What Income Risk Do Households Perceive? Subjective Expectations and Income Dynamics*

Henrique S. Basso
Banco de España and CEMFI

Olympia Bover
CEMFI

Julio Galvez
CUNEF Universidad

Laura Hospido
Banco de España and CEMFI

This draft: July 20, 2026

Abstract

Household consumption and saving decisions depend on perceived income risk, yet macroeconomic models typically 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 relies on subjective expectations 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 dynamics to be substantially more persistent and less dispersed than realized data imply, materially affecting life cycle consumption. Our findings support incorporating subjective probability data into structural models of income dynamics, consumption, and precautionary behavior.

Keywords: Subjective expectations, income risk, income dynamics, nonlinear persistence, consumption, precautionary saving, life cycle, panel data.

JEL codes: D15, D31, D84, E21, C23.

*A previous version of the paper circulated under the title "Subjective Income Expectations and Consumption Dynamics". We are grateful to Manuel Arellano, Richard Blundell, Stéphane Bonhomme, 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 thank Jose Maria Casado for his early contributions to the project and Jiayi Wen for excellent research assistance. We are also grateful to seminar and conference participants at CEMFI, CUNEF Universidad, Erasmus University Rotterdam, IER-Hitotsubashi University, 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), Workshop on consumption and saving over the life cycle (Laugarvatn, Iceland), IAAE (Oslo and Lisbon), the Workshop on Subjective Expectations (Lisbon), the Econometric Society World Congress (Seoul), the CEPR Workshop on Frontiers in Measurement and Survey Methods (Calabria), T2M (Montreal) and the BSE Summer Forum on Expectations in Dynamic Macroeconomic Models 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 the risk an econometrician recovers after the fact. Yet structural models typically discipline this uncertainty using statistical income processes estimated from realized income data, implicitly assuming that the risk inferred from past outcomes coincides with the risk households expect when making decisions. In this paper, we document that this assumption is invalid. Combining subjective income expectations and realized income data for the same households, we recover income processes whose persistence, dispersion, and nonlinearity differ systematically from those estimated from realizations alone. These differences are quantitatively large and have first-order consequences for precautionary saving, the life-cycle wealth accumulation, and the extent to which households insure consumption against income shocks.

Distinguishing between subjective expectations and realized income data is especially pressing in light of recent evidence that income dynamics are highly nonlinear. Realized income processes are found to exhibit 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 have become central ingredients of modern heterogeneous-agent models because they shape consumption insurance, wealth accumulation, and inequality (Kaplan and Violante, 2010; De Nardi et al., 2020; Gálvez and Paz-Pardo, 2026). Yet nearly all existing estimates of nonlinear income risk are inferred from realized income histories. As a result, it remains unclear whether the nonlinear risks recovered by econometricians correspond to the risks households perceive—and act upon—when making consumption and saving decisions.

The core difficulty is that the econometrician does not observe households' information sets (Jappelli and Pistaferri, 2010), nor the process by which households form expectations. Households often possess advance information about future income—such as promotions, bonuses, contract renewals, or impending layoffs—that the researcher cannot observe. At the same time, expectations may be shaped by departures from rational

benchmarks, with households overweighting salient or recent outcomes and underreacting to base rates (Bordalo, Gennaioli, and Shleifer, 2013, 2018). As a result, income innovations identified from realized data may combine true uncertainty with variation that households already anticipate, while perceived risk may reflect systematic biases. From a macroeconomic perspective, these distinctions are crucial: households smooth consumption and save against perceived uncertainty, not against income changes they anticipate or ex post realized risk. Estimating income processes solely from realized outcomes may therefore mismeasure persistence, volatility, and tail risk from the perspective relevant for household 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 the likelihood, dispersion, and asymmetry of future income changes, incorporating both private information and perceived risks. Subjective expectations are therefore closer to the object relevant for consumption, saving, and insurance decisions than income processes inferred solely from realized outcomes. 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 develop a unified framework for estimating income processes by combining probabilistic subjective expectations with realized income data, accommodating nonlinear persistence, non-normal shocks, and heterogeneity in both persistence and volatility. We apply the framework to the Spanish Survey of Household Finances (EFF), in which households distribute probability mass across five pre-specified intervals of next-year income growth. The EFF is particularly well suited to this exercise: it oversamples wealthy households, enabling a richer analysis of the upper tail than conventional surveys; it includes a rotating-panel component; and, crucially, it combines subjective expectations with detailed information on realized income, consumption, and wealth for the same households. Leveraging probabilistic subjective expectations data poses two key chal-

lenges. First, unlike realized income, which can be treated as a continuously distributed outcome observed in each period, subjective expectations are typically reported in a coarse format, with respondents allocating probability mass across a small number of broad income-growth bins. Second, responses may exhibit reporting bias or focal-point behavior, with a substantial fraction of households concentrating all probability mass in a single interval.

We begin by assuming that each household is characterized by a latent income process governing the distribution of future income conditional on current information. We consider two complementary specifications—a state-dependent AR(1) process and a fully nonlinear quantile-based process—to ensure that our conclusions do not hinge on a particular functional form. The AR(1) specification allows persistence and volatility to vary with observable characteristics, extending the standard linear model to capture heterogeneity across households and economic states (Chamberlain and Hirano, 1999; Meghir and Pistaferri, 2004; Browning, Ejrnaes, and Alvarez, 2010; Hospido, 2012; Gu and Koenker, 2017). The quantile-based specification instead models the entire conditional income distribution as a flexible function of current income and covariates, accommodating state-dependent persistence, heteroskedasticity, and asymmetric shocks (Arellano et al., 2017).

For each candidate income process, we derive the probability that future income falls within each survey interval conditional on current income and household characteristics, and treat each response as a draw from a multinomial distribution whose cell probabilities are implied by the underlying process. This maps coarse, interval-censored survey responses directly into a likelihood function, allowing us to estimate the income process by matching observed and model-implied probability allocations. The framework requires only that the estimated process be consistent with households' beliefs; it does not impose rational expectations and therefore accommodates both households that combine observed outcomes with private information and households whose perceptions are systematically biased.

A distinctive feature of the data is the prevalence of focal responses: on average,

more than 40% of households place all of their probability mass on the central interval, effectively reporting that their income will remain approximately stable. Such bunching may reflect genuinely low perceived risk, but it may also arise from cognitive limitations or question framing. To account for the prevalence of focal responses, we extend the framework with a two-regime hurdle model, building on approaches of de Bresser and van Soest (2013), Kleinjans and van Soest (2014), and Heiss, Hurd, van Rooij, Rossmann, and Winter (2022). In this setting, households first decide whether to report deterministically by assigning all probability mass to the central interval and, if not, allocate probability mass according to their perceived distribution. We identify the focal-point regime using an exclusion restriction based on households' responses to a parallel house-price expectations question, which shares a nearly identical probabilistic format and therefore plausibly captures the same reporting style without directly affecting the income distribution. This extension allows us to distinguish between genuinely low perceived risk and focal-point reporting driven by cognitive limitations, thereby improving both model fit and the interpretation of the estimated income processes.

In all specifications, we find that income processes estimated from subjective expectations differ markedly from those inferred from realized income for the same households. The gap between perceived and realized income risk is stark. In a state-dependent AR(1) specification, average perceived income persistence is close to one (0.999), whereas average realized persistence is 0.692. Average perceived volatility, in turn, is between six and ten times smaller than the 0.465 implied by realized data. The same conclusions emerge from the flexible nonlinear quantile model. Perceived income persistence remains high and exhibits little state dependence, whereas realized persistence is considerably lower and highly sensitive to income levels and shock magnitudes. Perceived conditional volatility is likewise at least ten times smaller than realized volatility. The only dimension along which perceived and realized risk broadly align is asymmetry: perceived skewness is somewhat more negative, but statistically indistinguishable from that observed in realized income.

We maintain a parametric and structural interpretation of the data—in contrast to

nonparametric bounds or partial-identification approaches to reporting errors—because our goal is to recover income processes that can be incorporated into models of consumption and saving. This allows us to move from measurement to economic consequences in a way that less parametric approaches do not readily permit. We show that these consequences are first-order. Embedding each estimated process in a standard life-cycle model with incomplete markets, we find that lower perceived risk leads households to save less and accumulate wealth more gradually. Because the high-variance, nonlinear process estimated from realized income generates particularly strong precautionary saving motives among high-income households, it produces substantially greater wealth accumulation early in life and a higher cross-sectional dispersion of wealth than the processes recovered from subjective expectations. The implications for consumption insurance are more subtle. Average insurance is similar under the realized and subjective processes, but this masks pronounced heterogeneity. Under realized risk, wealthier households self-insure against large shocks through asset accumulation, whereas under perceived risk both high- and low-income households hold thinner buffers and respond to shocks more uniformly.

What this means for welfare depends on *why* perceived and realized risk differ, and the evidence points in two opposing directions. If the gap reflects biased beliefs, households underinsure against risks they fail to anticipate. When we allow households to save as though income follows the low-risk perceived process while exposing them to the higher-variance realized process, consumption insurance deteriorates sharply, with the insurance coefficient falling from 0.78 to 0.65. If, instead, the gap reflects private information unavailable to the econometrician, then income processes estimated from realizations overstate the risk households actually face. Subjective expectations are then the more accurate input, and standard estimates of precautionary saving calibrated to realized data are too high. Our predictability exercises suggest that expectations do contain information: the perceived dispersion of income forecasts predicts both realized income and consumption beyond what is captured by current income. Yet the data cannot decisively distinguish between misperception and superior private information. What we establish is that the wedge between perceived and realized income risk is large and

economically meaningful. Quantifying the relative contribution of the forces underlying this wedge, rather than relying on either polar interpretation, is therefore the central task for future research.

Our approach is distinct from the existing literature linking subjective expectations to income risk. A long line of work combines expectations with realized data to study income processes, consumption, and portfolio choice (Guiso, Jappelli, and Terlizzese, 1996; Dominitz and Manski, 1997; Dominitz, 1998, 2001; Pistaferri, 2001; Guiso, Jappelli, and Pistaferri, 2002; Attanasio, Kovacs, and Molnar, 2020; Kovacs, Rondinelli, and Trucchi, 2021), while a more recent literature incorporates subjective expectations into macroeconomic models of perceived income risk (Rozsypal and Schlafmann, 2023; Stoltenberg and Uhlendorff, 2022; Wang, 2023; Caplin, Gregory, Lee, Leth-Petersen, and Sæverud, 2023; Kosar and Melcangi, 2025). This work typically relies on relatively simple, often linear or log-normal, income processes and focuses on low-order moments, whereas we estimate structural income processes that allow us to compare not only persistence and volatility but also nonlinearity.

The paper closest to ours is the contemporaneous study by Arellano, Attanasio, Crossman, and Sancibrián (2024), which also estimates nonlinear income processes from subjective expectations but differs along several dimensions. First, it does not compare perceived and realized processes for the same households. Second, it studies rural households in developing economies, where shocks are driven largely by agricultural and informal-sector events, whereas we study a developed economy with formal labor markets and social insurance. Third, it relies on household-specific min–max thresholds to elicit expectations, whereas our method applies directly to the fixed, common intervals used by the EFF and by surveys such as the New York Fed's Survey of Consumer Expectations. This design introduces two challenges absent from their setting: the coarse, interval-censored nature of reported probabilities and the high prevalence of focal-point responses, whereby a large share of households place all probability mass on a single interval. Our framework addresses both explicitly through a multinomial likelihood that accommodates coarse intervals and a hurdle model that separates genuine low-risk be-

liefs from reporting heuristics. As a result, our approach is applicable to the broad class of large-scale surveys that use fixed common intervals rather than the more specialized elicitation designs that permit household-specific thresholds. More broadly, we connect the subjective-expectations literature to the growing literature on nonlinear income dynamics (Arellano et al., 2017; Guvenen et al., 2021; De Nardi et al., 2020), while drawing on the literature on probabilistic survey responses (Manski and Molinari, 2010; Giustinelli, Manski, and Molinari, 2022a,b; de Bresser and van Soest, 2013; Heiss et al., 2022) to model focal-point behavior. Finally, the EFF panel combines information on expectations with data on realized income, consumption, and wealth for the same households. This allows us to compare perceived and realized income dynamics directly and to study how perceived income risk shapes household behavior, thereby connecting insights from the subjective expectations to both the literature on realized income dynamics and structural models of household behavior.

The rest of the paper is organized as follows. Section 2 describes the data and the probabilistic income expectations measure. Section 3 presents the income process specifications—a state-dependent AR(1) model and nonlinear quantile models—and defines the main risk measures we study. Section 4 details our estimation strategy for subjective income processes, introducing the multinomial likelihood framework and the hurdle model for reporting behavior. Section 5 presents the empirical comparison between income risk estimated from subjective expectations and from realized income data. Section 6 studies the implications of our findings for household consumption and wealth accumulation, while Section 7 examines the informational content of subjective expectations. Finally, Section 8 concludes. Additional material is provided in the online Appendix.

2 Data

We use data from the EFF, a representative survey designed to study household income and wealth in Spain. It contains detailed information on household assets, liabilities, income, consumption, and demographic characteristics. Because the survey was specifically designed to study household wealth, it oversamples wealthy households using information

from wealth tax records. This oversampling allows for a richer analysis of fluctuations in the upper tail of the income distribution than would be possible in a conventional survey. The EFF also includes an important rotating-panel component, with a refreshment sample introduced in each wave to preserve representativeness over time; households can be observed for between two and four waves. Since the question on subjective income expectations is only available starting in 2014, we focus on the 2014, 2017, 2020, and 2022 waves.[1] Following the literature on household income dynamics, we restrict the sample to working-age households (heads aged 25–65) with at least one working adult. Appendix A provides additional details on the variables, sample selection, and summary statistics.

2.1 Eliciting subjective income expectations

Starting in 2014, the EFF included a question designed to elicit households' expectations about future income. Household heads were asked to distribute ten points across five possible scenarios for total household income growth over the following twelve months.

The exact wording of the question is as follows:

> 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

Several features of the income expectations question are worth emphasizing. First, households are asked to allocate probability mass across five pre-specified income-growth

[1] The first six waves of the EFF were conducted every three years between 2002 and 2020. Starting with the 2020 wave, however, the survey frequency changed from triennial to biennial.

intervals. Second, respondents distribute 10 points rather than 100, reducing the cognitive burden associated with probabilistic reporting. Third, the question elicits beliefs using a probability-mass format rather than a cumulative distribution format.[2]

Although probabilistic survey questions may impose additional cognitive demands, a large body of evidence shows that respondents are generally willing and able to answer them meaningfully.[3] In particular, the literature has emphasized that probabilistic expectations tend to be sensible and internally consistent when questions concern well-defined events closely related to respondents' own lives (Van der Klaauw, Bruine de Bruin, Topa, Potter, and Bryan, 2008). Finally, the EFF includes a similar probabilistic expectations question regarding the future value of the household's primary residence, which we later use to help identify reporting behavior and focal responses.

2.2 Response behavior of households

Before discussing the estimation of income processes, we first examine households' response behavior to the subjective income expectations question. Following Bover (2015), we define four indicator variables. *Drop* equals one for households that allocate all 10 points to the first two bins of the income expectations question. *Increase* equals one for households that allocate all 10 points to the last two bins. *Bunching* equals one for households that place all 10 points on the middle bin. Finally, *Spread* equals one for households whose responses do not fall into any of the previous three categories.

Table 1 reports summary statistics on households' response behavior. About 41.7% of households bunch all probability mass on the middle bin across the four EFF waves, although this behavior declines over time—from 57.6% in 2014 to 31.8% in 2022.[4]

We next examine whether these response patterns are correlated with household

[2]The survey is administered through computer-assisted face-to-face interviews, with interviewers providing clarification when needed. Respondents are also given a paper version of the question and response options on which they can draft their answers. An automatic prompt appears whenever the responses entered by the interviewer do not sum to 10, in which case respondents are asked to revise their answers. A similarly designed question on house-price expectations is administered earlier in the survey.

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

[4]We also calculate the persistence of bunching in the middle, and find that persistence is 46.7%. Finally, we also examine the proportion of households bunching at any given bin. We find that 58% of households bunch at any bin, and the persistence of this behavior is approximately 61%.

Table 1: Household response behavior across EFF waves

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

Note: Proportion of households exhibiting the following response patterns in the subjective income expectations question. *Drop* equals one for households that allocate all 10 points to the two lowest income-growth intervals. *Increase* equals one for households that allocate all 10 points to the two highest income-growth intervals. *Spread* equals one for households that distribute probability mass across multiple bins. *Bunching* equals one for households that allocate all 10 points to the middle bin.

demographic characteristics using linear probability models for each indicator (Table 2). Households with secondary education are less likely to bunch and more likely to spread probability mass across bins, while female household heads are more likely to bunch. Bunching also becomes more prevalent with age and among households with permanent labor contracts, but is less common among the self-employed. Most notably, response behavior is highly persistent across domains: households that bunch in the house-price expectations question are about 25 percentage points more likely to bunch in the income expectations question. This strong cross-domain correlation is consistent with the presence of stable reporting styles or cognitive factors affecting probabilistic survey responses.

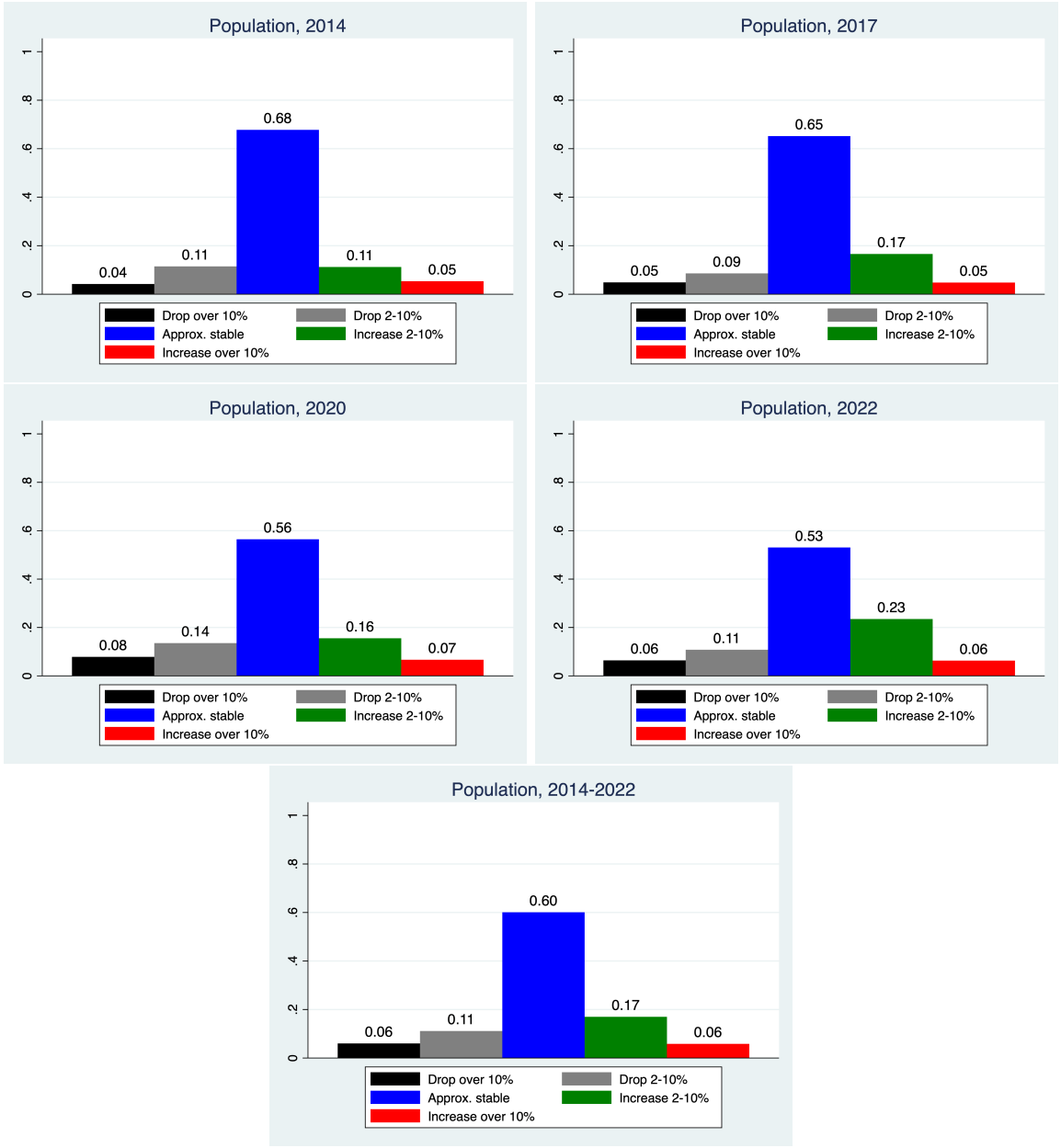
Finally, we plot the average probability distributions implied by households' responses, both across all waves and separately by survey wave (Figure 1). Overall, respondents place most probability mass on the approximately stable income-growth interval, which receives about 60% of the total probability mass on average. However, the concentration of beliefs around income stability declines markedly over time, falling from 68% in 2014 to 53% in 2022. At the same time, households increasingly assign probability mass to positive income-growth scenarios, particularly to moderate income increases. Although the perceived distribution remains centered around stable income growth, the responses display substantial heterogeneity and time variation in perceived risk.

