of reconciling coarse, highly bunched probability allocations with a smooth Gaussian process.

Household-specific dispersion. In Appendix D, we present the results for the AR(1) process with household-specific dispersion parameters. Although this model produces a better fit (with a RMSE of 0.05), the qualitative conclusions remain unchaned: households systematically overestimate persistence and underestimate risk.

5.1.2 Nonlinear income process

Figures 4 reports results from the nonlinear quantile-based income model without the hurdle component. First, the generalized persistence surface implied by realized income data displays rich nonlinearities: persistence varies with both the income level and the rank of the shock. By contrast, the subjective model generates an almost flat persistence surface, indicating that

Figure 2: Distributions of persistence and dispersion, and model fit.



Notes: These plots show the distribution of persistence (top row) and dispersion (middle row) for the state-dependent AR(1) process without modeling bunching behavior. The left columns correspond to the realized income process, while the right columns correspond to the subjective income process. The bottom row compares model fit by plotting the implied probabilities of the subjective income process to the observed probabilities.

households perceive shocks to be similarly persistent across the distribution, regardless of the type of shock that households receive. Second, subjective expectations again imply a compressed distribution of risk. Conditional volatility is uniformly smaller than what is inferred from realized income data, reinforcing the finding that households substantially under-perceive uncertainty. Third, while conditional skewness changes depending on household rank in realized income data, it does not appear to be the case in subjective income data. In particular, households perceive skewness to be quite negative, regardless of their rank in the income distribution. Compared to realized income data, the nonlinear income process estimated using subjective income data imply that households put weight in extremely negative events. The model also displays quite a good fit, as the RMSE is around 0.0045. Figure E.2 shows the 95% non-parametric bootstrap confidence intervals. As can be observed, the estimates are precisely estimated.

5.2 Hurdle model for bunching behavior

We then present the empirical results when we explicitly model the focal point answer via a hurdle model.

5.2.1 State-dependent AR(1) process

We report results for the following variation of the model with the state-dependent AR(1) process:

$$y_{it+1} = \rho y_{it} + \eta_i + \mathbf{X}_{it}\beta + \sigma(X_{it})u_{it},$$

as we find that for the subjective income process, the estimated distribution of $\rho(\mathbf{X}_{it})$ clusters around 0.99. We first report the estimates for persistence, and the moments of the distribution of dispersion.

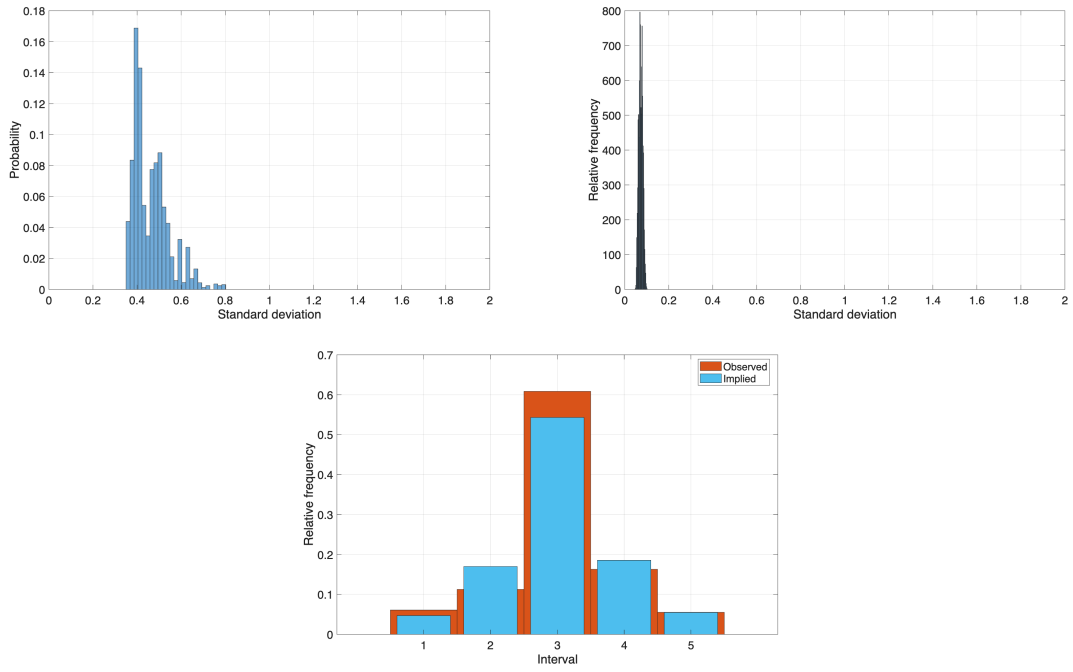
Allowing for focal-point reporting improves model fit substantially (Figure 3). The hurdle extension raises estimated dispersion to an average of 0.0732 while leaving persistence estimates essentially unchanged. The model now captures the shape of the observed bin probabilities much more closely, especially in the tails, reducing the RMSE to 0.12. Nevertheless, the middle bin remains difficult to match exactly: focal-point mass is sufficiently large that even a two-regime mixture cannot fully account for it.

Table 4: Mean and standard deviation of ρ_{it} and σ_{it}, state-dependent AR(1)

Panel A: Realized income process		
Persistence	0.6773	-
	[0.6505, 0.7036]	-
Volatility	0.4651	0.0808
	[0.4474, 0.4802]	[0.0669, 0.0976]
Panel B: Subjective income process		
	Mean	Standard Deviation
Persistence	0.9992	-
	[0.9539, 0.9994]	
Volatility	0.0732	0.0092
	[0.0620,0.1426]	[0.0074,0.5629]

Notes: This table shows estimates of the average and the standard deviation of persistence, $\mathbb{E}[\rho_{it}]$ and $\varsigma[\rho_{it}]$, and the average and the standard deviation of income volatility, $\mathbb{E}[\sigma_{it}]$ and $\varsigma[\sigma_{it}]$, for the realized and the subjective income processes. We include wave dummies and control for education and employment characteristics. We report in brackets the 95% block bootstrap confidence intervals, computed with 1,000 bootstrap replications.

Figure 3: Distributions of persistence and dispersion, and model fit.



Notes: These plots show the distribution of dispersion (top row) for the state-dependent AR(1) process with modeling bunching behavior. The left column corresponds to the realized income process, while the right column corresponds to the subjective income process. The bottom row compares model fit by plotting the implied probabilities of the subjective income process to the observed probabilities.

5.2.2 Nonlinear process

Most of the qualitative findings that we observe in Figure 4 are echoed in Figure 5, with some substantial differences. While the measure of conditional variance is still smaller than that implied by the realized income process, the average value increases from 0.02 in the perceived income process without a hurdle, to around 0.06 in the perceived income process correcting for hurdle behavior. The most remarkable change is with conditional skewness: the subjective income process implies the conditional skewness that is consistent with the patterns in Arellano et al. (2017), with high-income households perceiving negative skewness, while low income households perceiving no skewness in the future income distribution. These results highlight the importance of correcting for hurdle behavior. Figure E.3 shows the the 95% non-parametric bootstrap confidence intervals. As in the previous case, we find that our estimates are precisely estimated.

Taken together, the estimates from the nonlinear income process in the subjective expectations data confirm the main findings from the AR(1) setting: households view income as highly persistent and only weakly volatile, and subjective expectations substantially smooth over the nonlinear features present in realized income dynamics.

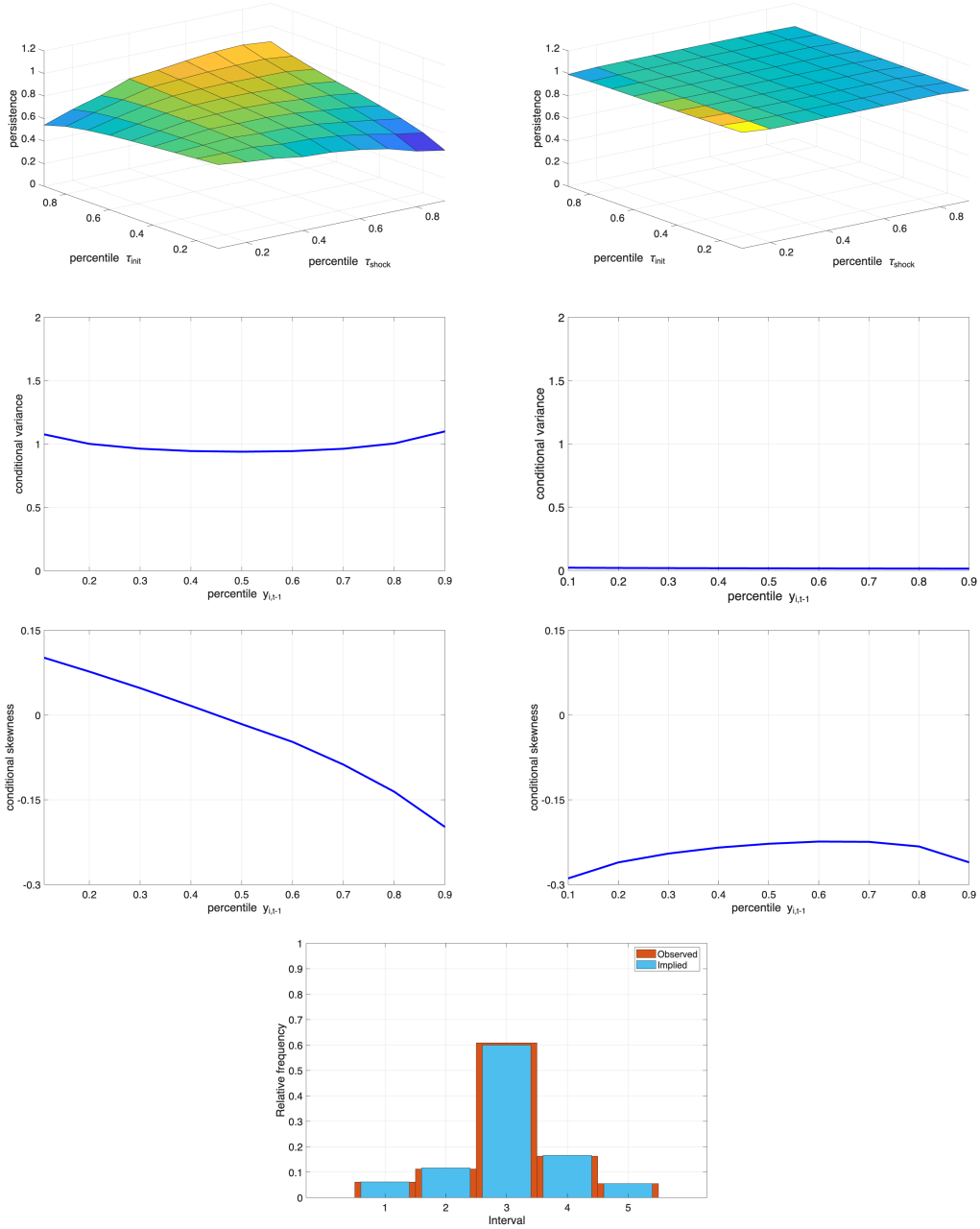
6 Household consumption insurance and wealth accumulation implications