Table 2: Demographic correlates of household response behavior

Variables	(1) Bunching	(2) Drop	(3) Increase	(4) Spread
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 household head	0.066***	-0.003	-0.018*	-0.046***
	(0.016)	(0.010)	(0.010)	(0.016)
Reference person aged 35–45	0.125***	0.005	-0.024	-0.105*
	(0.045)	(0.022)	(0.038)	(0.058)
Reference person aged 45–55	0.147***	0.039*	-0.034	-0.152***
	(0.043)	(0.022)	(0.038)	(0.057)
Reference person aged 55–65	0.232***	0.063**	-0.059	-0.237***
	(0.047)	(0.025)	(0.040)	(0.059)
Permanent labor contract	0.060***	-0.029**	-0.053***	0.021
	(0.023)	(0.013)	(0.016)	(0.022)
Self-employed	-0.072***	0.034**	0.009	0.029
	(0.022)	(0.016)	(0.019)	(0.025)
Risk averse	0.124***	-0.000	-0.009	-0.115**
	(0.043)	(0.023)	(0.029)	(0.056)
Medium risk tolerance	0.072	-0.008	-0.020	-0.044
	(0.045)	(0.025)	(0.031)	(0.059)
Bunching in house-price expectations	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: This table reports estimates from linear probability models relating household response patterns to demographic characteristics. *Drop* equals one for households that allocate all 10 points to the two lowest income-growth intervals. *Increase* equals one for households that allocate all 10 points to the two highest income-growth intervals. *Spread* equals one for households that distribute probability mass across multiple bins. *Bunching* equals one for households that allocate all 10 points to the middle bin. All specifications include wave fixed effects. Robust standard errors are reported 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%.

Figure 1 also indicates the prevalence of focal responses in the EFF, especially the tendency to place all probability mass on the middle. Such bunching may reflect genuinely stable income expectations, but it may also arise from cognitive limitations or reporting bias when answering probabilistic survey questions. The strong correlation between response patterns in income and house-price expectations—after controlling for household characteristics—is consistent with the presence of persistent reporting styles across domains. Because both interpretations are plausible, our empirical framework models the *subjective* income process jointly with a flexible representation of reporting behavior.

3 Modelling income dynamics

Income processes are central to understanding household consumption and saving decisions and are key inputs in quantitative life-cycle models. These models are widely used to study consumption insurance, precautionary saving, portfolio choice, and household welfare (see, e.g., Kaplan and Violante (2010), De Nardi et al. (2020), and Gálvez and Paz-Pardo (2026)). In this context, income processes summarize the uncertainty households face when making intertemporal decisions.

Our objective is to compare income dynamics inferred from realized income data with those perceived by households themselves. To do so, we consider two representations of income dynamics commonly used in quantitative macroeconomic models: a state-dependent autoregressive process and a fully nonlinear income process. We estimate each specification using both *realized* income data and *subjective* expectations data, and then compare the implied characteristics of the income processes, including persistence, dispersion, and asymmetry.

3.1 State-dependent AR(1) income process

In the state-dependent AR(1) specification, log household income for household i at time t, denoted by y_{it}, evolves according to the following law of motion:

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

where $\mathbf{X}_{it}$ denotes observable household characteristics, u_{it+1} is an income innovation, and η_i captures time-invariant unobserved heterogeneity. Both persistence, $\rho(\mathbf{X}_{it})$, and volatility, $\sigma(\mathbf{X}_{it})$, are allowed to vary with household characteristics. This specification therefore permits heterogeneity in income dynamics across households and over time.[5]

In this sense, our specification is closely related to those considered by Chamberlain and Hirano (1999), Meghir and Pistaferri (2004), Browning et al. (2010), Hospido (2012), Gu and Koenker (2017), and Almuzara (2020), which incorporate different dimensions of heterogeneity in income dynamics.[6]

The state-dependent AR(1) specification allows us to characterize the moments of the distributions of persistence and volatility. In particular, we compute the mean and the standard deviation of income persistence,

$$\mu_\rho = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho(\mathbf{X}_{it}), \quad \varsigma_\rho = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\big[\rho(\mathbf{X}_{it}) - \mu_\rho\big]^2} \tag{2}$$

and the mean and the standard deviation of income volatility:

$$\mu_\sigma = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sigma(\mathbf{X}_{it}), \quad \varsigma_\sigma = \sqrt{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\big[\sigma(\mathbf{X}_{it}) - \mu_\sigma\big]^2}. \tag{3}$$

Estimating the parameters of this process using *realized* income data can be carried out using standard likelihood-based methods. Specifically, we adopt a random-effects approach and maximize the average log-likelihood, following Alvarez and Arellano (2022). Additional details on the estimation procedure are provided in Appendix B.

[5] The canonical autoregressive process is a special case of equation (1), obtained when $\rho(\mathbf{X}_{it}) = \rho$ and $\sigma(\mathbf{X}_{it}) = \sigma$.

[6] A more general specification would allow ρ_{it} and σ_{it} to remain unrestricted:

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

Estimating such a model using realized income data requires sufficiently long panel dimensions (Bonhomme and Denis, 2025). Given the relatively short panel dimension of the EFF, we instead parameterize persistence and volatility as functions of observable household characteristics. This approach also allows us to identify the observable drivers of heterogeneity in income dynamics across both the cross-sectional and time-series dimensions.

Nonetheless, we also consider an alternative specification in which persistence is homogeneous while volatility is household-specific:

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

Details and empirical results for this alternative specification are reported in Appendix D.

3.2 Nonlinear income process

Recent work on income dynamics has challenged the long-standing view that household income processes can be adequately represented by linear models. In particular, Arellano et al. (2017) and Guvenen et al. (2021) provide evidence that, contrary to the implications of the standard AR(1) process, pre-tax household income exhibits skewness and volatility that vary with households' position in the income distribution and over the life cycle.

In light of this evidence, Arellano et al. (2017) propose a flexible representation of household income dynamics that accommodates nonlinearity, non-normality, and age dependence in the conditional income distribution. We consider a simplified version of their framework, in which next-period income y_{it+1} follows the quantile process

$$y_{it+1} = Q_y(u_{it+1} \mid y_{it}, \mathbf{X}_{it}; \theta), \qquad u_{it+1} \sim U(0,1), \quad t \geqslant 1. \tag{4}$$

Intuitively, the quantile function maps draws from a uniform distribution on $(0,1)$ (i.e., cumulative probabilities) into draws from the conditional distribution of household income (i.e., quantiles).[7]

One way to understand the role of nonlinearities within this framework is through a generalized notion of persistence, which can be interpreted as a measure of the *persistence of earnings histories*. Specifically, the persistence of income conditional on current income and household characteristics can be written as

$$\rho(y_{it}, \mathbf{X}_{it}, \tau) = \frac{\partial Q_y(\tau \mid y_{it}, \mathbf{X}_{it})}{\partial y_{it}}. \tag{5}$$

This expression measures how sensitive future income at quantile τ is to current income y_{it}. Unlike the linear AR(1) model, persistence is now allowed to vary with both the household's position in the income distribution and the sign and magnitude of shocks. This flexibility allows the model to capture situations in which unusually large positive or negative shocks—such as promotions, job losses, or career transitions—have different degrees of persistence.

Because model (4) places no parametric restrictions on the shape of the conditional income distribution, it also accommodates general forms of conditional heteroskedas-

[7] Both the standard AR(1) and the state-dependent AR(1) specifications can be interpreted as special cases of this more flexible income process.

ticity. In particular, for some $\tau \in (0.5, 1)$, we define a measure of conditional income volatility as

$$\sigma(y_{it}, \mathbf{X}_{it}, \tau) = Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) - Q_y(1-\tau \mid y_{it}, \mathbf{X}_{it}). \tag{6}$$

Similarly, conditional skewness can be measured as

$$sk(y_{it}, \mathbf{X}_{it}, \tau) = \frac{Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) + Q_y(1-\tau \mid y_{it}, \mathbf{X}_{it}) - 2Q_y(0.5 \mid y_{it}, \mathbf{X}_{it})}{Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) - Q_y(1-\tau \mid y_{it}, \mathbf{X}_{it})}. \tag{7}$$

Finally, we use the estimated process to simulate income paths and construct a discrete transition matrix. Specifically, letting $\Xi = \{y_{1,t+1}, \ldots, y_{n,t+1}\}$ denote a grid of income points, we compute transition probabilities $\Pr(y_{i,t+1} \mid y_{j,t})$ for all $i, j = 1, \ldots, n$, which together define the transition matrix $\mathbb{P}$.

Estimating this model using realized income data can be implemented using quantile regressions. Details of the estimation procedure are provided in Appendix B.

4 Estimating subjective income processes

In this section, we describe how to estimate household income processes using subjective income expectations together with observed current income and household characteristics. Unlike realized-income approaches, the data do *not* contain next-period income y_{it+1} at the time expectations are elicited. Instead, we observe current income y_{it}, covariates in the household information set, and respondents' reported probability mass over coarse bins of *future* income growth.

Estimating income processes from subjective expectations data presents two main challenges. First, the survey provides only coarse and discrete information about the conditional distribution of the continuous outcome y_{it+1} given $(y_{it}, \mathbf{X}_{it})$. Second, reported probabilities may reflect elicitation noise, focal responses, or other reporting behaviors potentially associated with cognitive constraints. Standard methods developed for realized income data are therefore not directly applicable to the estimation of *subjective* income processes.

We propose a framework that addresses both issues. In the first subsection, we show how to estimate the state-dependent and nonlinear income processes using subjective expectations data while abstracting from focal responses. To do so, we map the underlying

income process into observed survey probabilities and model responses as multinomial draws. In the second subsection, we extend the framework to account for focal-point responses using finite mixture models. The final subsection briefly discusses identification. Throughout, we interpret y_{it} as log household income, so that income growth is defined as

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

When realized income data are used to estimate income processes, a common implicit assumption is that households are rational and form expectations consistent with the econometrician's inferred income process. By contrast, our framework estimates income dynamics directly from subjective expectations data and therefore does not require rational expectations. Households observe income realizations and form perceptions about the future distribution of income, and we require only that the estimated process be consistent with these perceived beliefs. Our approach is therefore compatible both with fully rational households that use realized data together with private information to forecast future income and with households that form systematically biased perceptions of future income dynamics.

4.1 Coarse information

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

$$-\infty = h_0 < h_1 < \cdots < h_{J-1} < h_J = +\infty,$$

so that bin $j \in \{1, \ldots, J\}$ corresponds to the interval $(h_{j-1}, h_j]$. Households observe current income y_{it} and covariates $\mathbf{X}_{it}$ contained in their information set. We denote the observed income state by

$$\mathcal{W}_{it} \equiv (y_{it}, \mathbf{X}_{it}).$$

We assume that respondents form beliefs about the future distribution of income according to an underlying subjective income process. In what follows, we describe the estimation strategy for both the state-dependent AR(1) specification and the nonlinear quantile-based process.

4.1.1 State-dependent AR(1) process

Suppose households perceive income dynamics according to the state-dependent AR(1) process:

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

Respondents are assumed to form beliefs about the probability that future income falls into each of the pre-specified intervals implied by 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

$$\kappa_{j,it} \equiv y_{it} + h_j,$$

so that the event $\Delta y_{it} \in (h_{j-1}, h_j]$ is equivalent to $y_{it+1} \in (\kappa_{j-1,it}, \kappa_{j,it}]$. Then, the probability that future income falls in bin j is

$$\begin{aligned} p^*_{j,it}(\eta_i) &\equiv \Pr\big(h_{j-1} < \Delta y_{it} \leqslant h_j \,\big|\, \mathcal{W}_{it}, \eta_i\big) \\ &= \Pr(y_{it+1} \leqslant \kappa_{j,it} \mid \mathcal{W}_{it}, \eta_i) - \Pr(y_{it+1} \leqslant \kappa_{j-1,it} \mid \mathcal{W}_{it}, \eta_i) \\ &= \Phi\left(\frac{\kappa_{j,it} - \rho(\mathbf{X}_{it}) y_{it} - \mathbf{X}_{it}'\beta - \eta_i}{\sigma(\mathbf{X}_{it})}\right) - \Phi\left(\frac{\kappa_{j-1,it} - \rho(\mathbf{X}_{it}) y_{it} - \mathbf{X}_{it}'\beta - \eta_i}{\sigma(\mathbf{X}_{it})}\right), \end{aligned} \tag{8}$$

where $\Phi(\cdot)$ denotes the standard Normal cumulative distribution function (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 design as requiring respondents to express their beliefs using M discrete probability points, and model

$$\mathbf{d}_{it} \equiv (d_{1it}, \ldots, d_{Jit})' \sim \text{Multinomial}(M, \mathbf{p}^*_{it}(\eta_i))\,, \; \mathbf{p}^*_{it}(\eta_i) \equiv \big(p^*_{1,it}(\eta_i), \ldots, p^*_{J,it}(\eta_i)\big)'. \tag{9}$$

Importantly, this approach treats observed survey responses as noisy discretized representations of an underlying continuous subjective income distribution. The specification captures discretization-induced reporting noise and can be extended to accommodate additional forms of reporting behavior or rounding.[9]

[8] One can perform similar calculations assuming logistic shocks, following Arellano et al. (2024). We use the Normal distribution because it is the standard assumption in quantitative macroeconomic models of income dynamics.

[9] One may argue that households internally discretize probabilities using a number of points different from the survey value M. To account for this possibility, one could introduce an additional latent

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

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

independently across households. Conditional on η_i, the per-period likelihood contribution is given by the multinomial probability mass function. The integrated likelihood contribution for household i is

$$L_i(\theta) = \int \prod_{t=1}^{T} \left[\frac{M!}{\prod_{j=1}^{J} d_{jit}!} \prod_{j=1}^{J} \left(p^*_{j,it}(\eta_i)\right)^{d_{jit}} \right] \frac{1}{\sigma_\eta} \phi\left(\frac{\eta_i}{\sigma_\eta}\right) d\eta_i, \tag{10}$$

where $\phi(\cdot)$ denotes the standard Normal probability density function (PDF), and θ collects the parameters governing $\rho(\mathbf{X}_{it})$, $\sigma(\mathbf{X}_{it})$, β, and σ_η^2. The sample log-likelihood is therefore

$$\ell_N(\theta) = \sum_{i=1}^{N} \log L_i(\theta), \qquad \widehat{\theta} = \arg\max_{\theta} \ell_N(\theta). \tag{11}$$

Because the integrated likelihood in (10) does not admit a closed-form solution, we approximate the integral numerically using Gauss–Hermite quadrature with 12 nodes. Standard errors are computed using a non-parametric bootstrap with 500 replications.

Implementation. A key modeling choice concerns the specification of $\rho(\mathbf{X}_{it})$ and $\sigma(\mathbf{X}_{it})$. In our empirical implementation, we adopt a logistic specification for persistence and an exponential specification for volatility:[10]

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

where $\Lambda(z) = \exp(z)/(1+\exp(z))$. These functional forms ensure that persistence lies in the unit interval and that volatility remains strictly positive.

Motivated by Meghir and Pistaferri (2004) and Browning et al. (2010), who emphasize age and education as important dimensions of heterogeneity in income dynamics, the state variables in $\mathbf{X}_{it}$ include age and education dummies, together with year fixed effects

reporting layer (e.g., household-specific latent M_i). In our baseline specification, however, we treat M as fixed and known from the survey design.

[10] We also experimented with linear specifications for $\rho(\mathbf{X}_{it})$ and $\sigma(\mathbf{X}_{it})$. These alternatives provided a substantially poorer fit to the observed bin probabilities when estimated using subjective expectations data.

that capture aggregate shocks.[11] We also include log income y_{it} in the volatility equation to allow for potential nonlinearities, so that $\sigma(\cdot)$ can be rewritten as $\sigma(\mathbf{X}_{it}, y_{it})$.

4.1.2 Nonlinear income process

Motivated by recent work on nonlinear 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 modeled through the conditional quantile function

$$y_{it+1} = Q_y(u_{it+1} \mid y_{it}, \mathbf{X}_{it}; \theta), \qquad u_{it+1} \sim U(0,1), \quad t \geqslant 1, \tag{12}$$

which we parameterize following Arellano et al. (2017) as

$$Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) = \sum_{k=0}^{K} a_k(\tau)\, \psi_k(y_{it}) + \mathbf{X}_{it}'\beta(\tau), \tag{13}$$

where $\psi_k(\cdot)$ denotes 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 $\{\tau_1, \ldots, \tau_L\} \subset (0,1)$. To discipline the behavior of the quantile function near $\tau = 0$ and $\tau = 1$, we impose exponential-tail restrictions at the boundary knots with parameters λ_1 and λ_L.

If next-period income y_{it+1} were directly observed, the parameters of the income process could be estimated using standard quantile regressions. In our setting, however, the observed data consist only of probabilities assigned to intervals of the conditional distribution of y_{it+1} given y_{it} and $\mathbf{X}_{it}$. Moreover, estimating the model by standard maximum likelihoods as in the state-dependent AR(1) model is infeasible given the quantile representation, the nonlinearities present in the functional form, and the restrictions required to recover a proper distribution. To address these issues, we interpret each of the M reported probability points as corresponding to an underlying latent draw from the household's perceived income distribution. Specifically, we treat the latent outcomes $\{y_{it+1}^{(m)}\}_{m=1}^{M}$ as missing data and observe only their bin membership through $\mathbf{d}_{it}$. We then estimate the income process using a stochastic EM algorithm similar to that proposed by Arellano and Bonhomme (2016).

This approach is feasible because the parameterization of the quantile function, together with the closed-form tail restrictions, allows us to approximate the conditional

[11] We also estimate specifications including employment characteristics. Results are largely unchanged.

cumulative distribution function (CDF)

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

which in turn allows us to simulate latent realizations of y_{it+1} consistent with the observed survey responses. We outline the main steps of the algorithm below and relegate additional details to Appendix C.

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

$$f\big(y \mid y_{it}, \mathbf{X}_{it},\, y \in I_{j,it}; \theta^{(s)}\big) \propto f\big(y \mid y_{it}, \mathbf{X}_{it}; \theta^{(s)}\big)\ \mathbf{1}\{y \in I_{j,it}\}. \tag{14}$$

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

1. **Stochastic E-step.** For each household-period pair (i,t) and each bin j, we draw d_{jit} latent realizations from the truncated distribution in (14). Pooling these simulated observations across bins yields the augmented sample

   $$\{y_{it+1}^{(s,m)}, y_{it}, \mathbf{X}_{it}\}_{i,t,m},$$

   which contains exactly M latent draws for each (i,t) pair.

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

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

   where $\rho_\tau(\cdot)$ denotes the standard quantile-regression check function, $\rho_\tau(u) = u(\tau - \mathbf{1}(u < 0))$.

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

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

for a pre-specified tolerance level ε. As in the state-dependent AR(1) specification, standard errors are computed using a non-parametric bootstrap with 500 replications.

4.2 A hurdle model for bunching behavior

A large fraction of respondents allocate all M points to the middle bin. As discussed earlier, this response pattern may reflect genuinely low perceived income risk for some households, but it may also arise from cognitive difficulties or focal-point reporting behavior. To account for this possibility, we augment the model with a hurdle, or two-regime, reporting structure by adapting the finite-mixture approaches of de Bresser and van Soest (2013), Kleinjans and van Soest (2014), and Heiss et al. (2022) to our setting.[12]

We illustrate the hurdle model in the context of the state-dependent AR(1) specification and relegate the nonlinear extension to 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 respondents follow a sequential reporting process:

1. First, they enter either the deterministic focal-point regime or the non-focal reporting regime.

2. Conditional on entering the non-focal regime, they allocate the M points across bins according to the multinomial probabilities implied by their perceived income process.

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

$$\pi_{it}(b_i; \gamma) \equiv \Pr(R_{it} = 1 \mid \mathbf{Z}_{it}, b_i) = \Lambda(\gamma_0 + \mathbf{Z}'_{it}\gamma_1 + b_i)\,, \tag{15}$$

where

$$\gamma \equiv (\gamma_0, \gamma_1')'$$

collects the parameters of the reporting equation. The term b_i is a time-invariant respondent-specific random effect capturing the propensity to exhibit focal-point reporting. The vector $\mathbf{Z}_{it}$ contains covariates that may overlap with $\mathbf{X}_{it}$ as well as variables

[12] An alternative nonparametric approach to handling rounding behavior is proposed by Manski and Molinari (2010) and Giustinelli et al. (2022b), who transform point expectations into intervals based on survey responses. Because our objective is to estimate structural income processes, we adopt a more parametric approach.

excluded from the income process. In particular, we write

$$\mathbf{Z}_{it} = (\mathbf{X}'_{it}, \mathbf{V}'_{it})',$$

where $\mathbf{V}_{it}$ affects reporting behavior but does not directly affect the underlying perceived income process.

In the empirical implementation of the hurdle model, we use households' response patterns in the house-price expectations question as an exclusion restriction. Because the income and house-price expectations questions share a nearly identical probabilistic format, it is plausible that cognitive limitations or reporting styles affect responses similarly across domains, consistent with the evidence reported in Table 2.[13]

In the baseline specification, we assume that the household-specific random effect in the income process, η_i, and the random effect in the reporting equation, b_i, are independent:[14]

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

We also maintain that (η_i, b_i) is independent of the observed covariate histories.

Let θ denote the vector of income-process parameters introduced in Section 4.1, including σ_η^2. We collect all parameters of the hurdle model in

$$\vartheta \equiv \left(\theta', \gamma', \sigma_b^2\right)'.$$

Likelihood function. Conditional on η_i, define the probability that the non-focal multinomial reporting mechanism places all M points in the middle bin as

$$A_{it}(\eta_i; \theta) \equiv \Pr(d_{j^\star,it} = M \mid \mathbf{W}_{it}, \eta_i; \theta) = \left[p^*_{j^\star,it}(\eta_i; \theta)\right]^M,$$

where

$$\mathbf{W}_{it} \equiv (y_{it}, \mathbf{X}_{it})$$

and $p^*_{j^\star,it}(\eta_i; \theta)$ denotes the model-implied probability of the middle bin under the state-dependent AR(1) process in equation (1).

[13] Household income dynamics would be directly affected by house-price expectations only if both variables were driven by a common aggregate factor. The inclusion of year fixed effects in the income process absorbs such aggregate co-movement.

[14] Allowing for correlated random effects is straightforward by replacing the diagonal covariance matrix with a full bivariate Normal covariance matrix.

Conditional on (η_i, b_i), the per-period likelihood contribution is

$$L_{it}^{H}(\eta_i, b_i; \vartheta) = \begin{cases} \pi_{it}(b_i;\gamma) + [1 - \pi_{it}(b_i;\gamma)]\, A_{it}(\eta_i;\theta), & \text{if } S_{it} = 1, \\ [1 - \pi_{it}(b_i;\gamma)]\, \dfrac{M!}{\prod_{j=1}^{J} d_{jit}!} \displaystyle\prod_{j=1}^{J} \left[p_{j,it}^{*}(\eta_i;\theta)\right]^{d_{jit}}, & \text{if } S_{it} = 0. \end{cases} \tag{16}$$

The two cases admit a natural interpretation:

- If $S_{it} = 1$, the observed bunching response may arise either because the respondent is in the deterministic focal-point regime, with probability $\pi_{it}(b_i;\gamma)$, or because the respondent is in the non-focal regime and all M points happen to be allocated to the middle bin, with probability

$$[1 - \pi_{it}(b_i;\gamma)]\, A_{it}(\eta_i;\theta).$$

- If $S_{it} = 0$, the deterministic focal-point regime assigns zero probability to the observed allocation. The likelihood contribution is therefore the non-focal multinomial likelihood multiplied by

$$1 - \pi_{it}(b_i;\gamma).$$

The integrated likelihood contribution for household i is

$$L_i^{H}(\vartheta) = \iint \prod_{t=1}^{T_i} L_{it}^{H}(\eta_i, b_i; \vartheta) f_\eta(\eta_i;\sigma_\eta^2) f_b(b_i;\sigma_b^2)\, d\eta_i\, db_i, \tag{17}$$

where

$$f_\eta(\eta_i;\sigma_\eta^2) = \phi(\eta_i; 0, \sigma_\eta^2)$$

and

$$f_b(b_i;\sigma_b^2) = \phi(b_i; 0, \sigma_b^2).$$

The sample log-likelihood is

$$\ell_N^{H}(\vartheta) = \sum_{i=1}^{N} \log L_i^{H}(\vartheta),$$

and the maximum-likelihood estimator is

$$\widehat{\vartheta} = \arg\max_{\vartheta \in \mathcal{V}} \ell_N^{H}(\vartheta),$$

where $\mathcal{V}$ denotes the parameter space of the hurdle model.

We estimate the model by maximum likelihood and approximate the two-dimensional integrals in equation (17) using Gauss–Hermite quadrature. Standard errors are computed using a nonparametric bootstrap with 500 replications.

The hurdle model restricts deviations from unbiased probability reporting to arise through focal-point responses in which all probability mass is allocated to the middle bin. We also consider a more flexible specification in which households may systematically distort reported probabilities by tilting the perceived distribution toward optimism, pessimism, or excessive concentration around the center. Details of this *hurdle model with tilted probabilities* are presented in Appendix G. Estimation results are very similar to those obtained under the baseline hurdle model, suggesting that focal-point behavior is the primary source of elicitation bias in the data.

4.3 Identification

The object that we would like to recover is the perceived conditional distribution of future log income y_{it+1},

$$F_y(\cdot \mid \mathcal{W}_{it}), \tag{18}$$

where $\mathcal{W}_{it} \equiv (y_{it}, \mathbf{X}_{it})$ collects current household income and household characteristics. Households report their perceived distribution by allocating M probability points across pre-specified intervals of income growth. In this subsection, we briefly discuss the identification of the conditional distribution. A more formal analysis of identification is in Appendix E.

Given minimal assumptions on the data, the conditional distribution is only partially identified. This is because the data pin down the perceived CDF at the fixed cut-off points, while leaving its shape within each interval unrestricted. Multiple conditional distributions can therefore generate the same set of observed bin probabilities; any values within the interval defined by the bins can be generated by some conditional distribution consistent with the observed data, as in Manski and Tamer (2002). Hence, in the absence of additional assumptions, subjective expectations data alone do not nonparametrically

identify the full conditional distribution of future income or its associated moments.[15]

State-dependent AR(1) process. In the state-dependent AR(1) specification, the household-specific perceived income dynamics satisfy

$$y_{i,t+1} \mid W_{it}, \eta_i \sim N\left(\rho(X_{it})y_{it} + X_{it}'\beta + \eta_i, \sigma^2(X_{it})\right),$$

where

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

is a household-specific random effect. Under the maintained random-effects exogeneity assumption, integrating over η_i gives

$$y_{i,t+1} \mid W_{it} \sim N\left(\mu(W_{it}), \sigma_e^2(X_{it})\right),$$

where

$$\mu(W_{it}) = \rho(X_{it})y_{it} + X_{it}'\beta$$

and

$$\sigma_e^2(X_{it}) = \sigma^2(X_{it}) + \sigma_\eta^2.$$

The Normal location–scale restriction determines the full marginal CDF from $\mu(W_{it})$ and $\sigma_e(X_{it})$. At a given state, two distinct finite cutoffs whose cumulative probabilities are strictly between zero and one identify these two objects. Variation in current income conditional on X_{it}, together with the relevant full-rank conditions, then identifies the parameters governing $\rho(X_{it})$ and the conditional mean.

The cross-sectional distribution identifies only the total conditional variance $\sigma_e^2(X_{it})$. Under random-effects exogeneity and conditional independence of reporting draws across waves given the household effect, cross-wave dependence in the reported allocations identifies σ_η^2. The within-household innovation variance can then be recovered as

$$\sigma^2(X_{it}) = \sigma_e^2(X_{it}) - \sigma_\eta^2.$$

Thus, the panel separates persistent cross-household heterogeneity from the within-household uncertainty represented by the income innovations.

[15] One potential route to nonparametric identification would be to allow cutoff values to vary across households, as in the min-max elicitation approach of Arellano et al. (2024). If the support of the thresholds were sufficiently rich and the corresponding probabilities observed, the full conditional distribution could in principle be recovered.

Nonlinear income process. Under the nonlinear income process, we restrict the conditional quantile function $Q_y(\tau \mid W_{it})$ to a finite-dimensional sieve class. At a fixed sieve dimension, the observed allocations identify the model-implied CDF at the bin cutoffs. The parameters of the nonlinear process are point identified within the maintained sieve model provided that the mapping from those parameters to the collection of cutoff CDF values is injective. Monotonicity and the tail restrictions ensure that each admissible parameter vector defines a proper conditional distribution, while the sieve specification determines its shape within the bins.

Under these conditions, the conditional quantile function and the associated measures of persistence, volatility, and skewness are identified within the maintained model. This is a parametric or sieve-based identification result: the within-bin shape is supplied by the model restrictions rather than identified nonparametrically from the reported bin probabilities.

Hurdle model. The hurdle model introduces a second reporting regime. With probability $\pi_{it}(b_i)$, the household enters the deterministic focal-point regime and allocates all M points to the middle bin. With probability $1 - \pi_{it}(b_i)$, the allocation is generated by the underlying perceived income distribution.

Middle-bin bunching can therefore arise either from genuinely concentrated beliefs or from deterministic focal-point reporting. The excluded reporting variable shifts the probability of entering the focal regime while, conditional on the observed state and year effects, leaving the perceived income distribution unchanged. This excluded variation establishes relevance for the reporting component.

Point identification additionally requires that the non-focal allocation distribution identify the parameters of the underlying income process and that the mapping from the reporting equation parameters to focal-reporting probabilities be injective. Under these conditions, the income process parameters can first be recovered from the distribution of non-focal allocation patterns. Given the identified income process, the observed frequency of non-focal responses then identifies the focal-reporting probabilities and the parameters of the reporting equation.

5 Results

This section compares the income processes recovered from subjective expectations data with those estimated from realized income data, focusing on both the state-dependent AR(1) specification and the nonlinear quantile model. Across specifications, a consistent pattern emerges: households' subjective expectations imply substantially more persistent and less dispersed income dynamics than those inferred from realized income histories. These differences are economically meaningful and have important implications for precautionary saving, consumption smoothing, and the extent of self-insurance against income risk.

5.1 Coarse information

We begin by presenting results from the baseline specifications, which account for the coarse nature of the survey responses but abstract from systematic reporting biases or focal-point behavior.

5.1.1 State-dependent AR(1) process

Table 3 reports the estimated mean and dispersion of income persistence and volatility for the realized and subjective income processes. The results reveal substantial differences between perceived and realized income dynamics. Persistence estimated from realized income data is 0.692, whereas subjective expectations imply a much higher value, close to unity (0.999). In contrast, income volatility is substantially lower in the subjective income process: realized income data imply an average volatility of 0.465, while subjective expectations imply an average volatility of 0.041.

Figure 2 further illustrates these differences. The distributions of persistence and volatility differ markedly depending on whether the process is estimated from realized income data or from subjective expectations. Overall, the results point to two main empirical patterns: households perceive future income as substantially more persistent than estimates based on realized income histories imply,[16] and they perceive considerably

[16]This pattern is consistent with the evidence on over-persistence in subjective income expectations documented by Rozsypal and Schlafmann (2023).

Table 3: Persistence and volatility in the state-dependent AR(1) process

	Mean (μ_p, μ_σ)	Standard deviation $(\varsigma_p, \varsigma_\sigma)$
Panel A: Realized income process		
Persistence	0.6915	0.0407
	[0.6584,0.7155]	[0.0323,0.077]
Volatility	0.4651	0.0558
	[0.4543,0.4877]	[0.0428,0.0752]
Panel B: Subjective income process		
Persistence	0.9998	–
	[0.9539, 0.9999]	–
Volatility	0.0411	0.0105
	[0.0320, 0.1426]	[0.0084, 0.5629]

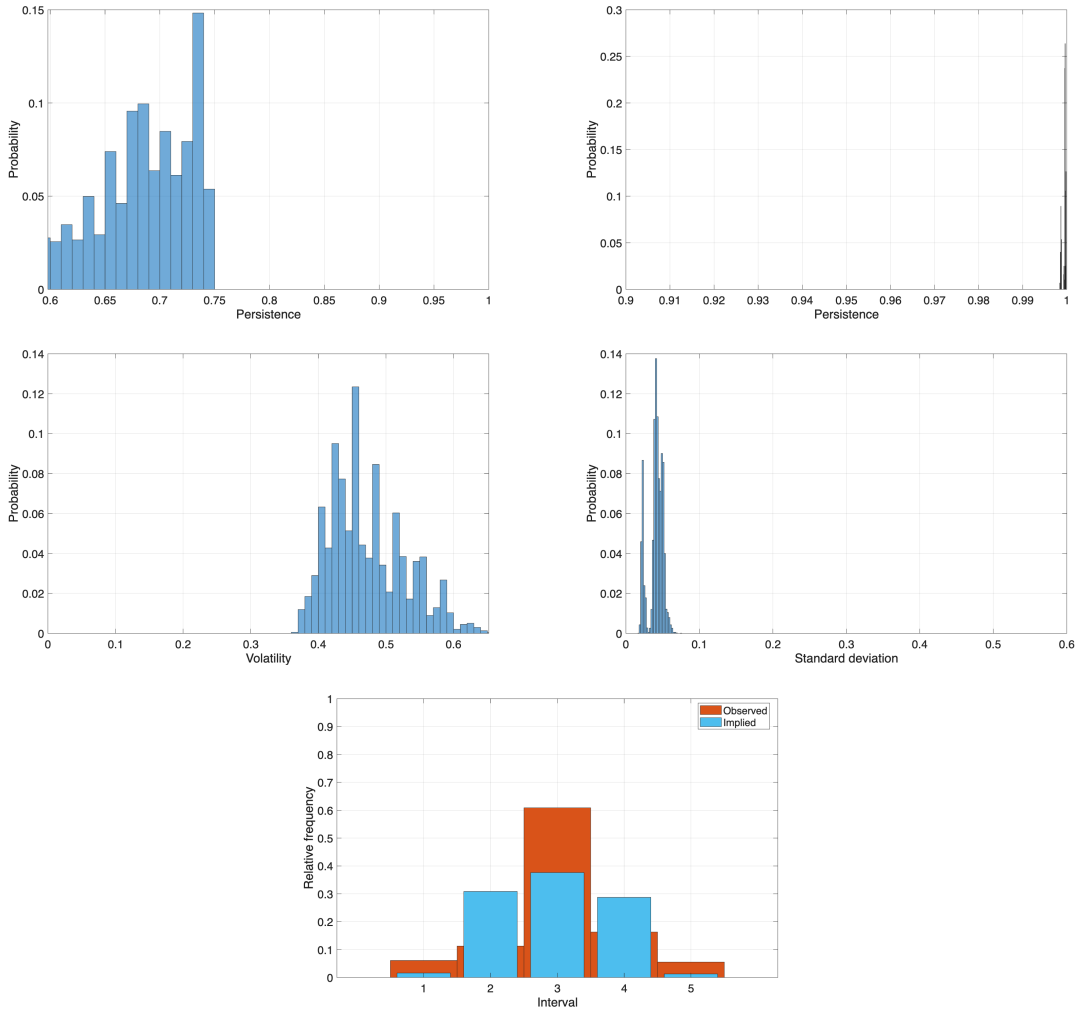
Notes: This table reports estimates of the mean and standard deviation of persistence, μ_p and ς_p in (2), and volatility, μ_σ and ς_σ in (3), for the realized and subjective income processes. The specifications include wave dummies and controls for education and employment characteristics. Brackets report 95% block-bootstrap confidence intervals computed using 500 bootstrap replications.

less volatility in future income than is observed in realized income dynamics.

The model fit further illustrates the limitations of the baseline specification. As the bottom panel of Figure 2 shows, the subjective AR(1) model fails to reproduce the sharp concentration of probability mass in the middle bin and underpredicts the probability assigned to more extreme income-growth outcomes, yielding a root-mean-squared error (RMSE) of 0.25. This relatively poor fit highlights the difficulty of reconciling coarse and highly bunched probability allocations with a smooth Gaussian income process and motivates the extension below that explicitly accounts for focal-point reporting behavior.

Household-specific volatility. Appendix D reports results for an alternative AR(1) specification that allows volatility to be household-specific. This additional flexibility substantially improves the fit of the subjective expectations data, reducing the RMSE from 0.25 in the baseline AR(1) specification to 0.05. Nevertheless, the qualitative conclusions remain unchanged: subjective expectations imply income dynamics that are much more persistent and substantially less volatile than those estimated from realized income histories.

Figure 2: Persistence, volatility, and model fit in the state-dependent AR(1) process



Notes: The top row shows the distribution of persistence, and the middle row shows the distribution of volatility for the state-dependent AR(1) process estimated 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 panel compares model-implied probabilities from the subjective income process with the observed probabilities in the survey data.

5.1.2 Nonlinear income process

Figure 3 reports results from the nonlinear quantile-based income model estimated without accounting for focal-point behavior. Three patterns stand out. First, the generalized persistence surface estimated from realized income data displays rich nonlinearities: persistence varies with both current income and the rank of the shock. By contrast, the subjective model generates an almost flat persistence surface, suggesting that households perceive shocks as similarly persistent across the income distribution and across different types of shocks. Second, subjective expectations again imply a more compressed distribution of income risk. Conditional volatility is uniformly lower than that inferred from realized income data, reinforcing the finding that households perceive future income as substantially less dispersed than realized income histories would suggest. Third, conditional skewness varies with household rank in the realized income process, while the subjective process implies a much flatter skewness profile. In particular, households perceive future income risk as negatively skewed across much of the income distribution.

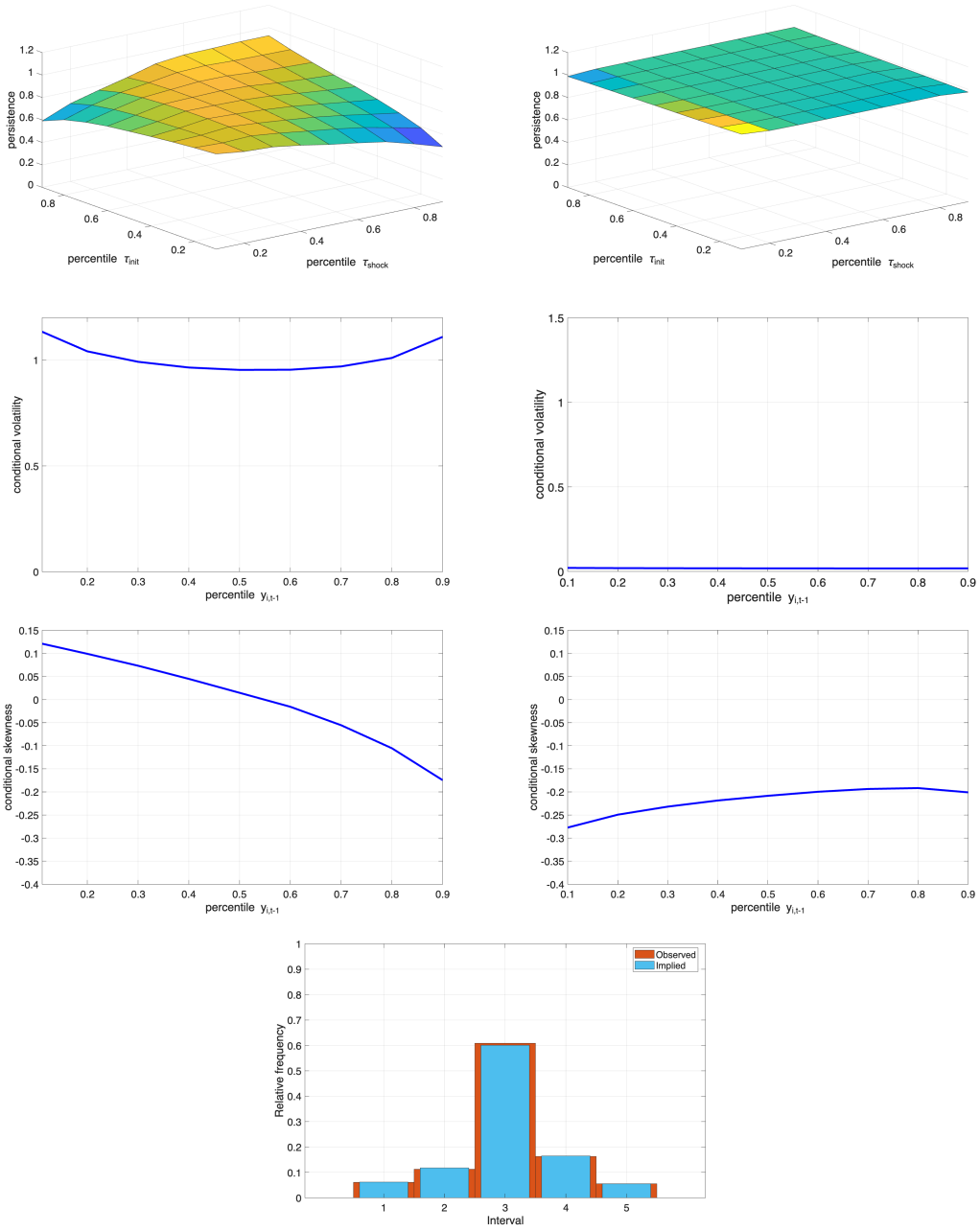
The nonlinear subjective model fits the observed probability allocations well, with an RMSE of around 0.0045. Figure I.4 reports 95% non-parametric bootstrap confidence intervals, showing that the estimated nonlinear objects are precisely estimated.