We employ a life-cycle model with heterogeneous agents and incomplete asset markets to study the key implications of the nonlinearities we uncover for wealth accumulation over the life cycle and for consumption insurance. We first present the main features of our model economy and describe its calibration, discussing how we embed the estimated income process into the framework. The main insights are presented next.

6.1 Model

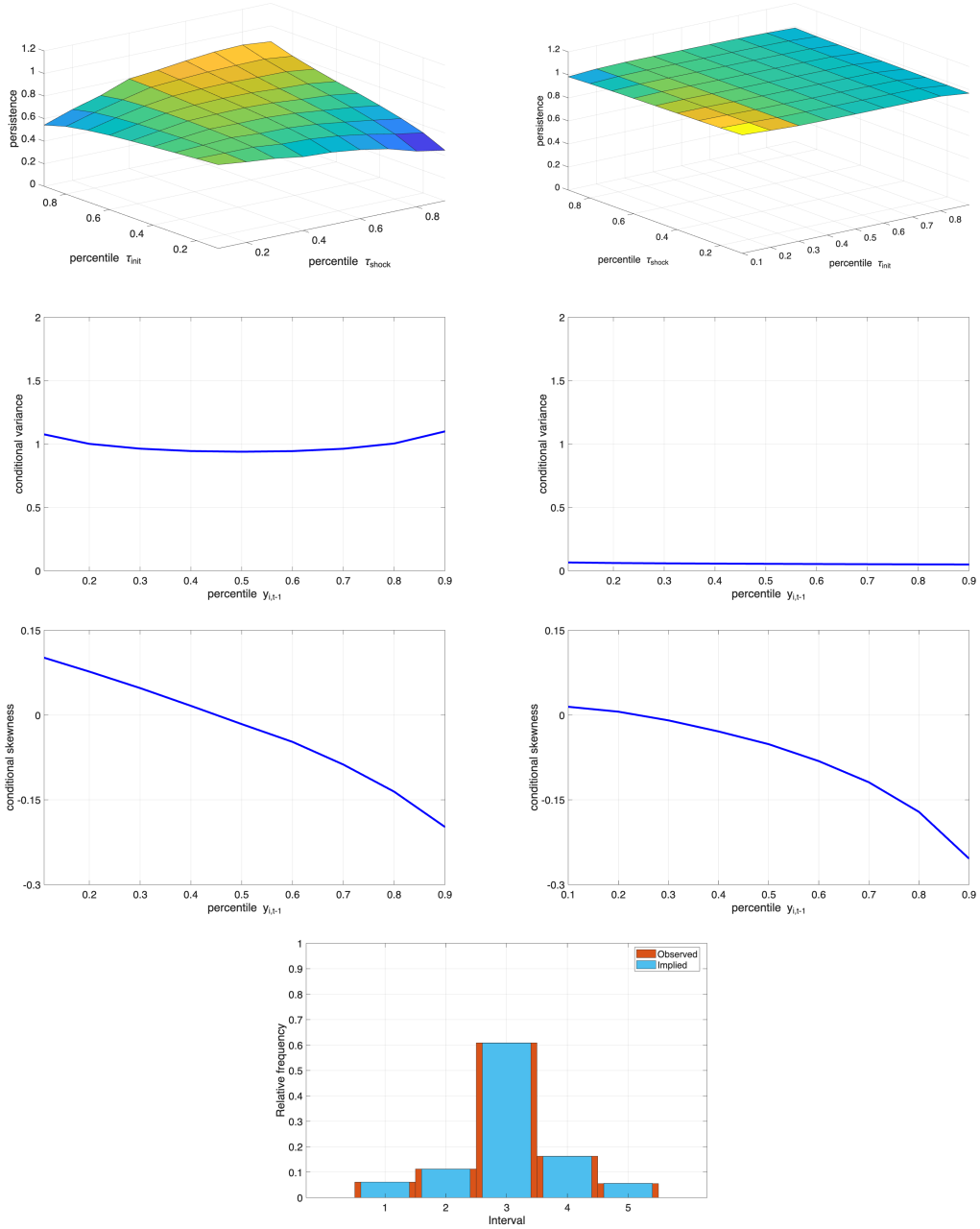
Each period a continuum of agents are born. Agents live a maximum of T periods, facing a probability s_j of surviving up to age j conditional upon being alive at age $j-1$ (survival

Figure 4: Nonlinear persistence, dispersion, skewness, and model fit.



Notes: These plots show the measures of nonlinear persistence (first row), conditional volatility (second row), and conditional skewness (third row) for the nonlinear process without hurdle model. The left columns correspond to the realized income process, while the right columns correspond to the subjective income process. The bottom row compares model fit by plotting the implied probabilities of the subjective income process to the observed probabilities.

Figure 5: Nonlinear persistence, dispersion, skewness, and model fit.



Notes: These plots show the measures of nonlinear persistence (first row), conditional volatility (second row), and conditional skewness (third row) for the nonlinear process with hurdle behavior. The left columns correspond to the realized income process, while the right columns correspond to the subjective income process. The bottom row compares model fit by plotting the implied probabilities of the subjective income process to the observed probabilities.

probability is zero at period $T+1$). Population grows at a constant rate g_n. [12] Furthermore, we assume agents are born with no assets and face a mandatory retirement age of $j = J_R$. The asset holdings of agents who die are discarded.

Age j agents select how much to consume c_j and next period asset holdings a_{j+1} (subject to a borrowing limit A) to maximize their utility $u(c_j)$. We assume retirees do not work and receive a pension PB_j that depends on their income before retirement. Agents in working age ($j \leqslant J_R$), supply one unit of labor inelastically at zero disutility, receiving income y_j, which is stochastic and follow the non-linear estimated process with realized data (NL) or the non-linear estimated process using subjective expectations with hurdling (NL-SE). During the working ages, an agent pays income taxes, denoted by T_j, depending on their total income $I_j = y_j + ra_j$, where r is the net return on asset holdings a_j. Agents pay income taxes $\tau(I_j)I_j$, where the tax rate $\tau(I_j)$ is a function of total income. Finally, agents may receive transfers $G(I_j)$. [13].

Workers - age $j < J_R$ solve the recursive problem

$$
\begin{aligned}
V(t,y,a) &= max_{a',c}\{c^{1-\sigma}/(1-\sigma) + \beta s_t E_t V(t+1,y',a')\},\\
s.t. \quad a' &= y - T_j(I_j) + G(I_j) - c + (1+r)a, a' \geqslant A\\
y_t &= Q_t^P(y_{t-1},u_t),\ u_t \sim U[0,1]\\
&\text{where } P = \{NL, \quad NL-SE\}
\end{aligned}
$$

Retirees - age $J_R \leqslant j \leqslant T$ solve the recursive problem

$$
\begin{aligned}
W(t,p,a) &= max_{a',c}\{c^{1-\sigma}/(1-\sigma) + \beta s_t E_t V(t+1,p,a')\},\\
s.t. \quad a' &= PB - c + (1+r)a, a' \geqslant A
\end{aligned}
$$

[12] We assume that age-j agents always constitute a fraction μ_j of the population at any point in time, as such the demographic structure is stationary. The weights μ_j are normalized to sum to 1, and are given by the recursion $\mu_{j+1} = \frac{s_{j+1}}{(1+g_n)}\mu_j$.

[13] We assume

$$
\tau(I_j) = \begin{cases} 0 \text{ if } I_j \leqslant \tilde{I} \\ max\{1 - \lambda(I_j/\bar{I})^{-\tau_I}, 0\} \text{ if } I_j > \tilde{I}, \end{cases} \tag{19}
$$

where $\bar{I}$ is the mean income and $\tilde{I}$ is a tax income threshold, and government transfers are given by

$$
G(I_j) = \begin{cases} g_0\bar{I} \text{ if } I_j = 0 \\ max\{g_1 - g_2(I_j/\bar{I}), 0\}\bar{I} \text{ if } I_j > 0. \end{cases} \tag{20}
$$

6.2 Calibration

Individuals start life at age 25, retire at age 65 and live up to a maximum possible age of 85. This implies that $J_R = 41$ and $T = 61$. The population growth rate is 0.6% per year, corresponding to the actual growth rate for the period 1990-2024 in Spain (Data Source: World Development Indicators). Survival probabilities (s_j, for $j = 25$ to 85) are set according to the data from the Spanish Statistical Office (INE) for the year 2023.

The coefficient of relative risk aversion is set to 2, a standard value. The risk-free rate is 2.5% and the discount factor β is calibrated to match a wealth to income ratio of 4 (calculate using EFF data). Finally, we set the exogenous borrowing limit A to 16% of average disposable household earnings, matching the liquid debt to income ratio of households who have liquid debt in the EFF.