The differences between income processes inferred from realized outcomes and from subjective expectations are substantial, regardless of whether we use the state-dependent AR(1) specification or the nonlinear quantile model. One potential concern is that the two sources of information refer to different horizons. The realized-income estimates use income observations from the 2014, 2017, and 2020 EFF waves,[17] so household income is observed at three-year intervals. By contrast, the subjective expectations question refers to expected income growth over the next twelve months.

Could this horizon difference explain the large gap between realized and subjective income processes? To assess this possibility, we use data from the Spanish Living Conditions Survey (ECV), which allows us to estimate both the state-dependent AR(1) and nonlinear income processes using one-year and three-year income changes. This exercise

[17] We obtain similar results when using waves from 2002–2020.

Figure 3: Nonlinear persistence, volatility, skewness, and model fit



Notes: The first row shows nonlinear persistence, the second row conditional volatility, and the third row conditional skewness for the nonlinear income process estimated without modeling focal-point behavior. The left column corresponds to the process estimated from realized income data, while the right column corresponds to the process estimated from subjective expectations data. The bottom panel compares model-implied probabilities from the subjective income process with the observed survey probabilities.

isolates how much estimated income dynamics change solely because of the length of the observation interval. The results are reported in Appendix F.

The evidence suggests that horizon differences are not driving our main findings. First, income processes estimated using three-year intervals in the ECV are very similar to those obtained from the EFF. Second, although persistence is slightly lower and volatility slightly higher when moving from one-year to three-year income changes, these differences are small relative to the uncovered gap between realized and subjective income processes.

While the absence of annual realized income data for the same EFF households prevents a definitive comparison, the ECV evidence indicates that differences in observation horizons are unlikely to account for the large discrepancies we document. In Section 6, when embedding the estimated processes into a life-cycle model, we explicitly account for the different horizons by adjusting the underlying transition matrices.

5.2 Hurdle model for bunching behavior

We now present the empirical results obtained when we explicitly account for focal-point reporting behavior through the hurdle model. These specifications allow systematic deviations from unbiased probability reporting to arise from the tendency of some households to allocate all probability mass to the middle bin.

5.2.1 State-dependent AR(1) process

We report results for the following state dependent AR(1) specification:

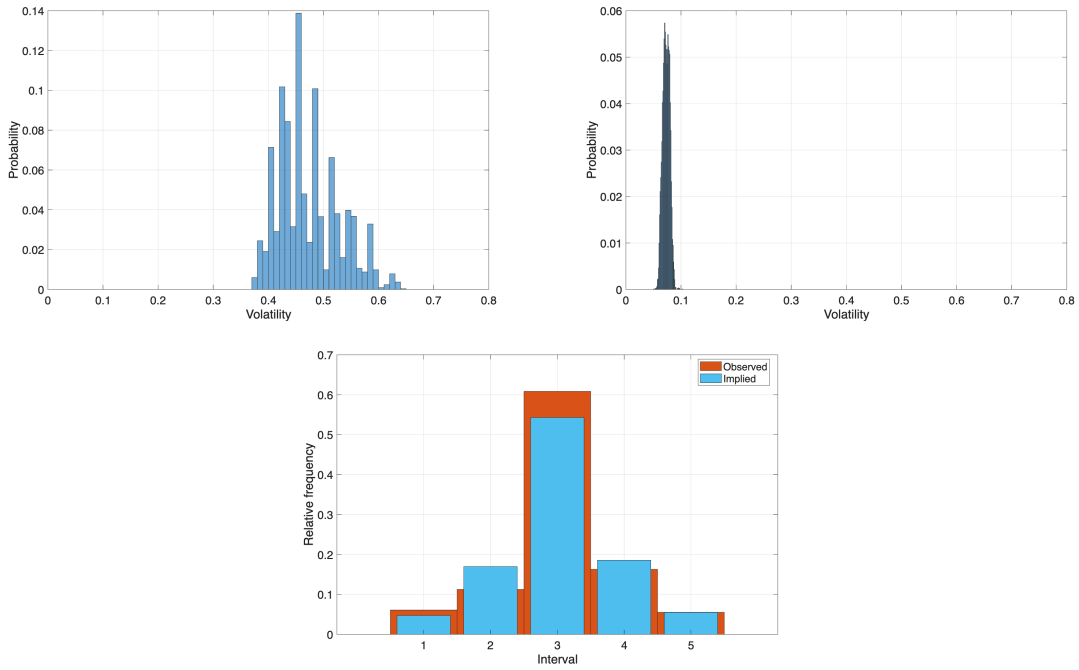
$$y_{it+1} = \rho y_{it} + \eta_i + \mathbf{X}'_{it}\beta + \sigma(\mathbf{X}_{it})u_{it+1}, \qquad u_{it+1} \sim N(0,1).$$

Relative to the baseline state-dependent AR(1) model, we restrict persistence to be homogeneous across households. This simplification is motivated by the fact that, in the baseline subjective income process, the estimated distribution of $\rho(\mathbf{X}_{it})$ is tightly concentrated around 0.999. We therefore focus on the single estimated persistence parameter and on the distribution of state-dependent volatility.

Allowing for focal-point reporting behavior improves model fit substantially (Figure 4). The hurdle extension increases the estimated volatility of the subjective income

process while leaving persistence essentially unchanged. The model also captures the shape of the observed bin probabilities more closely, especially in the tails, reducing the RMSE to 0.12. Nevertheless, the middle bin remains difficult to match exactly, suggesting that the degree of focal-point behavior is sufficiently large that even a two-regime mixture cannot fully reproduce the concentration of responses around income stability.

Figure 4: Volatility and model fit in the AR(1) process with hurdle behavior



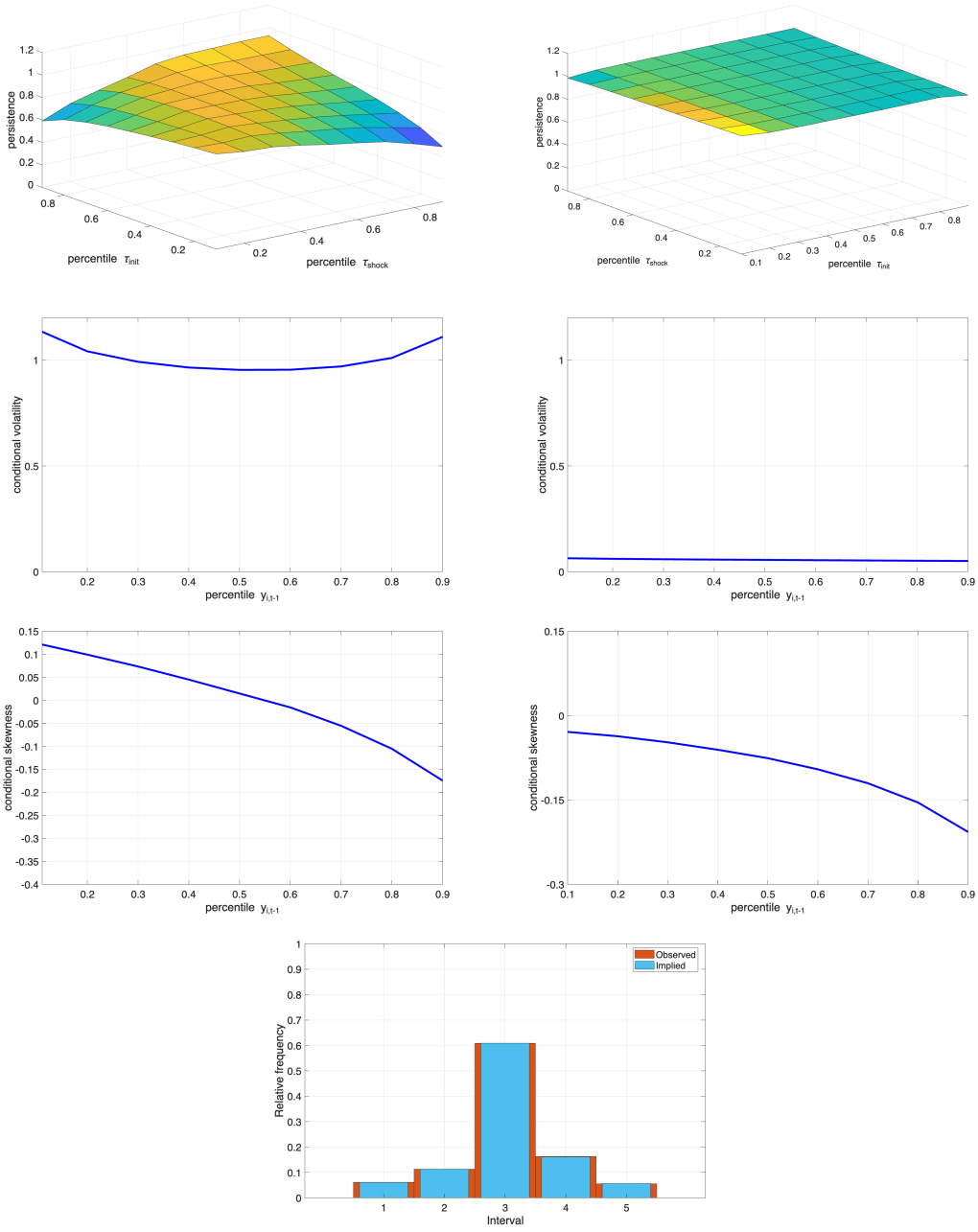
Notes: The top row shows the distribution of volatility for the AR(1) process estimated with the hurdle model for bunching behavior. The left panel corresponds to the process estimated from realized income data, while the right panel corresponds to the process estimated from subjective expectations data. The bottom panel compares model-implied probabilities from the subjective income process with the observed survey probabilities.

5.2.2 Nonlinear process

Figure 5 reports results from the nonlinear income process estimated with the hurdle model for focal-point behavior. The main qualitative findings from Figure 3 remain: subjective expectations imply income dynamics that are more persistent and less volatile than those inferred from realized income data. However, accounting for focal-point behavior leads to some important changes.

First, conditional volatility remains substantially lower in the subjective income pro-

Figure 5: Nonlinear persistence, volatility, skewness, and model fit with hurdle behavior



Notes: The first row shows nonlinear persistence, the second row conditional volatility, and the third row conditional skewness for the nonlinear income process estimated with the hurdle model for focal-point behavior. The left column corresponds to the process estimated from realized income data, while the right column corresponds to the process estimated from subjective expectations data. The bottom panel compares model-implied probabilities from the subjective income process with the observed survey probabilities.

cess than in the realized income process, but it increases relative to the specification that does not model bunching. In particular, the average value rises from around 0.02 in the baseline subjective process to approximately 0.06 once focal-point behavior is accounted for. Second, the most notable change concerns conditional skewness. With the hurdle correction, the subjective income process generates a skewness profile that is more consistent with the patterns documented by Arellano et al. (2017): high-income households perceive negative skewness in future income risk, whereas low-income households perceive little skewness. These results highlight the importance of distinguishing genuine beliefs about income risk from focal-point reporting behavior.

Figure I.5 reports the 95% non-parametric bootstrap confidence intervals. As in the previous specification, the estimated nonlinear objects are precisely estimated.

Taken together, the nonlinear estimates confirm the main message from the AR(1) analysis: households perceive income as highly persistent and substantially less volatile than realized income histories would imply. At the same time, accounting for focal-point behavior is important for recovering richer distributional features of subjective income risk, especially conditional skewness.

6 Implications for consumption insurance and wealth accumulation

We uncover substantial differences between the income processes inferred from realized income data and those recovered from subjective expectations. In this section, we employ a life cycle model with heterogeneous agents and incomplete asset markets to study the key implications of these differences for wealth accumulation over the life cycle and for consumption insurance.

Our aim is to assess whether each data-generating process, taken in isolation, leads to different economic outcomes, rather than to compare their relative performance within a structural framework in matching a specific set of moments. Given the distinct patterns of wealth accumulation and consumption insurance we obtain, such a comparison would only be informative if the underlying drivers of these differences were explicitly modeled

and accounted for—a task we leave for future research.[18]

We first present the main features of our model economy and describe its calibration, discussing how we embed the estimated income processes 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 probability is zero at period $T+1$). Population grows at a constant rate g_n.[19] 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 (y_{J_R}). Agents in working age ($j \leqslant J_R$), supply one unit of labor inelastically at zero disutility, receiving stochastic log income y_j, which follows either the nonlinear process estimated from realized data (NL) or the nonlinear process estimated from subjective expectations, with or without accounting for focal-point behavior (NL-SE Focal and NL-SE, respectively). Agents pay income taxes, denoted by T_j, depending on their total income $I_j = exp(y_j) + ra_j$, where r is the net return on asset holdings a_j. Income taxes are given by $\tau(I_j)I_j$, where the tax rate $\tau(I_j)$ is a function of total income. Finally, agents may receive transfers $G(I_j)$.[20]

[18] As discussed in the next section, we conjecture that the different data-generating processes we obtain may reflect that households have superior information about their income risk relative to the econometrician, or that they face cognitive limitations or exhibit biases when assessing the probability of future states of the world.

[19] 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 $\mu_{j+1} = \frac{s_{j+1}}{(1+g_n)}\mu_j$.

[20] 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

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

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

$$
\begin{aligned}
V_j(y,a) &= \max_{a',c} \left\{ \frac{c^{1-\sigma}}{1-\sigma} + \beta s_{j+1} \mathbb{E}\left[V_{j+1}(y',a') \mid y \right] \right\}, \\
\text{s.t.} \quad a' &= exp(y) - T_j(I_j) + G(I_j) - c + (1+r)a, \qquad a' \geqslant A, \\
y' &= Q_j^P(y,u'), \qquad u' \sim U[0,1], \\
P &\in \{NL,\ NL\text{-}SE,\ NL\text{-}SE \text{ Focal}\}.
\end{aligned}
$$

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

$$
\begin{aligned}
W_j(y_{J_R},a) &= \max_{a',c} \left\{ \frac{c^{1-\sigma}}{1-\sigma} + \beta s_{j+1} W_{j+1}(y_{J_R},a') \right\}, \\
\text{s.t.} \quad a' &= PB_j(y_{J_R}) - c + (1+r)a, \qquad a' \geqslant A.
\end{aligned}
$$

The model is solved recursively by backward induction over the life cycle holding interest rates fixed and then simulated for 100000 households given a set of structural parameters, discussed next.

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.

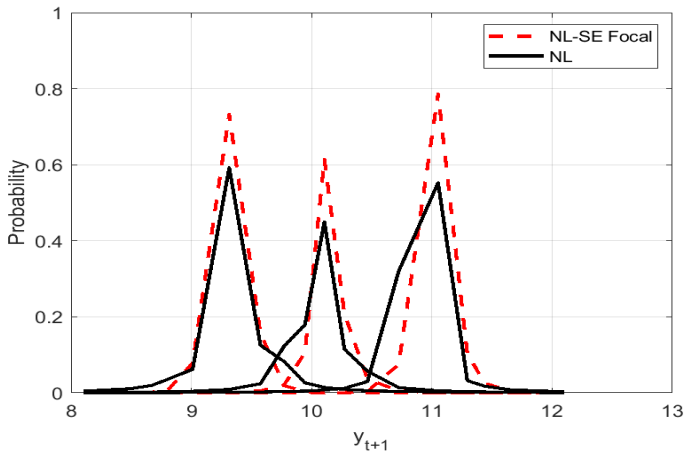
The coefficient of relative risk aversion is set to 2, a standard value. The risk-free rate is 2.5% and the discount factor β is calibrated to match a wealth to income ratio of 4 (calculate using EFF data). Finally, we set the exogenous borrowing limit A to 16% of average disposable household earnings, 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 nonlinear 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 $n = 18$.[21]

Figure 6: Transition Matrices



These matrices further illustrate the differences between the estimated processes using realized data (NL) and subjective expectations data accounting for focal behavior (NL-SE Focal). First, variances are smaller under NL-SE Focal. Second, the pro-

[21] The EFF 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. There is no unique way to adjust the transition matrices, thus, if 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 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). For robustness we also consider the model in which the period is set to three years, where such adjustments are not necessary.

cess 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 find little evidence of these nonlinearities in the *NL-SE* Focal processes.

Our aim is to compare wealth accumulation and consumption insurance across the different estimated income processes. Given that the variance of shocks under *NL-SE* Focal and *NL-SE* is lower, keeping the structural parameters unchanged while varying only the income process would result in asset-to-income ratios that are significantly lower than those obtained under *NL*, the process estimated using realized data. To avoid comparing economies with such different average saving rates, we keep all structural parameters constant but allow β to vary so that, in all cases, the asset-to-income ratio matches that observed in the data. We present the results with a fixed β in Appendix H.

Finally, we set the model period to one year, as is standard in life cycle models, and adjust the transition matrix of *NL* to account for the fact that the estimation is based on three-year intervals. In Appendix H, we also present an alternative specification in which the model period is set to three years and the transition matrices of *NL-SE* Focal and *NL-SE* are adjusted accordingly ($Q_3 = Q_1^3$, where Q_1 is the transition matrix obtained from the simulation exercises for these two income processes). Our main qualitative conclusions remain unchanged across all the alternative specifications.

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 with higher income save significantly more on average at the beginning of the life cycle. Under the estimated processes using subjective expectations data (*NL-SE* Focal and *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 a significantly higher cross section variance of wealth in the early part of the life cycle, as illustrated in Figure 8.

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[22]

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

Table 4 shows the results. We consider four versions of the model: one in which the process NL is assumed; one in which the process NL-SE is used; one in which NL-SE Focal is used, so focal-point behavior is accounted for and one in which households expect the income process to follow NL-SE Focal, 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 expectation formation. In the next section, we discuss the other main alternative explanation, which relies on households possessing superior private information.

We find that the degree of consumption insurance in the first three cases is very similar (around 0.78 in two cases, a bit lower for the NL-SE where the variance is the lowest and thus households save very little in the beginning of the life cycle). 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

[22]This measure can be obtained if the shock μ_{it} is known, which is the case in our model economy. An alternative 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.

Figure 7: Wealth Accumulation

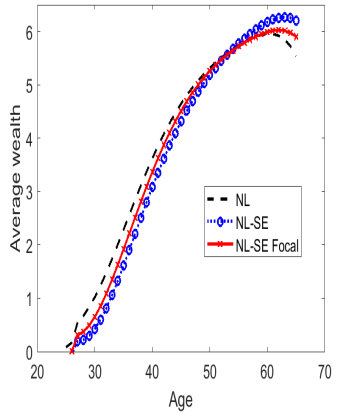


Figure 8: Cross-section Variance of Wealth

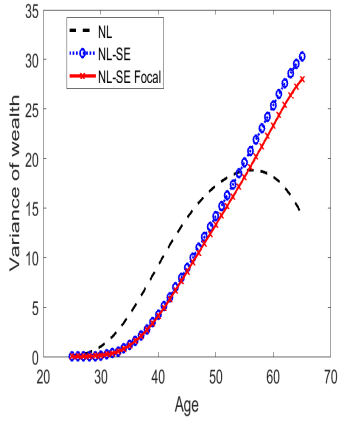


Table 4: Consumption Insurance

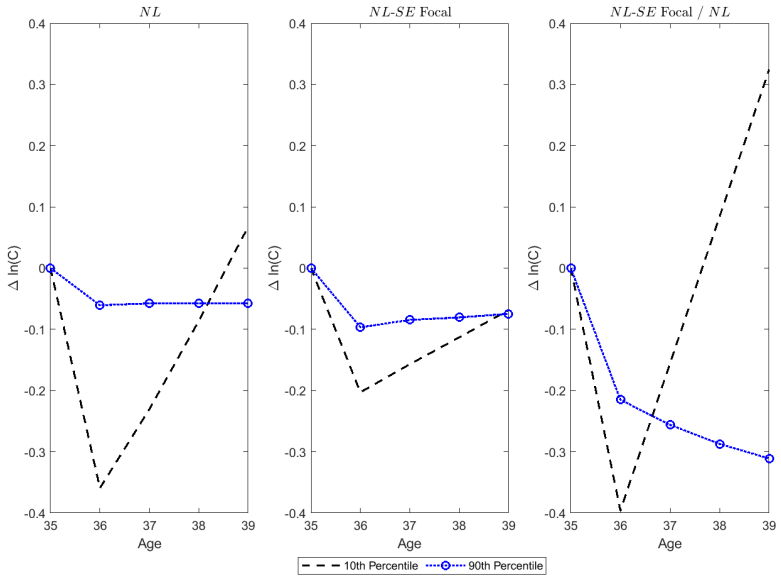
Income process	ψ
NL	0.781
NL-SE	0.730
NL-SE Focal	0.779
NL-SE Focal / NL	0.653

De Nardi et al. (2020)). Second, although the overall level of consumption insurance is similar across the first three 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.

To illustrate the different degrees of insurance that each specification affords across the distribution, we plot (Figure 9) the consumption response for households in the 10th and 90th percentiles of the wealth distribution after they face a negative shock equivalent to the 5th percentile of their future income distribution. We impose these shocks on 35-year-old individuals, assuming that throughout their life cycle up to this point, and thereafter, they receive(d) a series of shocks consistent with their income process.

Under the NL income process, rich households, anticipating the likelihood of nonlinear shocks, accumulate more assets and, despite receiving large negative shocks, are able to smooth consumption to a significant extent. Such shocks effectively wipe out

Figure 9: Consumption Impulse Responses



the memory of previous favorable shocks, permanently reducing consumption. Poor households hold significantly fewer assets and, when faced with large shocks, experience substantial consumption losses. However, given the relatively high likelihood of receiving favorable shocks that offset past negative shocks (the "good" aspect of the nonlinearity in NL), consumption eventually returns to its pre-shock level.

By contrast, under the NL-SE Focal process, both poor and rich households accumulate insufficient assets. As a result, even though they face relatively smaller shocks, they are similarly unable to smooth consumption, and consumption responses across the wealth distribution are more homogeneous than in the NL case. Therefore, while average consumption smoothing appears similar across the two income processes in the cross section, this masks substantial heterogeneity across households.

Finally, we present impulse responses for the case in which households choose their savings assuming a low-variance, high-persistence income process (NL-SE Focal), but instead experience shocks consistent with a more nonlinear, high-variance, and low-persistence process (NL). In this scenario, both rich and poor households incur significant consumption losses, implying that insurance is severely compromised. Due to the unexpected nonlinearity, rich households are unable to smooth consumption and continue to experience losses as they struggle to return to their previous income level. Poor

households, in contrast, eventually recover their consumption, as they are subsequently hit by unanticipated positive shocks.

7 Assessing the information content of subjective expectations

In the previous section, we showed that when income processes are consistent with households' subjective expectations, perceived income risk is low, allowing households to insure against risk while accumulating wealth more gradually over their lifetime. In contrast, when income processes are based on realized data, risk is higher and nonlinear, requiring households to accumulate substantial wealth early in life to insure against it.

We posit that if households expect a low-risk income process and plan their asset accumulation accordingly, while realized income exhibits higher variance and nonlinearities, consumption insurance will be negatively affected. As a result, consumption becomes more volatile and more closely tied to income fluctuations. The key assumption underlying the exercise is that households form biased expectations: although they face high-variance, nonlinear shocks, they do not fully incorporate them into their assessment of future income risk, possibly due to the limited salience of large shocks (see Bordalo et al. (2013)).

Nonetheless, the differences between the estimated processes may reflect that households possess private information about future income risk that is not observed by the econometrician, who relies on realized data. As a result, income risk may be overstated for some households. In particular, large fluctuations observed for a subset of households may be incorrectly generalized to others who face a low probability of such shocks—information that is instead captured in their subjective expectations. This raises the question: do households' subjective expectations contain additional information?

We begin by examining whether incorporating subjective expectations into a standard autoregressive model with household controls and time fixed effects improves income prediction (Table 5). Specifically, we include the mean and variance of the subjective income distribution. While the mean is not statistically significant—likely because it is

close to current income—the dispersion is both economically and statistically significant.

Table 5: 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 examine whether subjective income expectations are related to household consumption by analyzing consumption and elicited MPCs.[23] When controlling for current income, expected income adds little explanatory power, whereas the dispersion of future income is both economically and statistically significant, with a negative effect on consumption and elicited MPCs. These findings suggest that subjective expectations are meaningfully related to household decision-making.

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

[23]Elicited MPCs are measured using a standard lottery question 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 over the following twelve months, rather than save?* Households' responses range from 0 (saving all) to 100 (spending all).

Table 6: 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 - *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

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 substantially attenuated 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 key avenue for 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, S. Crossman, and V. Sancibrián (2024). Estimating flexible income processes from subjective expectations data: evidence from india and colombia. Technical report, National Bureau of Economic Research.

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., 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.

Dominitz, J. (1998). Earnings expectations, revisions, and realizations. *Review of Economics and statistics 80*(3), 374–388.

Dominitz, J. (2001). Estimation of income expectations models using expectations and realization data. *Journal of Econometrics 102*(2), 165–195.

Dominitz, J. and C. F. Manski (1997). Using expectations data to study subjective income expectations. *Journal of the American statistical Association 92*(439), 855–867.

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.

Guiso, L., T. Jappelli, and D. Terlizzese (1996). Income risk, borrowing constraints, and portfolio choice. *The American economic review*, 158–172.

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.

Kosar, G. and D. Melcangi (2025). Subjective uncertainty and the marginal propensity to consume. Technical report, CESifo Working Paper.

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.

Manski, C. F. and E. Tamer (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica 70*(2), 519–546.

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

Pistaferri, L. (2001). Superior information, income shocks, and the permanent income hypothesis. *Review of Economics and Statistics 83*(3), 465–476.

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

Stoltenberg, C. and A. Uhlendorff (2022). Consumption choices and earnings expectations: Empirical evidence and structural estimation. Technical report, Tinbergen Institute Discussion Paper.

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 "What Income Risk Do Households Perceive? Subjective Expectations and Income Dynamics"

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 8,822 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. 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

$$\eta_i \sim N(0, \sigma_\eta^2), \quad \eta_i \perp u_{it+1} \text{ for all } t, \quad \eta_i \perp \mathbf{X}_{it} \text{ for all } t. \tag{B2}$$

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:

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

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:

$$(\widehat{\alpha}(\tau_l), \widehat{\beta}(\tau_l)) = \arg \min_{a_k(\tau), \beta(\tau)} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho_{\tau_l} \left(y_{it+1} - \sum_{k=0}^{K} a_k(\tau)\, \psi_k(y_{it}) - \mathbf{X}'_{it}\beta(\tau) \right).$$

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 approximating cdf and pdf. Following Arellano and Bonhomme (2016), the conditional distribution of next-period income can be represented through the conditional quantile function. To simplify the discussion, we first ignore the conditioning on state variables $\mathbf{X}_{it}$. Recall that the conditional quantile function of y_{it+1} given y_{it} is

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

where $a_k(\tau)$ are piecewise-linear interpolating splines in τ.

Given this specification, we can write the approximating CDF and PDF implied by the quantile function. Let

$$A_{it}(\ell) \equiv Q_y(\tau_\ell \mid y_{it})$$

denote the value of the conditional quantile function at knot τ_ℓ. Then the approximating CDF $F_{it}(y)$ and PDF $f_{it}(y)$ are defined by three regions.

For the left tail, if $y < A_{it}(1)$,

$$F_{it}(y) = \tau_1 \exp\left\{\lambda_1\big(y - A_{it}(1)\big)\right\},$$

and therefore

$$f_{it}(y) = \lambda_1\tau_1 \exp\left\{\lambda_1\big(y - A_{it}(1)\big)\right\}.$$

For the interior segments, if $A_{it}(\ell) \leqslant y < A_{it}(\ell+1)$ for some $\ell = 1, \ldots, L-1$,

$$F_{it}(y) = \tau_\ell + \frac{\tau_{\ell+1} - \tau_\ell}{A_{it}(\ell+1) - A_{it}(\ell)}\big(y - A_{it}(\ell)\big).$$

This implies that the pdf is constant within each segment:

$$f_{it}(y) = \frac{\tau_{\ell+1} - \tau_\ell}{A_{it}(\ell+1) - A_{it}(\ell)}.$$

For the right tail, if $y > A_{it}(L)$,

$$F_{it}(y) = 1 - (1 - \tau_L)\exp\left\{-\lambda_L\big(y - A_{it}(L)\big)\right\},$$

and therefore

$$f_{it}(y) = \lambda_L(1-\tau_L)\exp\left\{-\lambda_L\big(y - A_{it}(L)\big)\right\}.$$

When the state variables $\mathbf{X}_{it}$ are included, the conditional quantile function becomes

$$Q_y(\tau \mid y_{it}, \mathbf{X}_{it}) = \sum_{k=0}^{K} a_k(\tau)\psi_k(y_{it}) + \mathbf{X}_{it}'\beta(\tau), \tag{C9}$$

and we write

$$A_{it}(\ell;\theta) \equiv Q_y(\tau_\ell \mid y_{it}, \mathbf{X}_{it};\theta).$$

The corresponding CDF and PDF are denoted by $F_{it}(y;\theta)$ and $f_{it}(y;\theta)$.

Observed data likelihood. Let $I_{j,it} \equiv (\kappa_{j-1,it}, \kappa_{j,it}]$ denote bin j for respondent i at time t. Given the nonlinear income process, the model-implied bin probabilities are

$$p_{j,it}(\theta) = \Pr(y_{it+1} \in I_{j,it} \mid y_{it}, \mathbf{X}_{it};\theta) = F_{it}(\kappa_{j,it};\theta) - F_{it}(\kappa_{j-1,it};\theta). \tag{C10}$$

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

$$\ell_N(\theta) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\log\frac{M!}{\prod_{j=1}^{J} d_{jit}!} + \sum_{j=1}^{J} d_{jit}\log\big(p_{j,it}(\theta)\big)\right]. \tag{C11}$$

C.2 Coarse information

Maximizing the observed-data log-likelihood (C11) directly is computationally demanding because of the constraints needed to guarantee monotonicity and because the resulting objective function is highly nonlinear. Hence, to estimate the parameters of the nonlinear income process, we appeal to a stochastic pseudo-EM algorithm in the spirit of Arellano and Bonhomme (2016). In particular, we interpret the M different points as possible realizations from the conditional distribution $f_{it}(y_{it+1}|y_{it};\theta)$, and estimate its parameters via quantile regressions. The estimated parameters of the conditional distribution are those that are used to evaluate the model-implied bin probabilities (C10), and the observed data log-likelihood (C11).

In practice, the algorithm works in the following manner. For a given initial parameter estimate $\theta^{(0)}$, we iterate on $s = 0, 1, 2, \ldots$ until convergence using the following two steps.

E-step. For each (i,t) and for each bin j, draw d_{jit} latent values from the posterior density

$$f_{it}(y \mid y \in I_{j,it}; \theta^{(s)}).$$

In practice, for each point in bin j, the steps to draw the latent value are:

1. Compute

$$a_{j,it}^{(s)} = F_{it}(\kappa_{j-1,it}; \theta^{(s)}), \qquad b_{j,it}^{(s)} = F_{it}(\kappa_{j,it}; \theta^{(s)}),$$

   where $F_{it}(\cdot; \theta^{(s)})$ is the approximating CDF defined in Subsection **??**.

2. Draw

$$u_{itjm}^{(s)} \sim U\Big[a_{j,it}^{(s)}, b_{j,it}^{(s)}\Big].$$

3. Set

$$y_{itjm}^{(s)} = Q_y\Big(u_{itjm}^{(s)} \mid y_{it}, \mathbf{X}_{it}; \theta^{(s)}\Big).$$

We finally collect the pooled augmented sample

$$\{y_{itjm}^{(s)}, y_{it}, \mathbf{X}_{it}\}_{i,t,j,m},$$

with exactly M simulated draws for each (i,t).

M-step. In the M-step, we update both the parameters of the quantile function and the tail parameters. 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

$$(\widehat{a}(\tau_\ell), \widehat{\beta}(\tau_\ell)) = \arg\min_{a(\tau_\ell), \beta(\tau_\ell)} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{J}\sum_{m=1}^{d_{jit}} \rho_{\tau_\ell}\left(y_{itjm}^{(s)} - \sum_{k=0}^{K} a_k(\tau_\ell)\psi_k(y_{it}) - \mathbf{X}_{it}'\beta(\tau_\ell)\right), \tag{C12}$$

where $\rho_\tau(\cdot)$ is the standard check function. After estimating the quantile regressions, we apply the ABB rearrangement step to enforce monotonicity of $\tau \mapsto Q_y(\tau \mid y_{it}, \mathbf{X}_{it})$.

We update the tail parameters using the exponential tail restrictions. Let

$$A_{it}^{(s+1)}(1) = Q_y(\tau_1 \mid y_{it}, \mathbf{X}_{it}; \theta^{(s+1)}), \qquad A_{it}^{(s+1)}(L) = Q_y(\tau_L \mid y_{it}, \mathbf{X}_{it}; \theta^{(s+1)})$$

denote the updated boundary quantiles. Index the pooled simulated draws by $q \equiv (i, t, j, m)$ and write

$$y_q \equiv y_{itjm}^{(s)}, \qquad A_q(1) \equiv A_{it}^{(s+1)}(1), \qquad A_q(L) \equiv A_{it}^{(s+1)}(L).$$

Define

$$N_{\text{left}} = \#\{q : y_q < A_q(1)\}, \qquad N_{\text{right}} = \#\{q : y_q > A_q(L)\}.$$

Then the MLE updates for the exponential tail rates are

$$\widehat{\lambda}_1^{(s+1)} = \left[\frac{1}{N_{\text{left}}} \sum_{q:\, y_q < A_q(1)} \big(A_q(1) - y_q\big) \right]^{-1}, \qquad \widehat{\lambda}_L^{(s+1)} = \left[\frac{1}{N_{\text{right}}} \sum_{q:\, y_q > A_q(L)} \big(y_q - A_q(L)\big) \right]^{-1}. \tag{C13}$$

C.3 Hurdle model in the nonlinear process

This subsection extends the hurdle reporting model to the nonlinear income process defined in subsections C.1 and C.2. The underlying income process is unchanged. For a given parameter vector θ, which includes the parameters of the conditional quantile function and the exponential-tail parameters, the model-implied probability of bin j is

$$p_{j,it}(\theta) = F_{it}(\kappa_{j,it}; \theta) - F_{it}(\kappa_{j-1,it}; \theta),$$

where $F_{it}(\cdot; \theta)$ is the approximating conditional CDF implied by the nonlinear conditional quantile function.

The additional component is a reporting equation that allows some respondents to place all M probability points in the middle bin deterministically. Let j^* denote the middle bin and define

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

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

Reporting equation. We use the non-centered random-effects representation

$$b_i = \sigma_b v_i, \qquad v_i \sim N(0, 1),$$

and specify

$$\pi_{it}(v_i;\gamma,\sigma_b) \equiv \Pr(R_{it}=1 \mid Z_{it}, v_i) = \Lambda\left(\gamma_0 + Z_{it}'\gamma_1 + \sigma_b v_i\right). \tag{C14}$$

The non-centered representation is equivalent to assuming $b_i \sim N(0,\sigma_b^2)$, but is more convenient for estimation because the Gauss–Hermite nodes associated with v_i remain fixed when σ_b is updated.

Under the non-deterministic component, the probability that all M points are allocated to the middle bin is

$$B_{it}(\theta) \equiv \left[p_{j^*,it}(\theta)\right]^M.$$

Conditional on v_i, the per-period observed-data likelihood contribution is

$$L_{it}^f(v_i;\Theta) = \begin{cases} \pi_{it}(v_i;\gamma,\sigma_b) + \left[1-\pi_{it}(v_i;\gamma,\sigma_b)\right] B_{it}(\theta), & S_{it}=1, \\ \left[1-\pi_{it}(v_i;\gamma,\sigma_b)\right] \dfrac{M!}{\prod_{j=1}^J d_{j,it}!} \displaystyle\prod_{j=1}^J \left[p_{j,it}(\theta)\right]^{d_{j,it}}, & S_{it}=0, \end{cases} \tag{C15}$$

where

$$\Theta = (\theta,\gamma,\sigma_b).$$

The individual likelihood contribution is

$$L_i^f(\Theta) = \int \prod_{t=1}^{T_i} L_{it}^f(v;\Theta)\phi(v)\,dv, \tag{C16}$$

where $\phi(\cdot)$ denotes the standard Normal density.

Stochastic pseudo-EM algorithm. As in Subsection C.2, direct maximization of the observed-data likelihood is computationally demanding. We therefore use a stochastic pseudo-EM algorithm that augments the observed allocations with latent income draws and the latent reporting regimes.

Let $\{v_r,\varpi_r\}_{r=1}^R$ denote fixed Gauss–Hermite nodes and weights, normalized to integrate with respect to the standard Normal distribution. Starting from an initial parameter vector

$$\Theta^{(0)} = \left(\theta^{(0)},\gamma^{(0)},\sigma_b^{(0)}\right),$$

iterate over $s=0,1,2,\ldots$.

E-step

Posterior quadrature-node probabilities. For each household i and node r, calculate

$$\omega_{ir}^{(s)} = \frac{\varpi_r \prod_{t=1}^{T_i} L_{it}^f\left(v_r; \Theta^{(s)}\right)}{\sum_{q=1}^{R} \varpi_q \prod_{t=1}^{T_i} L_{it}^f\left(v_q; \Theta^{(s)}\right)}, \qquad \sum_{r=1}^{R} \omega_{ir}^{(s)} = 1. \tag{C17}$$

The reporting probability inside $L_{it}^f(v_r; \Theta^{(s)})$ is evaluated as

$$\pi_{it}^{(s)}(v_r) = \Lambda\left(\gamma_0^{(s)} + Z_{it}'\gamma_1^{(s)} + \sigma_b^{(s)} v_r\right).$$

Posterior focal-regime probabilities. For an observation with $S_{it} = 1$, the response may have arisen either from the deterministic focal regime or from the non-focal multinomial component. Conditional on quadrature node v_r, define

$$\chi_{itr}^{(s)} \equiv \Pr\left(R_{it} = 1 \mid S_{it} = 1, v_r, \Theta^{(s)}\right) = \frac{\pi_{it}^{(s)}(v_r)}{\pi_{it}^{(s)}(v_r) + \left[1 - \pi_{it}^{(s)}(v_r)\right] B_{it}\left(\theta^{(s)}\right)}. \tag{C18}$$

For observations with $S_{it} = 0$, the deterministic regime is impossible, and we set

$$\chi_{itr}^{(s)} = 0.$$

Integrating over the posterior distribution of the reporting random effect, the posterior probability that observation (i, t) belongs to the deterministic focal regime is

$$\bar{r}_{it}^{(s)} = \sum_{r=1}^{R} \omega_{ir}^{(s)} \chi_{itr}^{(s)}. \tag{C19}$$

The posterior probability that the response was generated by the non-focal income distribution is therefore

$$w_{it}^{(s)} = 1 - \bar{r}_{it}^{(s)} = \sum_{r=1}^{R} \omega_{ir}^{(s)} \left[1 - \chi_{itr}^{(s)}\right]. \tag{C20}$$

Notice that $w_{it}^{(s)} = 1$ whenever $S_{it} = 0$.

Simulation of latent incomes. Conditional on the current income-process parameter vector $\theta^{(s)}$, simulate the latent income draws as in Subsection C.2. For each household-period pair (i, t) and each bin j, calculate

$$a_{j,it}^{(s)} = F_{it}\left(\kappa_{j-1,it}; \theta^{(s)}\right), \qquad b_{j,it}^{(s)} = F_{it}\left(\kappa_{j,it}; \theta^{(s)}\right).$$

For each of the $d_{j,it}$ points allocated to bin j, draw

$$u_{itjm}^{(s)} \sim U\left[a_{j,it}^{(s)}, b_{j,it}^{(s)}\right], \qquad m = 1, \ldots, d_{j,it},$$

and set

$$y_{itjm}^{(s)} = Q_y\left(u_{itjm}^{(s)} \mid y_{it}, X_{it}; \theta^{(s)}\right).$$

For observations with $S_{it} = 1$, all M simulated values lie in the middle bin, but their contribution to the income-process update is weighted by $w_{it}^{(s)}$. The portion attributed to the deterministic focal regime does not contribute to estimation of the income process.