Parameters λ, τ_I and $\tilde{I}$ determine the non-linear income tax function in the model. We set $\lambda = 0.91255$ and $\tau_I = 0.1209$ and $\tilde{I} = 0.47\bar{I}$, following García-Miralles, Guner, and Ramos (2019). We set pension benefits such that the average benefit ratio is equal to 64.1% (the benefit ratio in 2022; see the 2024 Ageing Report - European Commission). Furthermore, the minimum pension is set to match the income at the percentile 15% of the income distribution before retirement and the maximum at percentile 75% roughly matching the redistribution features of the pension system in Spain. The values of g_0, g_1 and g_2 of the transfer function $G(I)$ are estimated by Guner, Kaya, and Sánchez-Marcos (2024), using the data from the EU-Statistics on Income and Living Conditions. We follow their results and set $g_0 = 0.138$, $g_1 = 0.03$ and $g_2 = 0.0129$.

We use the estimation results of the income process presented above to simulate data for 10000 households. We then discretize $z \in [-\infty, \infty]$ into a vector $z = [z_1, ..., z_n]$ and utilize the simulated data to obtain the transition matrix $Q(z_i \mid z_h)$ for $i, h \in [1, n]$. As an illustration, Figure 6 depicts a subset of rows of the transition matrices for $n = 18$.[14]

[14]The Survey of Household Finance is conducted at three-year intervals; thus, the time period in the simulated data is three years. In the model, each period corresponds to one year, and hence we must adjust the transition matrix based on realized data. For the subjective expectations case the question explicitly refers to income variation in 12 months, no adjustment is needed. Let Q_3 denote the matrix from the empirical estimates and Q_1 the modified matrix to be used for the yearly transitions. To ensure consistency, we must have $Q_1^3 = Q_3$. We first calculate $\tilde{Q}_1 = Q_3^{(1/3)}$. However, given the nonlinearities present in the estimated income process and reflected in matrix Q_3, $\tilde{Q}_1$ may contain negative elements. When that occurs, we modify $\tilde{Q}_1$ such that, for each row (initial z), the expected income in the next period, $E(z' \mid z)$, and its variance, $VAR(z' \mid z)$, remain unchanged, and the

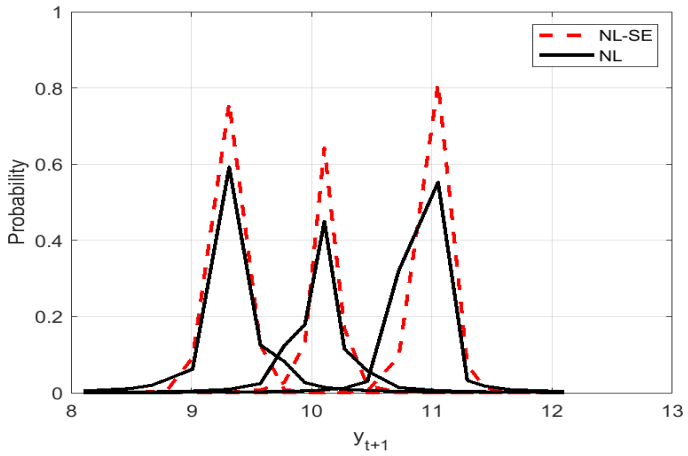


Figure 6: Transition Matrices

These matrices further illustrate the differences between the estimated processes using realized data (NL) and subjective expectations data (NL-SE). First, variances are smaller under NL-SE. Second, the process estimated from realized data shows a substantially greater risk of large negative shocks for high-income households and a higher probability of large positive shocks for low-income households, relative to the corresponding probabilities for a median-earning household; risks are nonlinear and depend on current income. We do not observe these nonlinearities in the NL-SE process.

6.3 Results

We focus on two key simulated outcomes of our life cycle model: wealth accumulation and consumption insurance.

Figure 7 plots average wealth over the life cycle. Under the estimated process based on realized data (NL), given the higher variance and, crucially, the presence of a large risk of negative shocks to high-income households, households save significantly more on average at the beginning of the life cycle. Under the estimated process using subjective expectations data (NL-SE), the variance of income shocks is substantially lower, so the pace of wealth accumulation is also significantly lower: households that are able to save engage in less precautionary saving. As the richer save more due to nonlinear risk, the economy under the NL also display

sum of the three elements around the diagonal remains the same (for row j, the elements in columns $j-1$, j, and $j+1$ are summed, while for the first and last rows, the two elements to the right and to the left of the diagonal, respectively, are used in the sum).

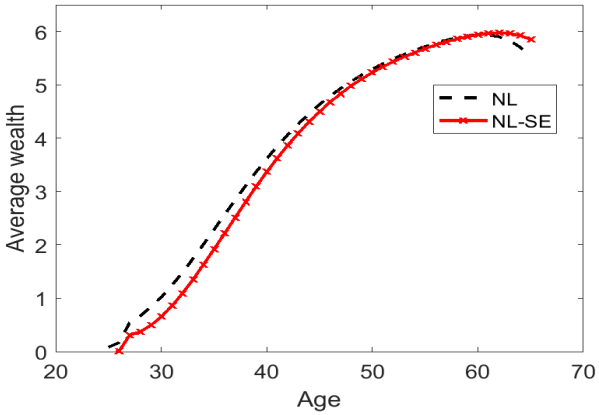


Figure 7: Wealth Accumulation

Income Process	ψ
NL	0.780
NL-SE	0.779
NL-SE / NL	0.653

Table 5: Consumption Insurance

a higher cross section variance of wealth in the early part of the life cycle.

Second, we focus on consumption insurance. We follow De Nardi et al. (2020) and calculate the insurance coefficient of household i's consumption (c_{it}) with respect to income shocks (μ_{it}), as proposed by Blundell, Pistaferri, and Preston (2008), which is given by[15]

$$\psi = 1 - \frac{\text{cov}(\Delta c_{it}, \mu_{it})}{\text{var}(\mu_{it})} \tag{21}$$

Table 5 shows the results. We consider three versions of the model: one in which the process NL is assumed; one in which the process NL-SE is used; and one in which households expect the income process to follow NL-SE, consistent with their subjective expectations, but realized income shocks follow the NL process. This exercise effectively assumes that the difference between the two estimated processes is entirely due to cognitive biases in household expectations. In the next section, we discuss the main alternative explanation, which relies on households possessing superior private information.

We find that the degree of consumption insurance in the first two cases is very similar (around 0.8 in both cases). First, note that in our model income shocks can be either transitory or persistent, while subjective expectations concern total income without distinguishing between the two. Accordingly, the implied degree of consumption insurance lies between the benchmark values typically associated with transitory shocks (around 0.4) and persistent shocks (around 0.9) (see De Nardi et al. (2020)). Second, although the overall level of con-

[15]This measure can be obtained if the shock μ_{it} is known, which is the case in our model economy. An equivalent measure under AR(1) shocks can be obtained by $\tilde{\psi} = 1 - \frac{\text{cov}(\Delta c_{it}, y_{it+1} - y_{it-2})}{\text{cov}(\Delta y_{it}, y_{it+1} - y_{it-2})}$. Results are largely unchanged using this alternative measure.

sumption insurance is similar across the two cases, the underlying mechanisms differ. In the first case, households rely on asset accumulation to smooth relatively large income shocks. In the second, income variance is much lower, so consumption fluctuations are correspondingly limited. Finally, if households save in a manner consistent with their subjective expectations—anticipating low income risk—but actual shocks follow the higher-variance, nonlinear process estimated from realized data, consumption insurance becomes more limited and income and consumption move more closely together. Thus, if cognitive biases distort households' assessment of extreme income risk and influence their saving decisions, consumption self-insurance may be impaired.

7 Assessing the information content in subjective expectations

In the previous section, we showed that if households' income process is consistent with their subjective expectations, perceived income risk is low, and households are able to insure against risk while accumulating wealth more gradually over their lifetime. In contrast, if the income process is consistent with estimates based on realized data, income risk is substantially higher and nonlinear, and households are able to insure against risk only by accumulating significant wealth early in life.

We then posit that if households expect the income process to involve low risk and plan their asset accumulation accordingly, but realized shocks exhibit the higher variance and nonlinearities estimated from realized data, consumption insurance will be significantly affected. In this case, household consumption becomes more volatile and responds more strongly to income changes. The key assumption underlying this exercise is that households exhibit biases in expectations formation. This may be related to the fact that large income shocks are not sufficiently salient, leading households to underestimate the probability that such shocks occur (see Bordalo et al. (2013)).

Nonetheless, the difference between the estimated processes could be related to the fact that households have superior private information regarding the distribution of future income shocks that is not available to the econometrician who relies solely on realized data. When

using realized data and observable household characteristics, the econometrician may overestimate income risk for a subset of households. In effect, observed income variation is used to infer risk for households that, in fact, have a very low probability of experiencing such variation—precisely the information conveyed by their subjective expectation responses. Do we have evidence that households' subjective expectations contain additional information?

We start by verifying whether adding subjective expectations data to an otherwise standard autoregressive model with controls for household characteristics improves the prediction of income (Table 6). More specifically, we add the mean and variance of the subjective distribution of future income and find that, although the mean is not significant (in many cases the mean is close to the current income, not adding additional information), the dispersion is both economically and statistically significant.

Table 6: Predictability of household income

VARIABLES	Future household income
Expected household income	0.047
	(0.067)
Household income	0.641***
	(0.068)
Expected dispersion of household income	-0.581***
	(0.210)
Constant	3.283***
	(0.216)
Observations	5,303
R-squared	0.608

Note: This table presents regression results that aim to predict household income using estimates of the mean and the dispersion of future household income, controlling for current income. The regression controls for demographic characteristics and wave dummies. Standard errors are clustered by household. Significance levels - *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Second, we verify whether subjective income expectations are related to household consumption decisions by examining consumption and elicited MPCs.[16] Once again, when current

[16]Elicited MPCs are measured based on a standard lottery question included in the EFF: *Imagine that in a raffle, you won an amount of money equal to your household's monthly income. What percentage would you spend in the following twelve months, rather than saving it?* Households' answers range from 0 (saving all) to 100 (spending all).

income is included, expected income does not provide additional explanatory power, whereas the dispersion of future income is both economically and statistically significant, negatively affecting both consumption and elicited MPCs.