M-step

Update of the reporting equation. Holding the posterior quantities $\omega_{ir}^{(s)}$ and $\chi_{itr}^{(s)}$ fixed, update the parameters of the reporting equation by solving

$$\left(\widehat{\gamma}^{(s+1)}, \widehat{\sigma}_b^{(s+1)}\right) = \arg\max_{\gamma,\, \sigma_b > 0} \sum_{i=1}^{N} \sum_{r=1}^{R} \omega_{ir}^{(s)} \sum_{t=1}^{T_i} \Bigg\{ \chi_{itr}^{(s)} \log \Lambda\left(\gamma_0 + Z_{it}'\gamma_1 + \sigma_b v_r\right) + \left[1 - \chi_{itr}^{(s)}\right] \log\left[1 - \Lambda\left(\gamma_0 + Z_{it}'\gamma_1 + \sigma_b v_r\right)\right] \Bigg\}. \tag{C21}$$

Unlike the previous formulation of (C21), this objective depends explicitly on σ_b. In the numerical implementation, it is convenient to optimize over

$$\alpha_b = \log \sigma_b, \qquad \sigma_b = \exp(\alpha_b),$$

so that the positivity restriction is automatically satisfied.

Update of the conditional quantile function. Let

$$z_{it} = \left(\psi_0(y_{it}), \ldots, \psi_K(y_{it}), X_{it}'\right)'.$$

For each spline knot τ_ℓ, update the corresponding coefficient vector $\delta(\tau_\ell)$ using weighted quantile regression:

$$\widehat{\delta}^{(s+1)}(\tau_\ell) = \arg\min_{\delta} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{j=1}^{J} \sum_{m=1}^{d_{j,it}} w_{it}^{(s)} \rho_{\tau_\ell}\left(y_{itjm}^{(s)} - z_{it}'\delta\right), \qquad \ell = 1, \ldots, L. \tag{C22}$$

Here,

$$\delta(\tau_\ell) = \left(a_0(\tau_\ell), \ldots, a_K(\tau_\ell), \beta(\tau_\ell)'\right)'.$$

After estimating the quantile regressions, apply the ABB rearrangement step to enforce monotonicity of

$$\tau \mapsto Q_y(\tau \mid y_{it}, X_{it}).$$

Update of the exponential-tail parameters. Let

$$A_{it}^{(s+1)}(1) = Q_y\left(\tau_1 \mid y_{it}, X_{it}; \theta^{(s+1)}\right)$$

and

$$A_{it}^{(s+1)}(L) = Q_y\left(\tau_L \mid y_{it}, X_{it}; \theta^{(s+1)}\right)$$

denote the updated boundary quantiles.

Index the simulated draws by

$$q = (i, t, j, m),$$

and write

$$y_q = y_{itjm}^{(s)}, \qquad w_q^{(s)} = w_{it}^{(s)},$$

$$A_q^{(s+1)}(1) = A_{it}^{(s+1)}(1), \qquad A_q^{(s+1)}(L) = A_{it}^{(s+1)}(L).$$

The weighted maximum-likelihood update for the left-tail rate is

$$\widehat{\lambda}_1^{(s+1)} = \frac{\sum_q w_q^{(s)} \mathbf{1}\left\{y_q < A_q^{(s+1)}(1)\right\}}{\sum_q w_q^{(s)} \left[A_q^{(s+1)}(1) - y_q\right] \mathbf{1}\left\{y_q < A_q^{(s+1)}(1)\right\}}. \tag{C23}$$

Similarly, the weighted maximum-likelihood update for the right-tail rate is

$$\widehat{\lambda}_L^{(s+1)} = \frac{\sum_q w_q^{(s)} \mathbf{1}\left\{y_q > A_q^{(s+1)}(L)\right\}}{\sum_q w_q^{(s)} \left[y_q - A_q^{(s+1)}(L)\right] \mathbf{1}\left\{y_q > A_q^{(s+1)}(L)\right\}}. \tag{C24}$$

Thus, both the number of effective tail observations and their total exceedances are weighted by the posterior non-focal probabilities.

If there are no simulated observations with positive weight in one of the tails, the corresponding tail parameter is left at its previous value for that iteration.

Convergence. At the end of each iteration, evaluate the quadrature approximation to the observed-data log-likelihood:

$$\ell^f(\Theta) = \sum_{i=1}^{N} \log \left[\sum_{r=1}^{R} \varpi_r \prod_{t=1}^{T_i} L_{it}^f(v_r; \Theta) \right]. \tag{C25}$$

The algorithm is stopped when

$$\frac{\left|\ell^f\left(\Theta^{(s+1)}\right) - \ell^f\left(\Theta^{(s)}\right)\right|}{1 + \left|\ell^f\left(\Theta^{(s)}\right)\right|} < \varepsilon$$

and the change in the parameter vector is also below a pre-specified tolerance.

Because the latent incomes are simulated and the conditional quantile function is updated through weighted quantile regressions followed by rearrangement, this procedure is a stochastic pseudo-EM rather than an exact EM algorithm. The observed-data log-likelihood therefore need not increase at every single iteration. In practice, the simulation noise can be reduced by using common random numbers across nearby iterations or by averaging the parameter estimates and likelihood over several final iterations.

Interpretation. The hurdle component does not alter the household's underlying non-linear perceived income process. It changes the estimation by separating observations that are likely to reflect deterministic focal-point reporting from observations generated by the subjective income distribution. Non-focal responses receive full weight, whereas middle-bin bunching observations contribute to the income-process estimation in proportion to their posterior probability of having arisen from the non-focal component.

D AR(1) process with lots of heterogeneity

In this appendix, we consider the following income process:

$$y_{it+1} = \rho y_{it} + \mathbf{X}_{it}'\beta + \sigma_i u_{it+1}, \qquad u_{it+1} \sim N(0,1). \tag{D14}$$

where σ_i 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 $e_{it}(\rho, \beta) = y_{it+1} - \rho y_{it} - \mathbf{X}'_{it}\beta$ 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:

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

The marginal likelihood for an individual i is

$$L_i(\theta) = \int_0^\infty \prod_{t=1}^{T} f(e_{it}|\sigma_i; \beta, \rho) g(\sigma_i; \delta) d\sigma. \tag{D15}$$

The sample log-likelihood is

$$\ell(\theta) = \sum_{i=1}^{N} \log L_i(\theta). \tag{D16}$$

To estimate this model, we appeal to maximum likelihood methods. We estimate this model assuming that the distribution of σ_i is lognormal[24] 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 households answer the income expectations question with income process (D14) in mind, the true probabilities $p^*_{j,it}(\sigma_i)$ are calculated by the households as:

$$p^*_{j,it}(\sigma_i) = \Phi\left(\frac{\kappa_{j,it} - \rho y_{it} - \mathbf{X}'_{it}\beta}{\sigma_i}\right) - \Phi\left(\frac{\kappa_{j-1,it} - \rho y_{it} - \mathbf{X}'_{it}\beta}{\sigma_i}\right). \tag{D17}$$

[24] 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.

The integrated likelihood contribution for household i is:

$$L_i(\theta) = \int \prod_{t=1}^{T} \left[\frac{M!}{\prod_{j=1}^{J} d_{jit}!} \prod_{j=1}^{J} \left(p^*_{j,it}(\sigma_i) \right)^{d_{jit}} \right] f_\sigma(\sigma_i; \delta)\, d\sigma_i. \tag{D18}$$

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 under-estimate the risk that they face as opposed to what is observed in the data. In particular, the persistence parameter is 0.7558 for the realized income process, and 1.0000 for the subjective income process. Meanwhile, on average, income volatility is 0.4189 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.

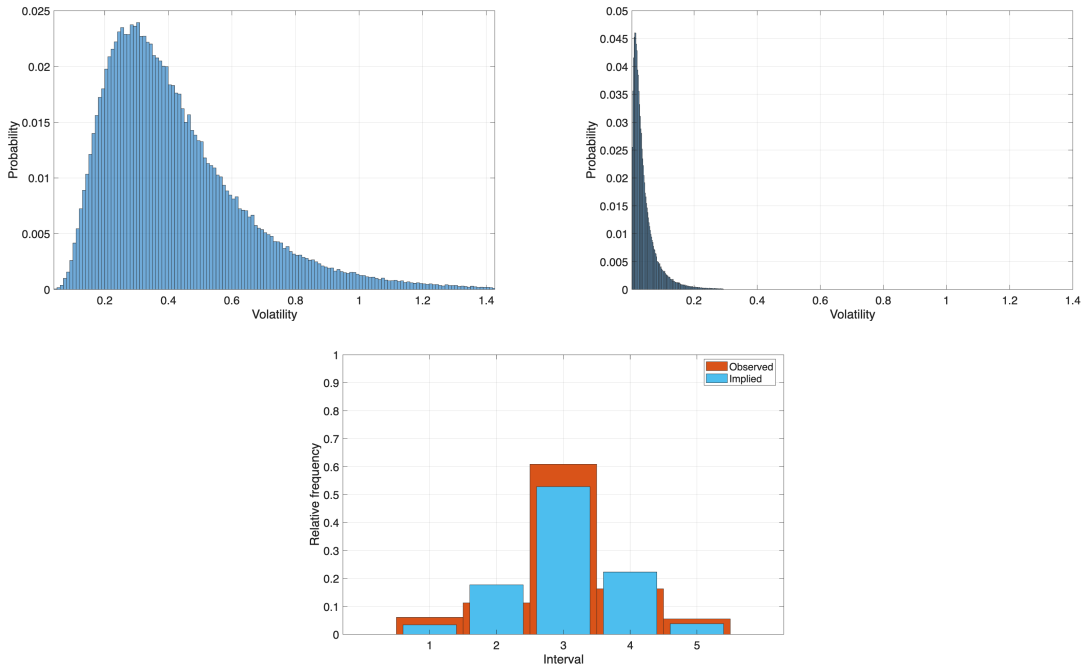
The bottom panel of Figure D.1 shows the model fit of the heterogeneous AR(1) process without a hurdle. We find that the model fit is good, with a RMSE of 0.0522, which is a remarkable improvement over the model with the state-dependent process. We do observe that the model underestimates the bunching 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	0.7558	-
	[0.7283,0.7797]	-
	Mean	Standard Deviation
Volatility	0.4189	0.2294
	[0.4133,0.4393]	[0.2200,0.2599]
Panel B: Subjective income process		
Persistence	1.000	-
	(*)	-
	Mean	Standard Deviation
Volatility	0.0426	0.0471
	[0.0378, 0.0527]	[0.0355,0.0475]

Notes: This table shows estimates of persistence, and the average and the standard deviation of income volatility, μ_σ and ς_σ, 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.

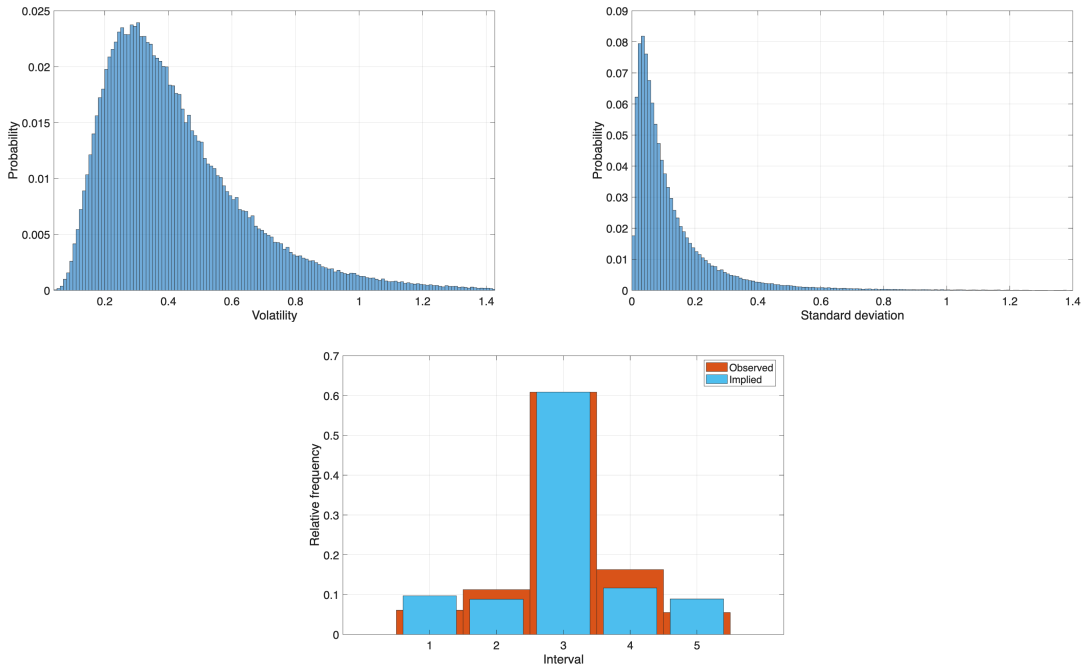
In sum, the results we obtain confirm the insight of the state-dependent AR(1) model; households over-estimate the persistence of their household income, and underestimate the dispersion of income shocks.

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

Panel A: Realized income process		
Persistence	0.7558	-
	[0.7283,0.7797]	-
	Mean	Standard Deviation
Volatility	0.4189	0.2294
	[0.4133,0.4393]	[0.2200,0.2599]
Panel B: Subjective income process		
Persistence	0.9862	-
	[0.9499, 0.9914]	-
	Mean	Standard Deviation
Volatility	0.1176	0.1073
	[0.0840,1.0426]	[0.0824,3.5679]

Notes: This table shows estimates of persistence, and the average and the standard deviation of income volatility, μ_σ and ς_σ, 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 Identification

In this appendix, we discuss identification of the conditional distribution of future log income perceived by household i at time t, given the state

$$W_{it} = (y_{it}, X_{it}),$$

which we denote by

$$F_y(c \mid W_{it}) = \Pr(y_{i,t+1} \leqslant c \mid W_{it}).$$

The observed data are

$$\{d_{it}, W_{it}\}_{t=1}^{T_i}, \qquad d_{it} = (d_{1,it}, \ldots, d_{J,it})', \qquad \sum_{j=1}^{J} d_{j,it} = M.$$

Let the income-growth cutoffs satisfy

$$-\infty = h_0 < h_1 < \cdots < h_{J-1} < h_J = +\infty.$$

The corresponding income-level cutoffs and bins are

$$\kappa_{j,it} = y_{it} + h_j, \qquad I_{j,it} = (\kappa_{j-1,it}, \kappa_{j,it}] = (y_{it} + h_{j-1}, y_{it} + h_j].$$

The allocation vector d_{it} provides interval-censored information about the perceived distribution of $y_{i,t+1}$. It reveals the probability assigned to each interval, but does not reveal how probability mass is distributed within an interval.

Assumption A1 (Correct reporting of subjective bin probabilities)**.** For every $j = 1, \ldots, J$,

$$E\left[\frac{d_{j,it}}{M} \,\middle|\, W_{it} = (y, X)\right] = F_y(y + h_j \mid y, X) - F_y(y + h_{j-1} \mid y, X).$$

Assumption A1 can be interpreted either as exact reporting of subjective probabilities or as unbiased reporting at the population level. The hurdle model considered below relaxes this assumption by allowing respondents to enter a deterministic focal-point regime.

E.1 Nonparametric content of the data

Define the cumulative reported probability

$$\Pi_j(y, X) \equiv E\left[\sum_{k=1}^{j} \frac{d_{k,it}}{M} \,\middle|\, W_{it} = (y, X)\right], \qquad j = 1, \ldots, J-1,$$

and set $\Pi_0(y, X) = 0$ and $\Pi_J(y, X) = 1$.

Proposition E.1 (Sharp partial identification)**.** *Under Assumption A1, the conditional CDF is identified at each finite bin cutoff:*

$$F_y(y + h_j \mid y, X) = \Pi_j(y, X), \qquad j = 1, \ldots, J-1.$$

Without additional restrictions on the within-bin shape of the distribution, the conditional quantile is set identified. In particular, if

$$\Pi_{j-1}(y, X) < \tau \leqslant \Pi_j(y, X),$$

then

$$Q_y(\tau \mid y, X) \in (y + h_{j-1}, y + h_j].$$

This identified set is sharp.

Proof. By Assumption A1,

$$E\left[\frac{d_{k,it}}{M}\,\middle|\, W_{it}=(y,X)\right]=F_y(y+h_k\mid y,X)-F_y(y+h_{k-1}\mid y,X).$$

Summing from $k=1$ to $k=j$ gives

$$\begin{aligned}\Pi_j(y,X)&=\sum_{k=1}^{j}E\left[\frac{d_{k,it}}{M}\,\middle|\, W_{it}=(y,X)\right]\\&=\sum_{k=1}^{j}\{F_y(y+h_k\mid y,X)-F_y(y+h_{k-1}\mid y,X)\}\\&=F_y(y+h_j\mid y,X)-F_y(y+h_0\mid y,X).\end{aligned}$$

Because $h_0=-\infty$,

$$F_y(y+h_0\mid y,X)=0,$$

and hence

$$F_y(y+h_j\mid y,X)=\Pi_j(y,X).$$

Now fix $\tau\in(0,1)$ and suppose

$$\Pi_{j-1}(y,X)<\tau\leqslant\Pi_j(y,X).$$

Since the CDF is nondecreasing,

$$F_y(y+h_{j-1}\mid y,X)<\tau\leqslant F_y(y+h_j\mid y,X).$$

Therefore, under the generalized-inverse definition

$$Q_y(\tau\mid y,X)=\inf\{c:F_y(c\mid y,X)\geqslant\tau\},$$

we have

$$Q_y(\tau\mid y,X)\in(y+h_{j-1},y+h_j].$$

To establish sharpness, fix any

$$q\in(y+h_{j-1},y+h_j].$$

Construct an alternative CDF $\widetilde{F}_y(\cdot\mid y,X)$ that agrees with $F_y(\cdot\mid y,X)$ at every cutoff $y+h_\ell$, but reallocates the probability mass inside bin j so that

$$\widetilde{F}_y(q-\mid y,X)<\tau\leqslant\widetilde{F}_y(q\mid y,X).$$

This reallocation leaves all bin probabilities unchanged and is therefore observationally equivalent to the original distribution. Its τ-quantile is q. Since every point in the interval can be attained in this way, the bound is sharp. □

Remark E.1 (Finite lower cutoffs). If the first cutoff h_0 is finite rather than $-\infty$, the correct identity is

$$F_y(y + h_j \mid y, X) = F_y(y + h_0 \mid y, X) + \Pi_j(y, X).$$

Thus, the CDF at the cutoffs is identified only if the probability below the first cutoff is known or separately modeled.

Remark E.2 (Additional panel waves). Additional waves identify the CDF at the cutoffs associated with the states observed in those waves. In the absence of restrictions linking the within-bin shape across states, however, additional waves do not tighten the within-bin quantile bounds at a fixed state (y, X).

E.2 State-dependent AR(1) model

We next impose the state-dependent Gaussian income process

$$y_{i,t+1} = \rho(X_{it})y_{it} + X_{it}'\beta + \eta_i + \sigma(X_{it})u_{i,t+1}, \qquad u_{i,t+1} \sim N(0,1),$$

where

$$\rho(X) = \Lambda(\rho_0 + X'\rho_1), \qquad \sigma(X) = \exp(\sigma_0 + X'\sigma_1).$$

Assumption A2 (Normal location–scale structure and random-effects exogeneity)**.** The household effect satisfies

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

and is independent of the full observed state history:

$$\eta_i \perp (W_{i1}, \ldots, W_{iT_i}).$$

Conditional on η_i and the observed state history, the income innovations are independent over time.

Assumption A3 (Support and rank)**.** For every relevant value of the discrete components of X, the conditional support of y contains a nondegenerate interval. Moreover, the relevant covariate design matrices have full column rank.

After integrating over η_i,

$$y_{i,t+1} \mid W_{it} \sim N\left(\mu(W_{it}), \sigma_e^2(X_{it})\right),$$

where

$$\mu(y, X) = \rho(X)y + X'\beta$$

and

$$\sigma_e^2(X) = \sigma^2(X) + \sigma_\eta^2.$$

E.2.1 Identification of the marginal distribution

Corollary E.1 (Identification of the marginal perceived distribution)**.** *Suppose Assumptions A1, A2, and A3 hold.*

Fix a state (y, X). If there are two distinct finite cutoffs $h_j \neq h_k$ such that

$$0 < \Pi_j(y, X) < 1, \qquad 0 < \Pi_k(y, X) < 1,$$

then $\mu(y, X)$ and $\sigma_e(X)$ are identified. Consequently, the full marginal CDF

$$F_y(\cdot \mid y, X)$$

is identified within the Normal location–scale model.

If this condition holds over a sufficiently rich set of states, then (ρ_0, ρ_1, β) are identified.

Proof. By Proposition E.1,

$$\Pi_j(y, X) = F_y(y + h_j \mid y, X).$$

Under the marginal Normal model,

$$\Pi_j(y, X) = \Phi\left(\frac{y + h_j - \mu(y, X)}{\sigma_e(X)}\right).$$

For an interior cutoff probability, define

$$z_j(y, X) = \Phi^{-1}\left(\Pi_j(y, X)\right).$$

Then

$$z_j(y, X) = \frac{y + h_j - \mu(y, X)}{\sigma_e(X)}.$$

For two distinct cutoffs,

$$z_k(y, X) - z_j(y, X) = \frac{h_k - h_j}{\sigma_e(X)}.$$

Because $h_k \neq h_j$ and $\sigma_e(X) > 0$,

$$\sigma_e(X) = \frac{h_k - h_j}{z_k(y, X) - z_j(y, X)}$$

is uniquely identified. The conditional mean is then

$$\mu(y, X) = y + h_j - \sigma_e(X) z_j(y, X).$$

Hence

$$F_y(c \mid y, X) = \Phi\left(\frac{c - \mu(y, X)}{\sigma_e(X)}\right)$$

is identified for every $c \in \mathbb{R}$.

It remains to identify the coefficients of the conditional mean. For a fixed X and two distinct income values $y_1 \neq y_2$ in the conditional support,

$$\rho(X) = \frac{\mu(y_2, X) - \mu(y_1, X)}{y_2 - y_1}.$$

Thus, $\rho(X)$ is identified over the support of X. Since Λ is known and strictly increasing,

$$\Lambda^{-1}(\rho(X)) = \rho_0 + X'\rho_1.$$

Full-rank variation in $(1, X')'$ identifies (ρ_0, ρ_1). Finally,

$$\mu(y, X) - \rho(X) y = X'\beta,$$

so full-rank variation in X identifies β. □

Remark E.3 (Interiority). The inverse Normal CDF is finite only on $(0, 1)$. Thus, two strictly interior cumulative probabilities are sufficient for pointwise identification of both location and scale. If both objects are unrestricted at a given state, a single interior cutpoint is not sufficient.

Remark E.4 (Interpretation under misspecification). If the true perceived distribution is not Gaussian, the Normal specification selects the location and scale parameters that best reconcile the model with the observed cutpoint probabilities. The implied within-bin distribution is then supplied by Normality rather than identified nonparametrically from the survey responses.

E.2.2 Separating within-household volatility and fixed heterogeneity

The cross-sectional distribution identifies

$$\sigma_e^2(X) = \sigma^2(X) + \sigma_\eta^2,$$

but does not separately identify $\sigma^2(X)$ and σ_η^2. Panel dependence provides the additional information required for this decomposition.

Define the cumulative allocation

$$C_{j,it} = \sum_{\ell=1}^{j} \frac{d_{\ell,it}}{M}.$$

Assumption A4 (Conditional independence across waves)**.** For $t \neq s$, the probability-point draws underlying d_{it} and d_{is} are independent conditional on

$$\eta_i \quad \text{and} \quad (W_{i1}, \ldots, W_{iT_i}).$$

Assumption A5 (Panel interiority)**.** For at least one pair $t \neq s$ and finite cutoffs h_j, h_k,

$$0 < \Pi_j(W_{it}) < 1, \qquad 0 < \Pi_k(W_{is}) < 1.$$

The within-type variances satisfy $\sigma^2(X) > 0$ over the relevant support.

Corollary E.2 (Identification of the variance components)**.** *Suppose Assumptions A1–A5 hold and $T_i \geqslant 2$ for a positive fraction of households. Then the joint panel distribution identifies*

$$\sigma_\eta^2, \quad \sigma^2(X), \quad (\sigma_0, \sigma_1).$$

Together with Corollary E.1, this identifies

$$(\rho_0, \rho_1, \beta, \sigma_0, \sigma_1, \sigma_\eta^2).$$

Proof. Fix $t \neq s$ and cumulative cutoffs j and k. Conditional on η_i,

$$E[C_{j,it} \mid W_{it}, \eta_i] = \Phi\left(\frac{y_{it} + h_j - \mu(W_{it}) - \eta_i}{\sigma(X_{it})}\right).$$

By Assumption A4,

$$\begin{aligned}&\operatorname{Cov}\left(C_{j,it}, C_{k,is} \mid W_{it}, W_{is}\right)\\ &\quad = \operatorname{Cov}_\eta\left(E[C_{j,it} \mid W_{it}, \eta_i], E[C_{k,is} \mid W_{is}, \eta_i] \mid W_{it}, W_{is}\right).\end{aligned}$$

Equivalently, consider two latent draws

$$Y_t^* = \mu(W_{it}) + \eta_i + \sigma(X_{it})u_t,$$

and

$$Y_s^* = \mu(W_{is}) + \eta_i + \sigma(X_{is})u_s,$$

where u_t and u_s are independent standard Normal variables. Conditional on (W_{it}, W_{is}), the pair (Y_t^*, Y_s^*) is bivariate Normal, with marginal variances

$$\sigma_e^2(X_{it}) \quad \text{and} \quad \sigma_e^2(X_{is}),$$

and correlation

$$r_{ts} = \frac{\sigma_\eta^2}{\sigma_e(X_{it})\sigma_e(X_{is})}.$$

Let

$$z_{j,it} = \Phi^{-1}(\Pi_j(W_{it})), \qquad z_{k,is} = \Phi^{-1}(\Pi_k(W_{is})).$$

It follows that

$$\begin{aligned}&\operatorname{Cov}\left(C_{j,it}, C_{k,is} \mid W_{it}, W_{is}\right)\\ &\quad = \Phi_2(z_{j,it}, z_{k,is}; r_{ts}) - \Phi(z_{j,it})\Phi(z_{k,is}),\end{aligned}$$

where $\Phi_2(a, b; r)$ is the standard bivariate Normal CDF with correlation r.

The left-hand side is identified from the joint panel distribution. The marginal quantities $z_{j,it}$, $z_{k,is}$, $\sigma_e(X_{it})$, and $\sigma_e(X_{is})$ are identified from Corollary E.1.

For finite a and b, the bivariate Normal CDF is strictly increasing in its correlation parameter:

$$\frac{\partial}{\partial r}\Phi_2(a, b; r) = \phi_2(a, b; r) > 0.$$

Therefore, the observed cross-wave covariance uniquely identifies r_{ts}. It then identifies

$$\sigma_\eta^2 = r_{ts}\,\sigma_e(X_{it})\sigma_e(X_{is}).$$

Having identified σ_η^2, the within-household innovation variance is

$$\sigma^2(X) = \sigma_e^2(X) - \sigma_\eta^2.$$

Finally,

$$\log \sigma^2(X) = 2\sigma_0 + 2X'\sigma_1.$$

Full-rank variation in $(1, X')'$ identifies (σ_0, σ_1). □

Remark E.5 (Role of the panel). The panel identifies σ_η^2 through persistent cross-wave dependence in allocations made by the same household. It does not identify the shape of the distribution within the survey bins; that shape continues to be supplied by the Normal location–scale assumption.

Remark E.6 (Overidentification). One pair of waves and one pair of interior cumulative cutoffs is sufficient in principle. Additional waves and cutoffs provide overidentifying covariance restrictions that can be used to assess the random-effects and conditional-independence assumptions.

E.3 Nonlinear income process

Let the conditional quantile function belong to a finite-dimensional sieve class

$$\mathcal{Q}_{L,K} = \{Q_y(\tau \mid y, X; \theta) : \theta \in \Theta_{L,K}\}.$$

Assumption A6 (Support)**.** The support $\mathcal{W}$ of (y, X) contains sufficient variation to identify the basis coefficients used in the quantile specification.

Assumption A7 (Correctly specified finite sieve)**.** For fixed sieve dimensions (L, K), there exists $\theta_0 \in \Theta_{L,K}$ such that

$$Q_y^0(\tau \mid y, X) = Q_y(\tau \mid y, X; \theta_0).$$

For every (y, X), the function

$$\tau \mapsto Q_y(\tau \mid y, X; \theta)$$

is continuous and strictly increasing.

For $j = 1, \ldots, J-1$, define

$$m_j(y, X; \theta) = F_y(y + h_j \mid y, X; \theta).$$

Assumption A8 (Cutpoint injectivity)**.** For fixed (L, K), the map

$$\theta \mapsto \{m_j(y, X; \theta) : j = 1, \ldots, J-1,\ (y, X) \in \mathcal{W}\}$$

is injective over $\Theta_{L,K}$. That is, if

$$m_j(y, X; \theta) = m_j(y, X; \theta')$$

for every j and almost every $(y, X) \in \mathcal{W}$, then

$$\theta = \theta'.$$

Corollary E.3 (Identification of the finite-dimensional nonlinear model)**.** *Suppose Assumptions A1, A6, A7, and A8 hold. For fixed sieve dimensions (L, K), θ_0 is point identified from the cross-sectional distribution of (d_{it}, W_{it}). Consequently, the conditional quantile function and the associated conditional CDF are identified within the maintained sieve model.*

Proof. By Proposition E.1, the data identify

$$\Pi_j(y, X) = F_y(y + h_j \mid y, X)$$

for every $j = 1, \ldots, J-1$ and almost every $(y, X) \in \mathcal{W}$.

Under the quantile specification and strict monotonicity,

$$F_y(c \mid y, X; \theta) = \sup\{\tau \in (0, 1) : Q_y(\tau \mid y, X; \theta) \leqslant c\}.$$

Therefore,

$$m_j(y, X; \theta_0) = \Pi_j(y, X)$$

for every cutoff and state.

Suppose that another parameter $\theta' \in \Theta_{L,K}$ generates the same observed cutpoint probabilities. Then

$$m_j(y, X; \theta') = m_j(y, X; \theta_0)$$

for all j and almost every $(y, X) \in \mathcal{W}$. By Assumption A8,

$$\theta' = \theta_0.$$

Thus, θ_0 is point identified. Since the parameter uniquely determines the quantile function, the full conditional quantile function and its associated CDF are identified within the maintained sieve class. □

Remark E.7 (Approximation versus identification). The preceding result concerns identification at a fixed sieve dimension under correct specification. If the true quantile function does not belong exactly to $\mathcal{Q}_{L,K}$, the fixed-dimensional model identifies a pseudo-true sieve approximation rather than the exact true distribution. Approximation results as $(L, K) \to \infty$ are therefore conceptually distinct from fixed-dimensional point identification.

Remark E.8 (Local identification). A sufficient condition for local identification is that the Jacobian of the collection of cutpoint probabilities with respect to θ has full column rank at θ_0. Global point identification requires the stronger injectivity condition in Assumption A8.

E.4 Hurdle model

Let

$$e^* = (0, \ldots, 0, M, 0, \ldots, 0)'$$

denote the allocation that places all M points in the middle bin. The observed bunching indicator is

$$S_{it} = 1\{d_{it} = e^*\}.$$

With probability

$$\pi(Z_{it}, b_i; \gamma) = \Lambda(\gamma_0 + Z_{it}'\gamma_1 + b_i),$$

the household enters the deterministic focal regime and reports e^*. Otherwise, it reports according to the underlying income-process model.

Let

$$f_\theta(d \mid W)$$

denote the probability of allocation d under the non-focal income-process component, after integrating over any income-process random effect. Define

$$A_\theta(W) = f_\theta(e^* \mid W).$$

Also define the marginal focal probability

$$q_{\gamma,\sigma_b}(Z) = \int \pi(Z, b; \gamma)\, dG_{\sigma_b}(b), \qquad b \sim N(0, \sigma_b^2).$$

Assumption A9 (Exclusion and random-effects exogeneity)**.** The reporting shifter V_{it} is contained in Z_{it} but not in W_{it}. Conditional on W_{it} and year effects, V_{it} does not affect the perceived income distribution.

Moreover,

$$(b_i, \eta_i) \perp \{Z_{it}, W_{it}\}_{t=1}^{T_i}, \qquad b_i \perp \eta_i.$$

Assumption A10 (Positivity and nondegeneracy)**.** Over the relevant support,

$$0 < q_{\gamma,\sigma_b}(Z) < 1$$

and

$$A_\theta(W) < 1.$$

Assumption A11 (Identification from off-focal allocations)**.** The map from θ to the conditional distribution of non-focal allocation patterns,

$$\left\{ \frac{f_\theta(d \mid W)}{1 - A_\theta(W)} : d \neq e^*,\ W \in \mathcal{W} \right\},$$

is injective. For models with an income-process random effect, this condition applies to the corresponding joint cross-wave distribution of off-focal allocations.

Assumption A12 (Identification of the reporting equation)**.** The map

$$(\gamma, \sigma_b^2) \mapsto \{q_{\gamma,\sigma_b}(Z) : Z \in \mathcal{Z}\}$$

is injective over the maintained parameter space. The excluded variable V has nondegenerate conditional variation and shifts the focal-point probability.

Proposition E.2 (Identification of the hurdle model)**.** *Suppose Assumptions A9–A12 hold and the underlying income-process component satisfies the relevant identification conditions in Corollary E.2 or Corollary E.3. Then*

$$\theta, \quad \gamma, \quad \sigma_b^2$$

are jointly point identified.

Proof. Under Assumption A9, the observed conditional probability of an allocation d is

$$g(d \mid Z, W) = \begin{cases} q_{\gamma,\sigma_b}(Z) + [1 - q_{\gamma,\sigma_b}(Z)]\, A_\theta(W), & d = e^*, \\ [1 - q_{\gamma,\sigma_b}(Z)]\, f_\theta(d \mid W), & d \neq e^*. \end{cases}$$

Consider the distribution conditional on observing a non-focal allocation, $S = 0$. For $d \neq e^*$,

$$\begin{aligned} \Pr(d \mid S = 0, Z, W) &= \frac{[1 - q_{\gamma,\sigma_b}(Z)] f_\theta(d \mid W)}{[1 - q_{\gamma,\sigma_b}(Z)][1 - A_\theta(W)]} \\ &= \frac{f_\theta(d \mid W)}{1 - A_\theta(W)}. \end{aligned}$$

Thus, conditioning on $S = 0$ eliminates the reporting probability. The resulting distribution is independent of the excluded reporting shifter V. By Assumption A11, this distribution identifies θ.

Once θ is identified, $A_\theta(W)$ is known. Moreover,

$$\Pr(S = 0 \mid Z, W) = [1 - q_{\gamma,\sigma_b}(Z)]\,[1 - A_\theta(W)]\,.$$

Since $A_\theta(W) < 1$,

$$q_{\gamma,\sigma_b}(Z) = 1 - \frac{\Pr(S = 0 \mid Z, W)}{1 - A_\theta(W)}$$

is identified over the support of Z. Assumption A12 then identifies (γ, σ_b^2). Hence all income-process and reporting-process parameters are jointly point identified. □

Remark E.9 (Role of the exclusion restriction). The exclusion restriction provides variation in the mixture probability while holding the income-process component fixed. For a continuous excluded variable,

$$\frac{\partial}{\partial V} \Pr(S = 1 \mid Z, W) = \frac{\partial q_{\gamma,\sigma_b}(Z)}{\partial V}\,[1 - A_\theta(W)]\,.$$

For a discrete excluded variable, the analogous expression is a finite contrast. Because $A_\theta(W) < 1$, a genuine shift in the focal probability changes observed middle-bin bunching. This establishes relevance of the exclusion. Point identification additionally requires the component and reporting-equation injectivity conditions stated above.

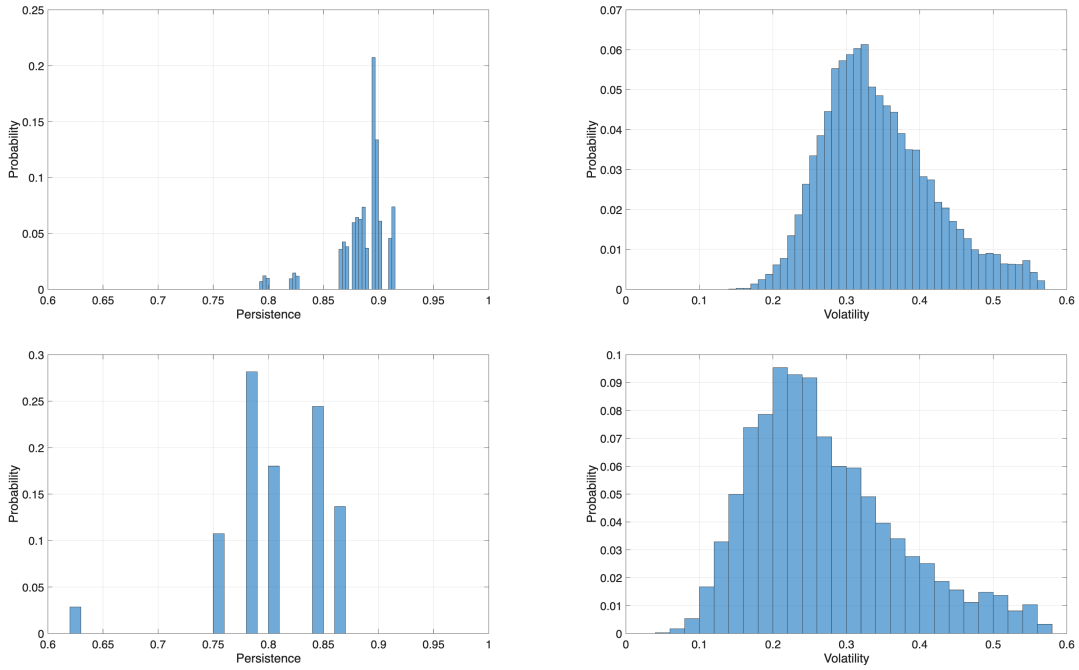
F Results from the ECV

As we discussed in section 5, one potential reason for the differences between the realized and the subjective income processes is the time intervals between the observations in the realized and in the subjective expectations data. In particular, the estimation using realized income data uses household income data observed in three year intervals, while the subjective income expectations questions relate to annual changes in household income. To verify whether these differences explain the disparity in the two answers, we use data from the *Encuesta de Condiciones de Vida* (ECV), or the Living Conditions Survey, that is conducted by the Spanish Statistical Institute (INE), and estimate both one-year and three-year intervals. To ensure that we have comparable samples, we restrict the data to households with heads from 25 to 65 years of age, who have at least one active household member working, and whose main source of income does not come from pensions. We are left with a balanced panel of 3,863 households who we observe for four consecutive periods from 2021-2024.

Figure F.1 shows the results of the estimation of the state-dependent AR(1) process for data observed in one- and three-year intervals. As the results indicate, on average persistence decreases from approximately 0.85 for the one-year income process to 0.72 for the three-year income process, while average dispersion has a slight increase. These changes though, are nowhere near the differences observed between realized and subjective income expectations data.

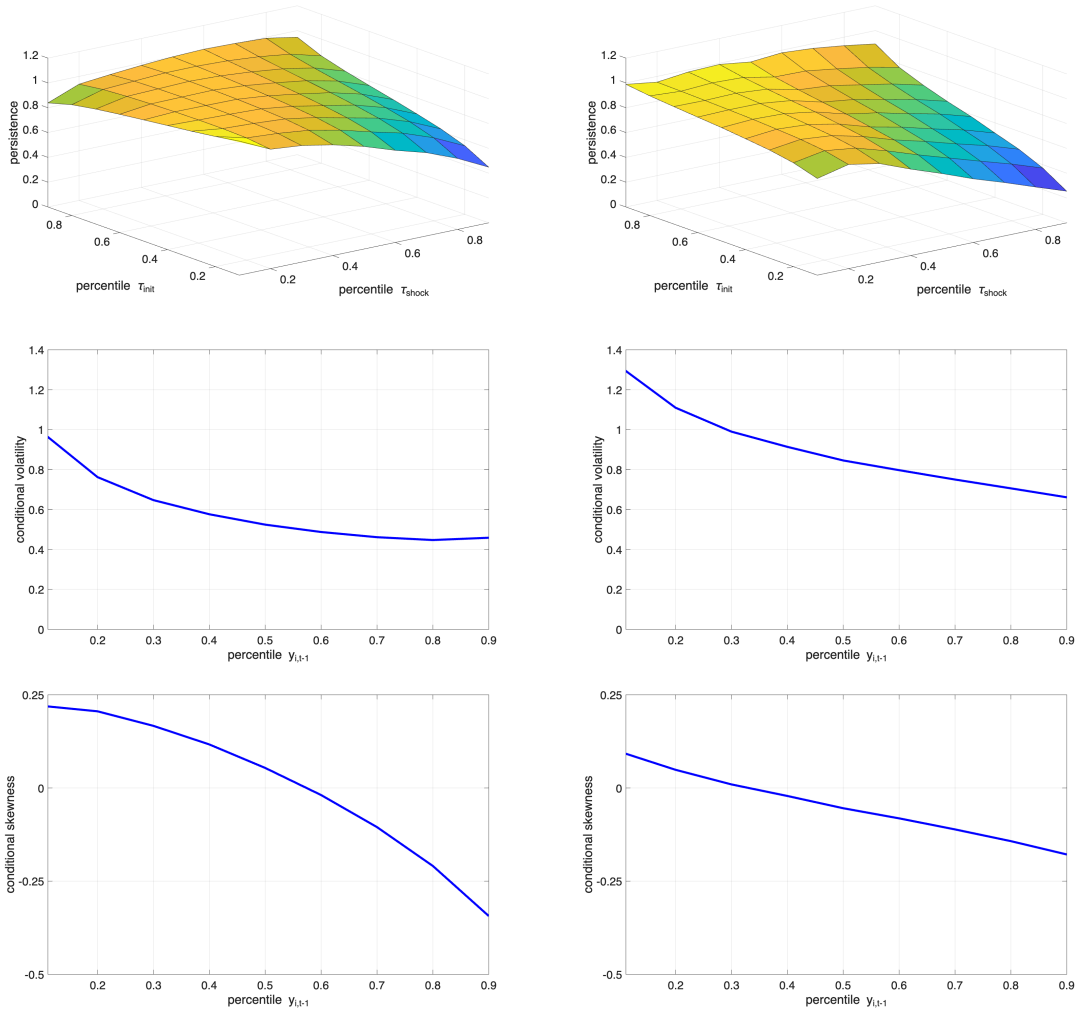
Figure F.2, meanwhile, shows the results of the estimation of the nonlinear income process for data observed in one-year and three-year intervals. We observe similar patterns with respect to persistence and variance, while for skewness, we find that the estimates are wider in magnitude.