Table 7: Subjective income expectations and household consumption

VARIABLES	(1) Log consumption	(2) Elicited MPC
Expected household income (in logs)	-0.035	0.067
	(0.038)	(0.062)
Household income (in logs)	0.451***	-0.037
	(0.042)	(0.064)
Expected dispersion of household income	-0.293**	-0.342*
	(0.130)	(0.192)
Household wealth (in logs)	0.034***	-0.013**
	(0.006)	(0.007)
Constant	4.675***	0.015
	(0.144)	(0.180)
Observations	8,822	7,221
R-squared	0.460	

Note: The first column is an OLS regression of log household consumption on the income expectations measures, wealth, income, demographic characteristics and time dummies. The second column is a Tobit regression of the elicited MPC measure from the Spanish EFF on the income expectations measures, wealth, income, demographic characteristics, and time dummies. Standard errors are clustered by household. Significance levels - *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

8 Conclusions

This paper develops and estimates flexible income processes using probabilistic subjective expectations data and contrasts them with processes inferred from realized income alone. Across both the state-dependent AR(1) and nonlinear quantile specifications, we document systematic and economically meaningful differences between perceived and realized income dynamics. Households perceive income to be substantially more persistent and less dispersed than implied by ex-post realizations; the nonlinearities present in realized income data are not present in subjective expectation data. Embedding these alternative income processes into a life-cycle model reveals that such differences have first-order implications for precautionary saving, wealth accumulation, and the degree of consumption insurance.

An important open question is whether these discrepancies reflect biases in expectation formation or instead arise because households possess private, household-specific information that is not observed by the econometrician estimating income processes from realized data alone. Distinguishing between these two interpretations is crucial for both measurement and policy analysis: if the gap is driven by cognitive distortions, models calibrated on subjective expectations may misrepresent actual risk; if it reflects superior private information, models based solely on realized income may overstate uncertainty relevant for household decision-making. As such, developing a framework that disentangles biased belief formation from informational advantages is a promising path of future research.

References

Almuzara, M. (2020). Heterogeneity in transitory income risk. Technical report, Federal Reserve Bank of New York.

Alvarez, J. and M. Arellano (2022). Robust likelihood estimation of dynamic panel data models. *Journal of Econometrics 226*(1), 21–61.

Arellano, M., O. Attanasio, B. Augsburg, and S. Crossman (2022). Estimating flexible income processes from subjective expectations data: evidence from colombia and india. Technical report, CEMFI.

Arellano, M., R. Blundell, and S. Bonhomme (2017). Earnings and consumption dynamics: a nonlinear panel data framework. *Econometrica 85*(3), 693–734.

Arellano, M. and S. Bonhomme (2016). Nonlinear panel data estimation via quantile regressions. *The Econometrics Journal 3*(19), C61–C94.

Attanasio, O. and B. Augsburg (2016). Subjective expectations and income processes in rural india. *Economica 83*(331), 416–442.

Attanasio, O., A. Kovacs, and K. Molnar (2020). Euler equations, subjective expectations and income shocks. *Economica 87*(346), 406–441.

Blundell, R., L. Pistaferri, and I. Preston (2008). Consumption inequality and partial insurance. *American Economic Review 98*(5), 1887–1921.

Blundell, R., L. Pistaferri, and I. Saporta-Eksten (2016). Consumption inequality and family labor supply. *American Economic Review 106*(2), 387–435.

Bonhomme, S. and A. Denis (2025). Fixed effects and beyond. bias reduction, groups, shrinkage and factors in panel data. Technical report, Banco de España.

Bordalo, P., N. Gennaioli, and A. Shleifer (2013). Salience and consumer choice. *Journal of Political Economy 121*(5), 803–843.

Bordalo, P., N. Gennaioli, and A. Shleifer (2018). Diagnostic expectations and credit cycles. *The Journal of Finance 73*(1), 199–227.

Bover, O. (2015). Measuring expectations from household surveys: new results on subjective probabilities of future house prices. *SERIEs 6*(4), 361–405.

Browning, M., M. Ejrnaes, and J. Alvarez (2010). Modelling income processes with lots of heterogeneity. *The Review of Economic Studies 77*(4), 1353–1381.

Caplin, A., V. Gregory, E. Lee, S. Leth-Petersen, and J. Sæverud (2023). Subjective earnings risk. Technical report, National Bureau of Economic Research.

Chamberlain, G. and K. Hirano (1999). Predictive distributions based on longitudinal earnings data. *Annales d'Economie et de Statistique*, 211–242.

de Bresser, J. and A. van Soest (2013). Survey response in probabilistic questions and its impact on inference. *Journal of Economic Behavior & Organization 96*, 65–84.

De Nardi, M., G. Fella, and G. Paz-Pardo (2020). Nonlinear household earnings dynamics, self-insurance, and welfare. *Journal of the European Economic Association 18*(2), 890–926.

Gálvez, J. and G. Paz-Pardo (2026). Richer earnings dynamics, consumption and portfolio choice over the life cycle. *Journal of Financial Economics 176*, 104206.

García-Miralles, E., N. Guner, and R. Ramos (2019, November). The spanish personal income tax: facts and parametric estimates. *SERIEs: Journal of the Spanish Economic Association 10*(3), 439–477.

Giustinelli, P., C. F. Manski, and F. Molinari (2022a). Precise or imprecise probabilities? evidence from survey response related to late-onset dementia. *Journal of the European Economic Association 20*(1), 187–221.

Giustinelli, P., C. F. Manski, and F. Molinari (2022b). Tail and center rounding of probabilistic expectations in the health and retirement study. *Journal of Econometrics 231*(1), 265–281.

Gu, J. and R. Koenker (2017). Unobserved heterogeneity in income dynamics: An empirical bayes perspective. *Journal of Business & Economic Statistics 35*(1), 1–16.

Guiso, L., T. Jappelli, and L. Pistaferri (2002). An empirical analysis of earnings and employment risk. *Journal of Business & Economic Statistics 20*(2), 241–253.

Guner, N., E. Kaya, and V. Sánchez-Marcos (2024, August). Labor market institutions and fertility. *International Economic Review 65*(3), 1551–1587.

Guvenen, F., F. Karahan, S. Ozkan, and J. Song (2021). What do data on millions of us workers reveal about lifecycle earnings dynamics? *Econometrica 89*(5), 2303–2339.

Heiss, F., M. Hurd, M. van Rooij, T. Rossmann, and J. Winter (2022). Dynamics and heterogeneity of subjective stock market expectations. *Journal of Econometrics 231*(1), 213–231.

Hospido, L. (2012). Modelling heterogeneity and dynamics in the volatility of individual wages. *Journal of Applied Econometrics 27*(3), 386–414.

Jappelli, T. and L. Pistaferri (2010). The consumption response to income changes. *Annu. Rev. Econ. 2*(1), 479–506.

Kaplan, G. and G. L. Violante (2010). How much consumption insurance beyond self-insurance? *American Economic Journal: Macroeconomics 2*(4), 53–87.

Kleinjans, K. J. and A. van Soest (2014). Rounding, focal point answers and nonresponse to subjective probability questions. *Journal of Applied Econometrics 29*(4), 567–585.

Kovacs, A., C. Rondinelli, and S. Trucchi (2021). Permanent versus transitory income shocks over the business cycle. *European Economic Review 139*, 103873.

Lee, W. (2025). Identification and estimation of dynamic random coefficient models. *arXiv preprint arXiv:2505.01600*.

Manski, C. F. and F. Molinari (2010). Rounding probabilistic expectations in surveys. *Journal of Business & Economic Statistics 28*(2), 219–231.

Meghir, C. and L. Pistaferri (2004). Income variance dynamics and heterogeneity. *Econometrica 72*(1), 1–32.

Rozsypal, F. and K. Schlafmann (2023). Overpersistence bias in individual income expectations and its aggregate implications. *American Economic Journal: Macroeconomics 15*(4), 331–371.

Van der Klaauw, W., W. Bruine de Bruin, G. Topa, S. Potter, and M. F. Bryan (2008). Rethinking the measurement of household inflation expectations: preliminary findings. *FRB of New York Staff Report* (359).

Wang, T. (2023). Perceived versus calibrated income risks in heterogeneous-agent consumption models. Technical report, Bank of Canada.

Online Appendix for the paper "Income uncertainty, nonlinear dynamics and consumption: a subjective income expectations framework"

A Data description and summary statistics

The EFF is a household survey that has been administered by the Banco de España since 2002. The primary goal of the EFF is to track household consumption, income and wealth. The survey has a rotating panel component, wherein a fraction of households are followed over time up to four waves. It is this component that we use for our empirical analysis.

Starting from 2011, the EFF has administered questions that elicit household expectations on diverse questions such as future house prices, household income, unemployment, and the marginal propensity to consume. As the household income expectations question was first asked in 2014, we focus our empirical study from the 2014 wave onwards. To construct the panel dataset that we use for the empirical study, we follow the criterion in Blundell, Pistaferri, and Saporta-Eksten (2016). We focus on households whose heads are aged 25-65 years old. We impose that there should be at least one member of the household active in the labor market. Moreover, we remove households whose main source of income is pensions. We also impose that we observe households for at least two waves of the EFF. This leaves us with a panel of 8,823 household-year observations for the analysis.

Variable definitions Apart from the subjective income expectations question, we also consider other variables for the empirical analysis. Realized income is the total pre-tax household income, including transfers. In the estimation of the realized income process, we use residuals from a first stage regression of the logarithm of total household income on a set of demographics, which include age of the household head, education categories, family composition (family size, number of children, dummy for children away), indicators for income recipients other than the household head (one additional income recipient or more than one) and year dummies. Household (net) wealth is constructed as the sum of total real and financial assets (including real estate value, car value, cash, pension funds, stocks and bonds) net of total debt (mortgages and other debt). Household consumption is the sum of expenditure in food and

other non-durable goods. We also use the subjective house price expectations question that was described in Bover (2015). Like the income expectations question, the house price expectations questions aims to elicit households' views on the probability distribution of their growth of the value of their house (regardless of whether they own it or are renting the house).

Summary statistics The following table describes the summary statistics of the households in the sample we are using for the estimation.

Table A.1: Summary statistics, EFF 2014-2022 panel.

Variable	Mean	Standard Deviation	Minimum	Maximum
Household income	45355.34	45645.72	104.46	7885194
Household wealth	264038	1161995	1.0111	2.54e+08
Non-durable consumption	15089.32	9472.724	1007.056	662566
Age of head	48.2669	9.298195	25	65
Head is woman	.4090239	.4916816	0	1
Number of household members	3.108887	1.300956	1	9
Second adult income earner	.5275598	.4992682	0	1
Third (or more) adult income earner	.1197938	.3247387	0	1

Notes: All monetary amounts are in 2017 euros. These statistics were calculated for the sample who satisfy the sample selection criteria.

B Estimation of the realized income processes

In this section, we outline the estimation of the income processes we described in section 3 using both realized income data and subjective income data.

B.1 State-dependent AR(1) process

Recall the specification of the state-dependent AR(1) process:

$$y_{it+1} = \rho(\mathbf{X}_{it})\, y_{it} + \mathbf{X}_{it}'\beta + \eta_i + \sigma(\mathbf{X}_{it})\, u_{it+1}, \qquad u_{it+1} \sim N(0,1), \tag{B1}$$

To estimate this model, we follow the approach in Alvarez and Arellano (2022) and consider a random effects specification for η_i. In particular, we assume that

$$\eta_i \sim N(0, \sigma^2), \quad \eta_i \perp u_{it} \text{ for all } t, \quad \eta_i \perp \mathbf{X}_{it} \text{ for all } t. \tag{B2}$$

Let us define $\mu_{it}(\theta, \eta_i) \equiv \rho(\mathbf{X}_{it}; \pi) y_{it} + \mathbf{X}_{it}'\beta + \eta_i$, and $\sigma_{it}(\theta) \equiv \sigma(\mathbf{X}_{it}; \gamma)$. Then, the conditional distribution of y_{it+1} given y_{it}, $\mathbf{X}_{it}$ and η_i is

$$y_{it+1}|y_{it}, \mathbf{X}_{it}, \eta_i \sim N(\mu_{it}(\theta, \eta_i), \sigma_{it}^2(\theta))$$

with density

$$f(y_{it+1}|y_{it}, \mathbf{X}_{it}, \eta_i; \theta) = \frac{1}{\sigma_{it}(\theta)} \phi\left(\frac{y_{it+1} - \mu_{it}(\theta, \eta_i)}{\sigma_{it}(\theta)}\right), \tag{B3}$$

where $\phi(\cdot)$ is the standard Normal pdf. Assuming conditional independence over t given η_i,

$$L_i(\theta|\eta_i) = \prod_{t=1}^{T} f(y_{it+1}|y_{it}, \mathbf{X}_{it}, \eta_i; \theta). \tag{B4}$$

Integrating over the distribution of η_i, we obtain the unconditional likelihood contribution for household i:

$$L_i(\theta) = \int \prod_{t=1}^{T} f(y_{it+1}|y_{it}, \mathbf{X}_{it}, \eta_i; \theta) \frac{1}{\sigma_\eta} \phi\left(\frac{\eta_i}{\sigma_\eta^2}\right) d\eta_i \tag{B5}$$

The maximum likelihood estimator of the parameters of the income process can be obtained by maximizing over the average log-likelihood

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta), \text{ where } \ell(\theta) = \sum_{i=1}^{N} \log L_i(\theta). \tag{B6}$$

Because the optimization problem requires integrating over the distribution of η, we evaluate the integral numerically by Gauss-Hermite quadrature methods. In our empirical implementation, we assume that $\rho(\mathbf{X}_{it})$ is a logistic function and that $\sigma(\mathbf{X}_{it})$ is an exponential function.

B.2 Nonlinear income process

Following Arellano et al. (2017), we approximate the conditional quantile function of y_{it+1} given y_{it} and covariates $\mathbf{X}_{it}$ via the following

$$Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) = \sum_{k=0}^{K} a_k(\tau)\, \psi_k(y_{it}) + \mathbf{X}_{it}'\beta(\tau), \tag{B7}$$

where $a_k(\tau)$ and $\beta(\tau)$'s are piecewise-interpolating polynomial splines, and ψ_k is a low-order Hermite polynomial. Given this specification, we solve, for a given set of quantiles the following optimization problem:

$$(\widehat{\alpha}(\tau_l), \widehat{\beta}(\tau_l)) = \arg \min_{a_k(\tau), \beta(\tau)} \rho_{\tau_l} \left(y_{it+1} - \sum_{k=0}^{K} a_k(\tau)\, \psi_k(y_{it}) - \mathbf{X}_{it}'\beta(\tau) \right).$$

C Estimating the nonlinear income process using subjective expectations data

C.1 Approximation of the conditional distributions

In this subsection, we write the analytical form of the cdf and the pdf.

Following Arellano and Bonhomme (2016), notice that we can write the approximating cdf of the income process. To simplify the discussion, we will ignore the conditioning on state variables $\mathbf{X}_{it}$. Recall that we can write the conditional quantile function of y_{it+1} given y_{it} as

$$Q_y(\tau \mid y_{it}) = \sum_{k=0}^{K} a_k(\tau)\psi_k(y_{it}), \tag{C8}$$

where $a_k(\tau)$ are piecewise-linear interpolating splines in τ. Given this specification, we can write the approximating cdf $F(y_{it+1}|y_{it})$ and the pdf $f(y_{it+1}|y_{it})$. In particular, we can distinguish three cases. To make the notation compact, let us define $A_y(l) = Q_y(\tau_l \mid y_{it})$ for a given quantile τ_l.

- For $y < A_y(1) : F(y_{it+1}|y_{it}) = \tau_1 \exp\{\lambda_1(y - A_y(1))\}$. This implies that

$$f(y|y_{it}) = \lambda_1\tau_1 \exp\{\lambda_1(y - A_y(1))\}.$$

- For the interior segments:

$$F(y|y_{it}) = \tau_l + \frac{\tau_{l+1} - \tau_l}{A_y(l+1) - A_y(l)}(y - A_y(l)).$$

  This implies that the pdf is:

$$f(y|y_{it}) = \frac{\tau_{l+1} - \tau_l}{A_y(l+1) - A_y(l)}.$$

- For $y > A(\tau_L) : F(y|y_{it}) = 1 - (1 - \tau_L) \exp\{-\lambda_L(y - A_y(L))\}$. This implies that

$$f(y|y_{it}) = \lambda_L(1 - \tau_L) \exp\{-\lambda_L(y - A_y(L))\}.$$

Because we can write the conditional distributions analytically, it becomes easier to sample latent y_{it+1} values in the stochastic E-step of the estimation algorithm.

C.2 Coarse information

E-step. For each (i,t) and for each bin j, draw d_{jit} latent values from the posterior density $f(y|y_{it}, \mathbf{X}_{it}, y \in I_{jit}; \theta^{(s)})$. In practice, for each point in bin j, the steps to draw the latent values are:

1. Compute $a_{jt} = F_{it}(s_{j-1,it}; \theta^{(s)})$ and $b_{jt} = F_{it}(s_{j,it}; \theta^{(s)})$, where $F_{it}(\cdot)$ is the approximating cdf in Appendix C.1.
2. Draw $u \sim U[a_{jt}, b_{jt}]$.
3. Set $y_{it+1}^{(s,m)} = Q_y(u|y_{it}, \mathbf{X}_{it}; \theta^{(s)})$.

We finally collect the pooled augmented sample $\{y_{i,t+1}^{(s,m)}, y_{it}, \mathbf{X}_{it}\}_{i,t,m}$ with exactly M draws per (i,t).

M-step. In the M-step, we update both the parameters of the quantile function and the tails. To update the parameters of the quantile function, we estimate quantile regressions at the spline knots $\{\tau_\ell\}_{\ell=1}^{L}$ on the augmented sample. For each knot τ_ℓ, estimate

$$(\widehat{a}(\tau_\ell), \widehat{\beta}(\tau_\ell)) = \arg\min_{a(\tau_\ell),\beta(\tau_\ell)} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \rho_{\tau_\ell}\left(y_{i,t+1}^{(s,m)} - \sum_{k=0}^{K} a_k(\tau_\ell)\psi_k(y_{it}) - \mathbf{X}_{it}'\beta(\tau_\ell)\right),$$

where $\rho_\tau(\cdot)$ is the standard check function.

We update the tail parameters with imposed exponential tail restrictions. Let $A_{it}(1) = Q_y(\tau_1 \mid y_{it}, \mathbf{X}_{it}; \theta^{(s+1)})$ and $A_{it}(L) = Q_y(\tau_L \mid y_{it}, \mathbf{X}_{it}; \theta^{(s+1)})$ denote the updated boundary quantiles. We index pooled simulated draws by $l \equiv (i,t,m)$ and write $y_l \equiv y_{i,t+1}^{(s,m)}$, with $A_{y_l}(1) \equiv A_{it}(1)$ and $A_{y_l}(L) \equiv A_{it}(L)$ for the corresponding (i,t). Define

$$N_L = \#\{l : y_l < A_{y_l}(1)\}, \qquad N_R = \#\{l : y_l > A_{y_l}(L)\}.$$

Then the MLE updates for the exponential tail rates are

$$\widehat{\lambda}_1 = \left[\frac{1}{N_L}\sum_{l:\, y_l < A_{y_l}(1)} \big(A_{y_l}(1) - y_l\big)\right]^{-1}, \widehat{\lambda}_L = \left[\frac{1}{N_R}\sum_{l:\, y_l > A_{y_l}(L)} \big(y_l - A_{y_l}(L)\big)\right]^{-1}.$$

C.3 Hurdle model in the nonlinear process

This appendix extends the hurdle reporting model in the main text to the case where the subjective income process is nonlinear and specified following Arellano et al. (2017). The key difference relative to the baseline is that the nonlinear income process is estimated via a stochastic EM algorithm. Moreover, the only random effect is the respondent-specific effect b_i in the hurdle equation.

C.3.1 Hurdle structure

Let $j^\star$ denote the index of the middle bin and define the observed bunching indicator

$$S_{it} \equiv \mathbf{1}\{d_{j^\star,it} = M\}.$$

Let $R_{it} \in \{0,1\}$ indicate whether respondent i at time t is in the deterministic focal-point regime. We model the focal-point propensity as

$$\pi_{it}(b_i) \equiv \Pr(R_{it} = 1 \mid \mathbf{Z}_{it}, b_i) = \Lambda(\gamma_0 + \mathbf{Z}_{it}'\gamma_1 + b_i)\,, \tag{C9}$$

where b_i is a time-invariant respondent-specific random effect, with

$$b_i \sim N(0, \sigma_b^2).$$

Conditional on b_i, define the probability of placing all M points in the middle bin under the (non-deterministic) multinomial mechanism as

$$A_{it} \equiv \Pr\left(d_{j^\star,it} = M \mid \mathcal{I}_{it}, R_{it} = 0\right) = \left(p^*_{j^\star,it}\right)^M,$$

where $p^*_{j^\star,it}$ is implied by the nonlinear quantile process below. Then the per-period likelihood contribution is

$$L_{it}(b_i) = \begin{cases} \pi_{it}(b_i) + \big(1 - \pi_{it}(b_i)\big)\, A_{it}, & \text{if } S_{it} = 1, \\ \big(1 - \pi_{it}(b_i)\big)\, \dfrac{M!}{\prod_{j=1}^J d_{jit}!} \displaystyle\prod_{j=1}^J \left(p^*_{j,it}\right)^{d_{jit}}, & \text{if } S_{it} = 0, \end{cases} \tag{C10}$$

and the individual integrated likelihood is one-dimensional:

$$L_i^f(\theta) = \int \prod_{t=1}^T L_{it}(b_i)\, f(b_i)\, db_i, \qquad f(b_i) = \phi(b_i; 0, \sigma_b^2). \tag{C11}$$

C.3.2 Nonlinear subjective income process: ABB-style conditional quantiles

Let y^*_{it+1} denote latent subjective next-period income. The survey bins are defined over subjective income growth $\Delta y^*_{it} \equiv y^*_{it+1} - y_{it}$ via fixed intervals $\{\mathcal{I}_j\}_{j=1}^J$, so that bin j corresponds to

$$\Delta y^*_{it} \in \mathcal{I}_j \quad \Longleftrightarrow \quad y^*_{it+1} \in \mathcal{J}_{j,it} \equiv y_{it} + \mathcal{I}_j = [\underline{c}_{j,it}, \overline{c}_{j,it}).$$

Conditional quantile function. We parameterize the conditional quantiles of y^*_{it+1} given $(y_{it}, \mathbf{X}_{it})$ as

$$Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) = \sum_{k=0}^{K} a_k(\tau)\, \psi_k(y_{it}) + \mathbf{X}'_{it}\beta(\tau), \tag{C12}$$

and enforce monotonicity in τ via rearrangement (ABB).

Latent rank representation. Let $U_{it} \sim \text{Unif}(0,1)$ be a latent rank. Then

$$y^*_{it+1} = Q_y(U_{it} \mid y_{it}, \mathbf{X}_{it}), \qquad \Delta y^*_{it} = y^*_{it+1} - y_{it}. \tag{C13}$$

Implied per-point bin probabilities. Because $Q_y(\cdot \mid y_{it}, \mathbf{X}_{it})$ is (after rearrangement) monotone in τ, define rank cutoffs for each bin endpoint:

$$\begin{aligned} \underline{\tau}_{j,it} &\equiv \sup\Big\{\tau \in (0,1) : Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) \leqslant \underline{c}_{j,it}\Big\}, \\ \overline{\tau}_{j,it} &\equiv \sup\Big\{\tau \in (0,1) : Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) \leqslant \overline{c}_{j,it}\Big\}, \end{aligned} \tag{C14}$$

so that the per-point probability of falling in bin j is

$$p^*_{j,it} = \Pr\big(y^*_{it+1} \in \mathcal{J}_{j,it} \mid y_{it}, \mathbf{X}_{it}\big) = \overline{\tau}_{j,it} - \underline{\tau}_{j,it}. \tag{C15}$$

(Operationally, $\underline{\tau}_{j,it}$ and $\overline{\tau}_{j,it}$ are obtained by numerically inverting $\tau \mapsto Q_y(\tau \mid y_{it}, \mathbf{X}_{it})$ at the endpoints.)

C.3.3 Stochastic EM algorithm

Let $\Theta \equiv (\theta, \gamma, \sigma_b^2)$ denote all parameters, where θ collects the objects parameterizing $\{a_k(\cdot), \beta(\cdot)\}$ in (C12). We iterate $k = 0, 1, 2, \ldots$:

E-step: posterior regime weights and latent ranks (and integration over b_i).

(i.) *Quadrature over the hurdle random effect b_i.* Approximate the one-dimensional integral over b_i using Gauss–Hermite nodes $\{b^{(r)}\}_{r=1}^{R}$ with weights $\{w_b^{(r)}\}_{r=1}^{R}$. Compute normalized posterior node weights

$$\omega_{ir}^{(k)} \propto \left[\prod_{t=1}^{T} L_{it}\Big(b^{(r)}; \widehat{\Theta}^{(k)}\Big)\right] w_b^{(r)}, \qquad \sum_{r=1}^{R} \omega_{ir}^{(k)} = 1. \tag{C16}$$

(ii.) *Posterior probability of the deterministic regime.* For $S_{it} = 0$, the deterministic regime is impossible, hence $\Pr(R_{it} = 1 \mid \cdot) = 0$. For $S_{it} = 1$, conditional on a quadrature node $b^{(r)}$,

$$\Pr(R_{it} = 1 \mid S_{it} = 1, b^{(r)}; \widehat{\Theta}^{(k)}) = \frac{\pi_{it}(b^{(r)}; \widehat{\gamma}^{(k)})}{\pi_{it}(b^{(r)}; \widehat{\gamma}^{(k)}) + \big(1 - \pi_{it}(b^{(r)}; \widehat{\gamma}^{(k)})\big)\, A_{it}(\widehat{\theta}^{(k)})}. \tag{C17}$$

Define the node-averaged posterior

$$\overline{R}_{it}^{(k)} \equiv \Pr(R_{it} = 1 \mid \text{data}; \widehat{\Theta}^{(k)}) = \mathbf{1}\{S_{it} = 1\} \sum_{r=1}^{R} \omega_{ir}^{(k)} \Pr(R_{it} = 1 \mid S_{it} = 1, b^{(r)}; \widehat{\Theta}^{(k)}). \tag{C18}$$

(iii.) *Simulate latent ranks consistent with the observed bins (only for the non-deterministic component).* Using current $\widehat{\theta}^{(k)}$, compute rank cutoffs $\{\underline{\tau}_{j,it}, \overline{\tau}_{j,it}\}_{j=1}^{J}$ from (C14). For each bin j, simulate d_{jit} latent ranks

$$U_{itjm}^{(k)} \sim \text{Unif}\Big(\underline{\tau}_{j,it},\ \overline{\tau}_{j,it}\Big), \qquad m = 1, \ldots, d_{jit}, \tag{C19}$$

and map them into latent incomes

$$y_{itjm}^{*,(k)} = Q_y\Big(U_{itjm}^{(k)} \mid y_{it}, \mathbf{X}_{it}; \widehat{\theta}^{(k)}\Big). \tag{C20}$$

Each simulated draw receives weight

$$w_{itjm}^{(k)} \equiv \big(1 - \overline{R}_{it}^{(k)}\big), \tag{C21}$$

so that only the non-deterministic component updates the nonlinear income process.[17]

[17] If desired, one can also propagate the uncertainty in b_i into this step by multiplying by $\sum_r \omega_{ir}^{(k)}$, which equals one.

M-step: update γ, σ_b^2, and θ.

(i.) *Gate parameters* γ. Update (γ, σ_b^2) by maximizing the expected gate log-likelihood, integrating over b_i by quadrature:

$$(\widehat{\gamma}^{(k+1)}, \widehat{\sigma}_b^{2,(k+1)}) = \arg\max_{\gamma,\sigma_b^2} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{r=1}^{R} \omega_{ir}^{(k)} \Big[\overline{R}_{it}^{(k)} \log \pi_{it}(b^{(r)};\gamma) + \big(1-\overline{R}_{it}^{(k)}\big) \log\big(1-\pi_{it}(b^{(r)};\gamma)\big)\Big]. \tag{C22}$$

(ii.) *Quantile-process parameters* θ. Let $\mathbf{z}_{it} \equiv \big(\psi_0(y_{it}), \ldots, \psi_K(y_{it}), \mathbf{X}_{it}'\big)'$ and write

$$Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) = \mathbf{z}_{it}'\delta(\tau), \quad \delta(\tau) \equiv \big(a_0(\tau), \ldots, a_K(\tau), \beta(\tau)'\big)'.$$

For a grid of quantile indices $\tau \in \mathcal{T}$, update $\delta(\tau)$ pointwise by weighted quantile regression using the simulated latent incomes:

$$\widehat{\delta}^{(k+1)}(\tau) = \arg\min_{\delta} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{m=1}^{d_{jit}} w_{itjm}^{(k)}\, \rho_\tau\Big(y_{itjm}^{*,(k)} - \mathbf{z}_{it}'\delta\Big), \tag{C23}$$

where $\rho_\tau(u) = u(\tau - \mathbf{1}\{u < 0\})$ is the check loss. Finally, enforce monotonicity of $\tau \mapsto Q_y(\tau \mid y_{it}, \mathbf{X}_{it})$ via rearrangement (ABB step).

Convergence. At each iteration, evaluate the observed-data log-likelihood

$$\ell^f(\widehat{\Theta}^{(k)}) = \sum_{i=1}^{N} \log L_i^f(\widehat{\Theta}^{(k)})$$

(using Gauss–Hermite quadrature for (C11)) and iterate until convergence.

Remark. The hurdle likelihood (C10) is identical to the main text; the EM algorithm is used to handle the nonlinear ABB-style income process by augmenting each histogram with latent ranks consistent with the observed bins, while the only random effect is the respondent-specific b_i in the hurdle (logit) equation.

D AR(1) process with lots of heterogeneity

In this appendix, we consider the following income process:

$$y_{it+1} = \rho_i y_{it} + \mathbf{X}_{it}\beta + \sigma_i u_{it+1}, u_{it+1} \sim N(0,1). \tag{D24}$$

where σ is household-specific. Estimating individual-specific parameters is not feasible, as we have a short panel (see Bonhomme and Denis (2025) for a discussion of issues related to the estimation of σ_i via fixed effects methods, and Lee (2025) for a discussion of identification and estimation of random coefficient models of this type via fixed effects.) However, we can estimate features of the distribution of σ_i, such as the mean and the variance of the parameters. To do so, we take a random effects approach, and assume parametric distributions for σ_i. We discuss the estimation for both the realized and subjective income processes.

D.1 Realized income process

Define $e_{it}(\rho, \beta) = y_{it+1} - \rho y_{it} - X_{it}'\beta$ for a given household. Note, that for a given individual i at a given time t,

$$e_{it}|\sigma_i \sim N(0, \sigma_i^2), \text{i.i.d. over } t.$$

The conditional density for an individual i at period t is:

$$f(e_{it}|\sigma_i; \beta, \rho) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{e_{it}^2}{\sigma_i^2}\right).$$

The marginal likelihood for an individual i is

$$L_i(\theta) = \int_0^\infty \prod_{t=1}^T f(e_{it}|\sigma_i; \beta, \rho) g(\sigma_i; \delta) d\sigma. \tag{D25}$$

The sample log-likelihood is

$$\ell(\theta) = \sum_{i=1}^N \log L_i(\theta). \tag{D26}$$

To estimate this model, we appeal to maximum likelihood methods. We estimate this model assuming that the distribution of σ_i is lognormal[18] and integrate the log-likelihood via Gauss-Hermite quadrature methods.

D.2 Subjective income process

We discuss the estimation for the baseline model, as the estimation of the model with the hurdle behavior is an extension of the model in the main text. Under the assumption that

[18] We also estimate the model under the assumption that the distribution is inverse-Gamma, which is convenient because we can derive the analytical formulas for the maximum likelihood estimators of the parameter estimates.

households answer the income expectations question with income process (D24) in mind, the true probabilities $p^*_{jit}(\theta)$ are calculated by the households as:

$$p^*_{ijt}(\sigma_i) = \Phi\left(\frac{s_{(j+1),it} - \rho y_{it} - \mathbf{X}'_{it}\beta}{\sigma_i}\right) - \Phi\left(\frac{s_{jit} - \rho y_{it} - \mathbf{X}'_{it}\beta}{\sigma_i}\right). \tag{D27}$$

The integrated likelihood, for a given household i at period t, is:

$$L_{it}(\sigma_i) = \int \prod_{t=1}^{T} \prod_{j=1}^{J} [p^*_{ijt}(\rho, \sigma_i)]^{d_{jit}} f(\sigma_i) d\sigma_i. \tag{D28}$$

In practice, we assume that σ_i is distributed lognormally, and numerically integrate the likelihood via Gauss-Hermite quadrature methods.

D.3 Results

We first discuss the results of the realized and the subjective income processes without taking bunching behavior into account. Table D.1 shows the estimates of persistence and the mean and standard deviation of the distribution of income volatility. As in the state-dependent AR(1) model, we find that households over-estimate the persistence of their income and underestimate the risk that they face as opposed to what is observed in the data. In particular, the persistence parameter is 0.7567 for the realized income process, and 1 for the subjective income process. Meanwhile, on average, income volatility is 0.4229 in the realized income process, while it is 0.0426 in the subjective income process. The top panels of Figure D.1 also demonstrate the differences in dispersion between the implied distribution of σ_i according to realized income data and subjective income data. As the results indicate, the distribution of σ_i under the realized income process is more dispersed than under the subjective income process.

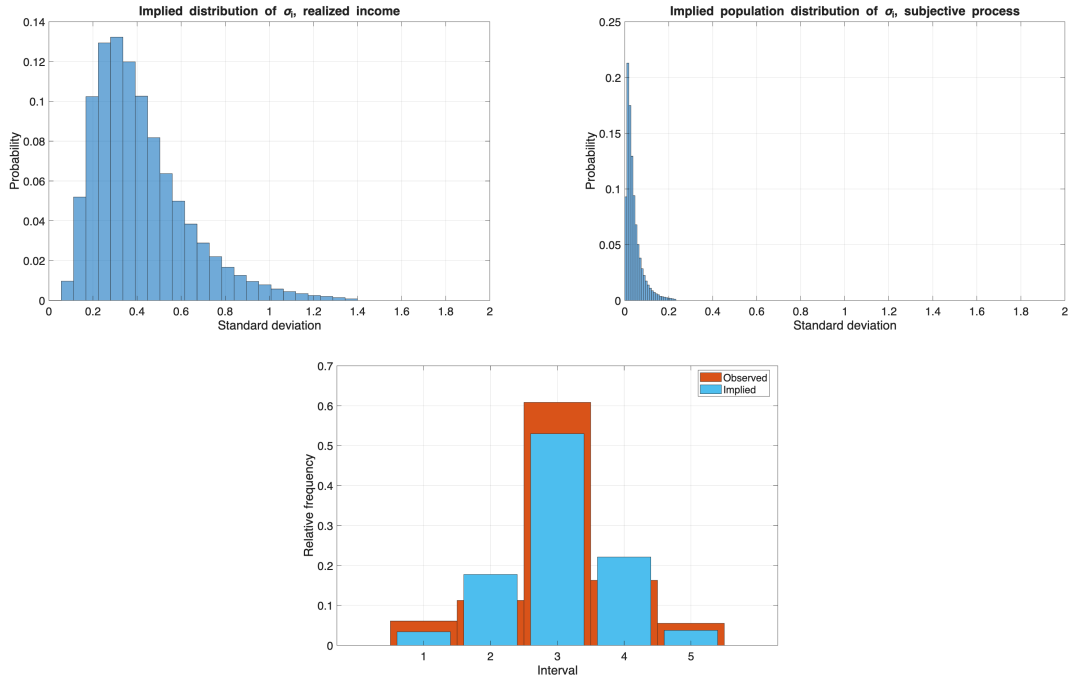
The bottom panel of Figure D.1 shows the model fit of the heterogeneous AR(1) process without a hurdle. We find that the model fit is good, with a RMSE of 0.0522, which is a remarkable improvement over the model with the state-dependent process. We do observe that the model underestimates the bunching in the middle, while it spreads out more points in the left and right bins.

Table D.1: Persistence and distribution of σ_i, heterogeneous AR(1) without hurdle

Panel A: Realized income process		
Persistence	0.7567	-
	[0.7158,0.7663]	-
	Mean	Standard Deviation
Volatility	0.4229	0.2326
	[0.4094, 0.4359]	[0.2126,0.2518]
Panel B: Subjective income process		
Persistence	1.0000	-
	(*)	-
	Mean	Standard Deviation
Volatility	0.0426	0.0471
	[0.0370, 0.0510]	[0.0384, 0.0529]

Notes: This table shows estimates of persistence, and the average and the standard deviation of income volatility, $\mathbb{E}[\sigma_{it}]$ and $\varsigma[\sigma_{it}]$, for the realized and the subjective income processes. We include wave dummies and control for education and employment characteristics. We report in brackets the 95% block bootstrap confidence intervals.

Figure D.1: Distributions of dispersion, and model fit.



Notes: These plots show the distribution of dispersion (top row) for the individual-specific AR(1) process without modeling bunching behavior. The left column corresponds to the realized income process, while the right column corresponds to the subjective income process. The bottom row compares model fit by plotting the implied probabilities of the subjective income process to the observed probabilities.

We then look at how the results change when the subjective income process is estimated with the hurdle model. Table D.2 presents the results associated to the subjective income process with the hurdle model. The results that we obtain indicate that, while persistence decreases from 1.00 to 0.98, and the estimates of the distributional characteristics of dispersion increase, the estimates are still far from the ones implied by realized income data. Figure D.2 plots the implied and observed bin probabilities of the income process with heterogeneous σ_i. We get a better model fit, as the RMSE is 0.0432. We also observe, though, that while we are able to target well the middle bin, it comes at the expense of the left and right bins.

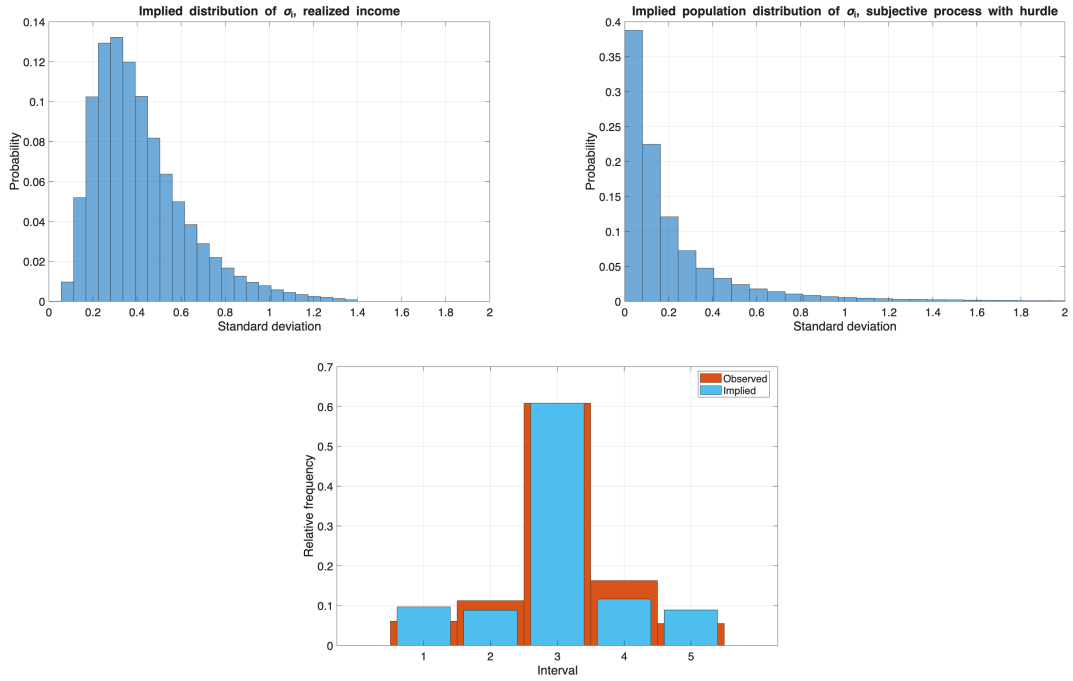
In sum, the results we obtain confirm the insight of the state-dependent AR(1) model; households over-estimate the persistence of their household income, and underestimate the dispersion of income shocks.

Table D.2: Persistence and distribution of σ_i, heterogeneous AR(1) with hurdle

Panel A: Realized income process		
Persistence	0.7567	-
	[0.7158,0.7663]	-
	Mean	Standard Deviation
Volatility	0.4229	0.2326
	[0.4094, 0.4359]	[0.2126,0.2518]
Panel B: Subjective income process		
Persistence	0.9862	-
		-
	Mean	Standard Deviation
Volatility	0.1176	0.1073

Notes: This table shows estimates of persistence, and the average and the standard deviation of income volatility, $\mathbb{E}[\sigma_i]$ and $\varsigma[\sigma_i]$, for the realized and the subjective income processes. We include wave dummies and control for education and employment characteristics. We report in brackets the 95% block bootstrap confidence intervals.

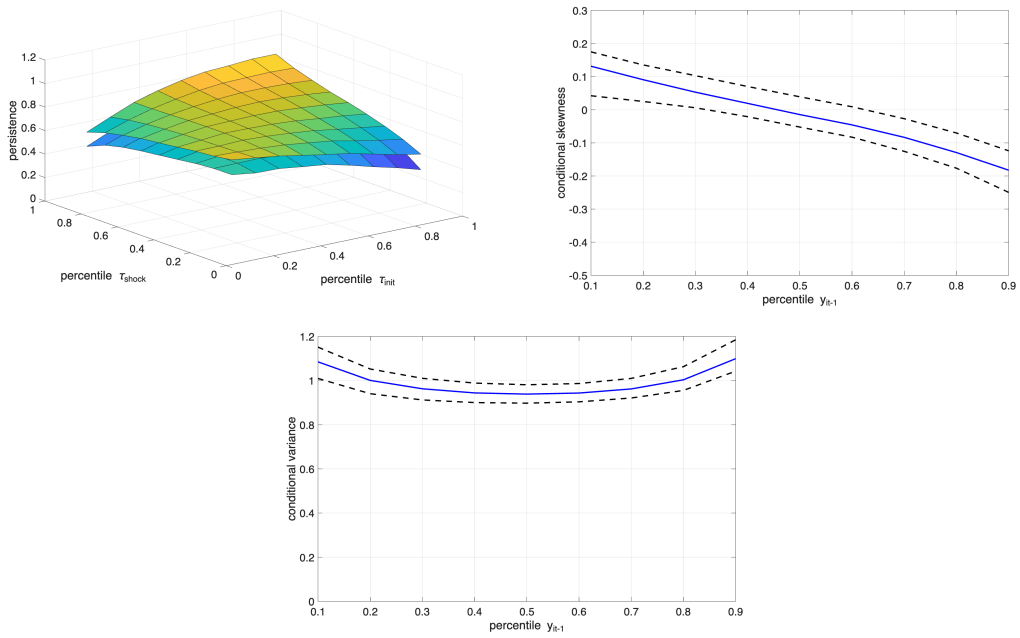
Figure D.2: Distributions of dispersion, and model fit.



Notes: These plots show the distribution of dispersion (top row) for the individual-specific AR(1) process with modeling bunching behavior. The left column corresponds to the realized income process, while the right column corresponds to the subjective income process. The bottom row compares model fit by plotting the implied probabilities of the subjective income process to the observed probabilities.

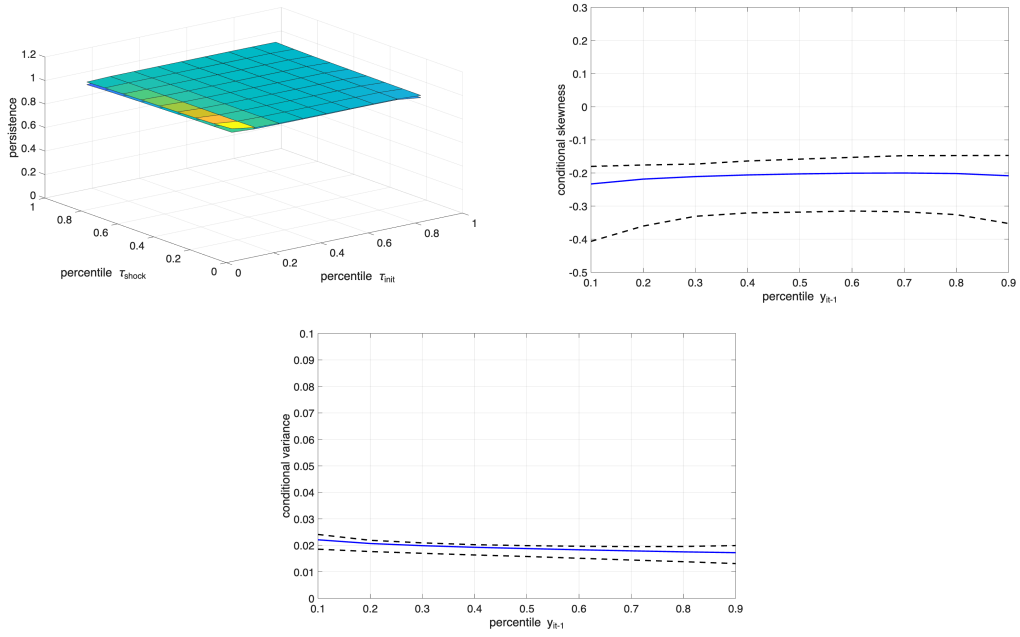
E Additional figures

Figure E.1: Non-parametric bootstraps, nonlinear income process with realized data



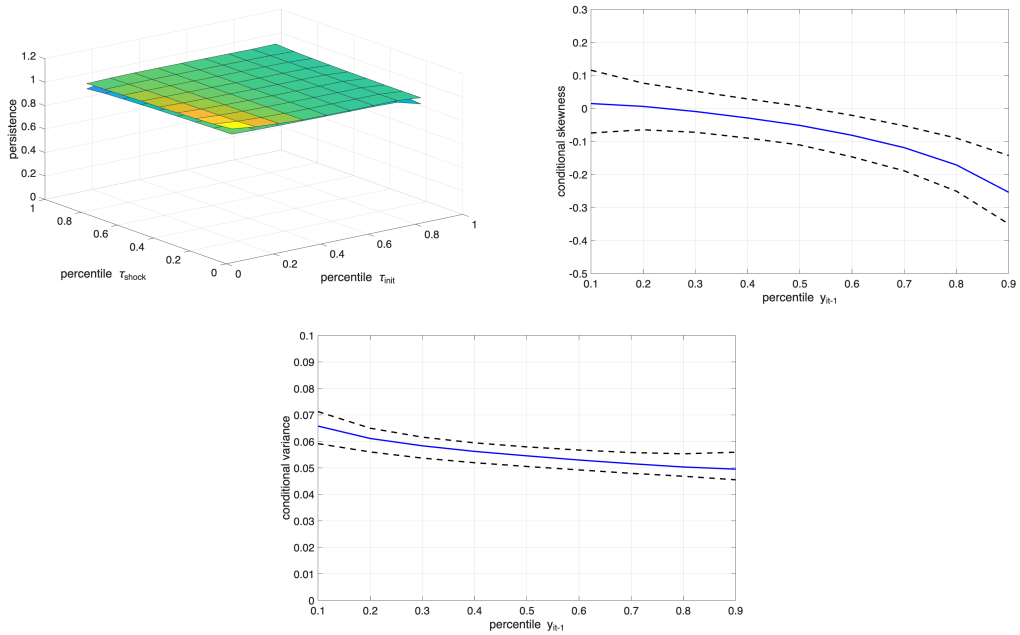
Notes: These plots show the 95% confidence bands associated with the nonlinear persistence measure (top left), conditional skewness (top right), and the conditional volatility (bottom), for the realized income process. These were constructed from 1,000 block bootstrap replications.

Figure E.2: Non-parametric bootstraps, nonlinear income process with subjective income data



Notes: These plots show the 95% confidence bands associated with the nonlinear persistence measure (top left), conditional skewness (top right), and the conditional volatility (bottom), for the subjective income process estimated without the hurdle model. These were constructed from 1,000 block bootstrap replications.

Figure E.3: Non-parametric bootstraps, nonlinear income process with subjective income data with hurdle model



Notes: These plots show the 95% confidence bands associated with the nonlinear persistence measure (top left), conditional skewness (top right), and the conditional volatility (bottom), for the subjective income process estimated with hurdle behavior. These were constructed from 1,000 block bootstrap replications.