Figure F.1: Distributions of persistence and dispersion, state-dependent AR(1) process, ECV



Notes: These plots show the distribution of persistence (left panel) and dispersion (right panel) for the state-dependent AR(1) process using one-year (top row) and three-year (bottom row) income data.

Figure F.2: Distributions of persistence and dispersion, state-dependent AR(1) process, ECV



Notes: These plots show estimates of persistence (top), conditional variance (middle), and conditional skewness (bottom) for the nonlinear income process using one-year (left panel) and three-year data (right panel).

G Model with tilted probabilities

G.1 Model specification and estimation

An alternative to the hurdle model for bunching behavior is a model wherein households report their probabilities via a tilted distribution, in which the tilts are due to some characteristics of the expectations question that could be particularly salient for household respondents. We discuss the model set-up and the estimation for the state-dependent AR(1) model as an illustration, and discuss the results for the three models that we estimate.

Suppose that households use the state-dependent AR(1) model to answer the subjective income expectations question:

$$y_{it+1} = \rho(\mathbf{X}_{it})y_{it} + \mathbf{X}'_{it}\beta + \eta_i + \sigma(\mathbf{X}_{it})u_{it+1}, \qquad u_{it+1} \sim N(0,1). \tag{G19}$$

However, different households might respond to the income expectations question in different ways. In particular, a fraction ζ_{it} of households will draw their responses from a Multinomial distribution with probabilities defined from the underlying income process $p^*_{j,it}(\eta_i)$, while the rest draw their points from a tilted probability distribution $\omega_{j,it}(\eta_i, \lambda)$, which is equal to:

$$\omega_{j,it}(\eta_i, \lambda) = \frac{p^*_{j,it}(\eta_i)w_j(\lambda)}{\sum_{k=1}^{J} p^*_{k,it}(\eta_i)w_k(\lambda)} \tag{G20}$$

where $w_j(\lambda) = \exp\{\lambda_1(j - j^\star) + \lambda_2(j - j^\star)^2\}$. The tilted probabilities are governed by two parameters λ_1, which is related to directional bias (over or under pessimism) and λ_2, which is related to the shape of the distribution (centering or spreading). This particular form of the probability tilts is inspired by the idea that some events could be particularly salient to households, which leads them to focus on less bins of the income expectations question as a result.

Let $T_{it} = 1$ denote truthful probability reporting. We model the probability of truthful reporting as

$$\zeta_{it} \equiv \Pr(T_{it} = 1 \mid \mathbf{Z}_{it}) = \Lambda(\mathbf{Z}'_{it}\gamma), \tag{G21}$$

where T_{it} is a latent indicator that is equal to one when households report their responses via the true probabilities.

Conditional on η_i, the likelihood for individual i at time t is:

$$L_{it}(\eta_i) = \zeta_{it} L_{it}^U(\eta_i) + (1 - \zeta_{it}) L_{it}^B(\eta_i),$$

where

$$L_{it}^U(\eta_i) = \prod_{j=1}^{J} \left[p_{j,it}^*(\eta_i)\right]^{d_{jit}}$$

is the likelihood if she reports truthfully, and

$$L_{it}^B(\eta_i) = \prod_{j=1}^{J} \left[\omega_{j,it}(\eta_i, \lambda)\right]^{d_{jit}}$$

is the likelihood if she reports with the probabilistic tilt.

The conditional panel likelihood for an individual i is:

$$L_i(\eta_i) = \prod_{t=1}^{T} L_{it}(\eta_i), \tag{G22}$$

To complete the model, we need a specification for the unobserved heterogeneity. The integrated likelihood, hence, is:

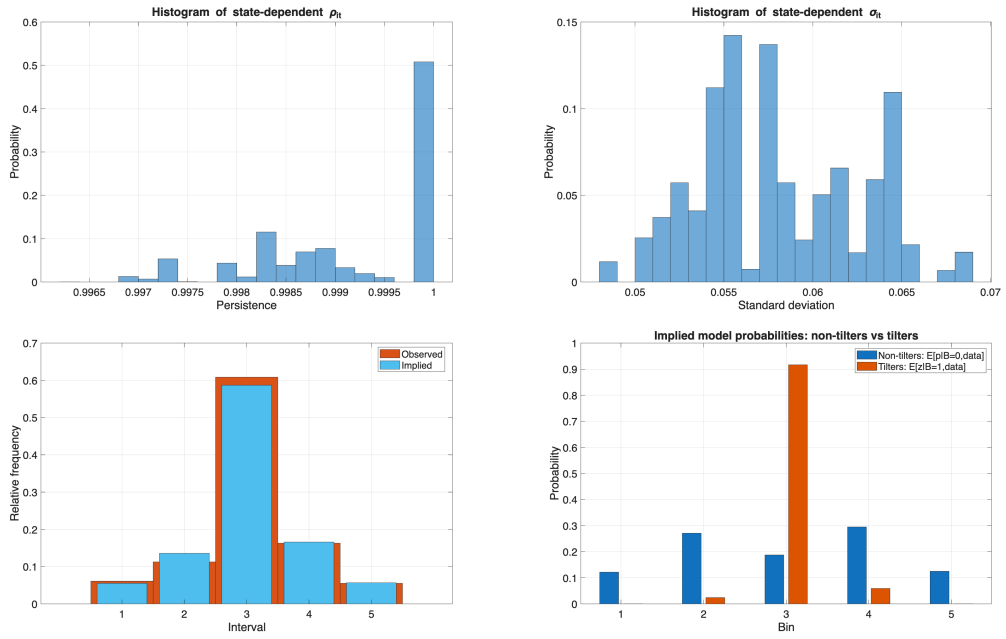
$$L_i = \int \prod_{t=1}^{T} L_{it}(\eta_i) f(\eta_i) d\eta_i. \tag{G23}$$

We estimate this model via the EM algorithm, and conduct non-parametric bootstrap to compute standard errors.

G.2 Results

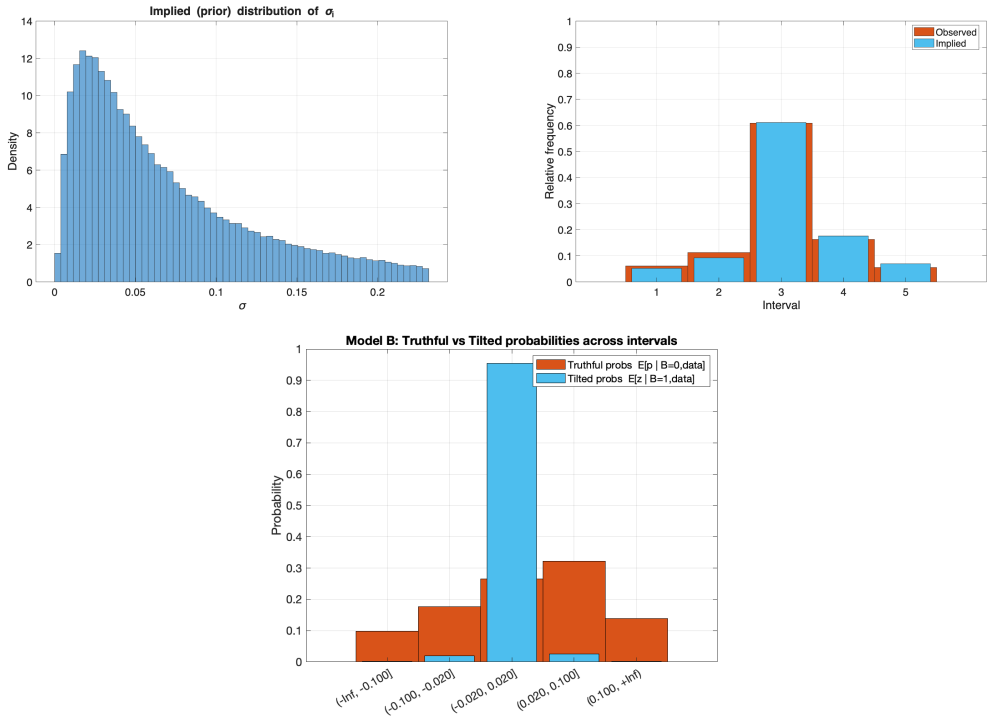
Figures G.1-G.3 present the results from the estimated models with tilted probabilities. As the results indicate, the empirical results are qualitatively the same as in the hurdle model. Moreover, the central message of the empirical analysis remains: households overestimate the persistence of their income and underestimate the risk that they face. It seems to be the case as well, that the primary mode of cognitive bias is heaping all of the points into the middle bin. These results indicate that the hurdle model with bunching in the middle is already sufficient enough to capture most of the potential cognitive biases present in the data.

Figure G.1: Results of the state-dependent AR(1) process, model with tilted probabilities



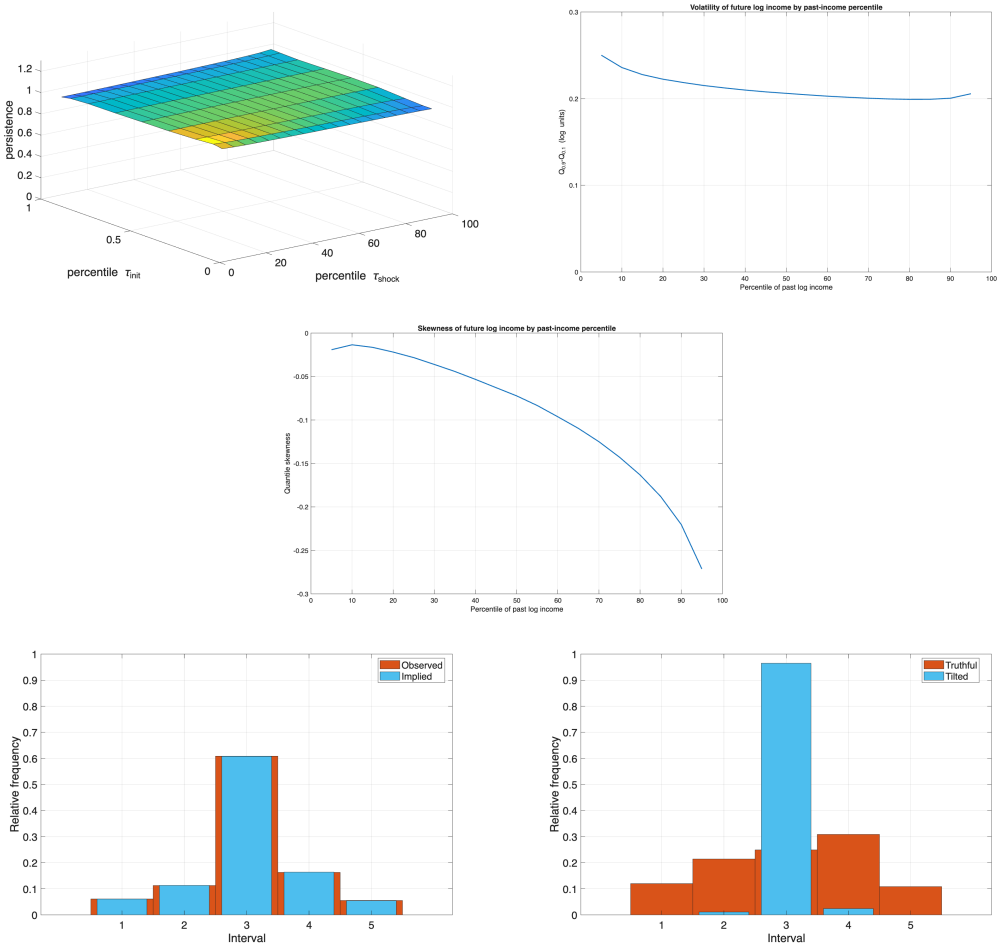
Notes: These plots show estimates of persistence (top left), volatility (top right), the comparison between the observed and model-implied bin probabilities (bottom left), and the decomposition of bin probabilities between truthful and tilted households (bottom right).

Figure G.2: Results of the heterogeneous AR(1) process, model with tilted probabilities



Notes: These plots show estimates of volatility (top left), the comparison between the observed and model-implied bin probabilities (top right), and the decomposition of bin probabilities between truthful and tilted households (bottom).

Figure G.3: Results of the nonlinear process, model with tilted probabilities



Notes: These plots show estimates of persistence (top left), conditional volatility (top right), skewness (middle), the comparison between the observed and model-implied bin probabilities (bottom left), and the decomposition of bin probabilities between truthful and tilted households (bottom right).

H Additional Results - Life Cycle Model

We report additional results from the life cycle model that complement the baseline exercises described in the main text. As discussed in the last paragraphs of Section 6, we consider two alternative specifications to assess the robustness of our findings to modeling choices regarding time aggregation and preference calibration.

First, we present results for the case in which the discount factor β is kept constant across income processes, rather than being adjusted to match the asset-to-income ratio observed in the data. This exercise allows us to isolate the implications of the estimated income processes for wealth accumulation and consumption insurance without imposing identical average saving rates across specifications. Second, we consider an alternative timing structure in which the model period is set to three years, consistent with the frequency of income observations in the EFF, and adjust the transition matrices of the income processes accordingly. In particular, for the processes estimated using subjective expectations data, we construct the three-year transition matrix as $Q_3 = Q_1^3$, where Q_1 is the one-year transition matrix obtained from the estimation.

The results (displayed in Figures H.1, H.2 and Table H.1) confirm the main qualitative findings of the baseline analysis. When β is held fixed, differences in wealth accumulation across income processes become even more pronounced. In particular, income processes estimated using realized data continue to generate substantially higher precautionary savings, as households respond to higher perceived income risk by accumulating wealth early in life. In contrast, the processes estimated using subjective expectations imply much flatter wealth profiles, reflecting the lower perceived risk embedded in those processes. These differences mirror the patterns documented in the main text and highlight the central role of perceived income risk in shaping life cycle saving behavior.

Similarly, the alternative specification with a three-year model period yields results in general qualitatively consistent with the baseline (see Figures H.3, H.4 and Table H.2). Adjusting the transition matrices to account for the longer horizon slightly reduces the persistence of income dynamics, but the key conclusions remain. Lower perceived risk leads households to save less and accumulate wealth more gradually. Nonlinear pro-

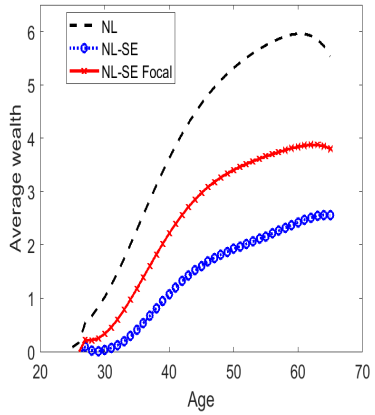


Figure H.1: Wealth Accumulation - Constant β

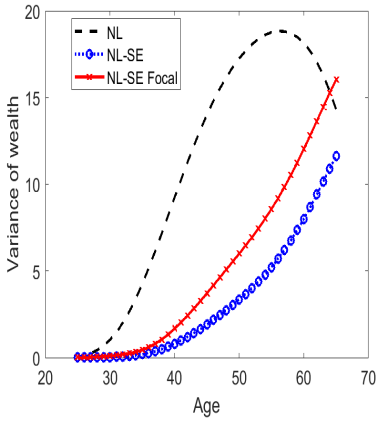


Figure H.2: Cross-section Variance of Wealth - Constant β

Income Process	ψ
NL	0.781
NL-SE	0.676
NL-SE Focal	0.756
NL-SE Focal / NL	0.609

Table H.1: Consumption Insurance - Constant β

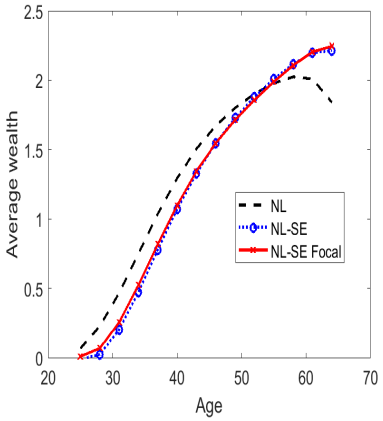


Figure H.3: Wealth Accumulation - 3 Year

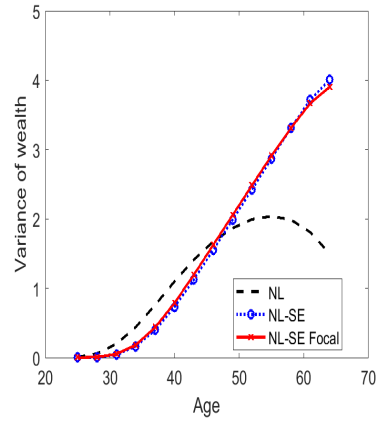


Figure H.4: Cross-section Variance of Wealth - 3 Year

Income Process	ψ
NL	0.70
NL-SE	0.56
NL-SE Focal	0.60
NL-SE Focal / NL	0.57

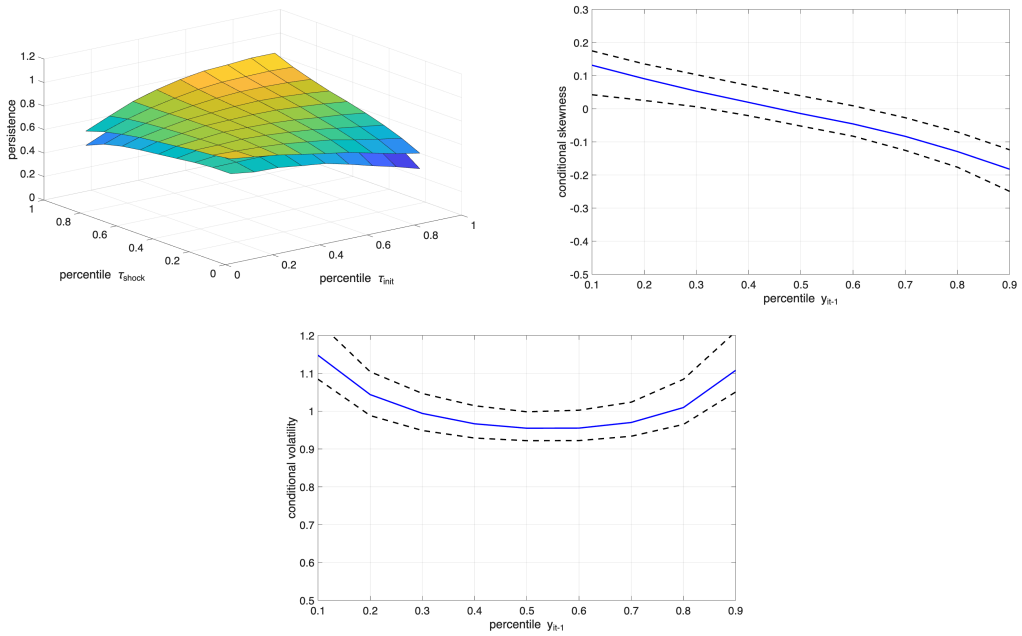
Table H.2: Consumption Insurance - 3 Year

cess estimated from realized income generates particularly strong precautionary saving motives among high-income households, it produces substantially greater wealth accumulation early in life and a higher cross-sectional dispersion of wealth than the processes recovered from subjective expectations.

Overall, these additional exercises reinforce the robustness of our main conclusions: differences in the inferred income processes—particularly in terms of perceived dispersion and nonlinearities—translate into economically meaningful differences in consumption insurance and wealth accumulation over the life cycle. These findings are not driven by specific assumptions regarding the calibration of preferences or the time aggregation of income dynamics.

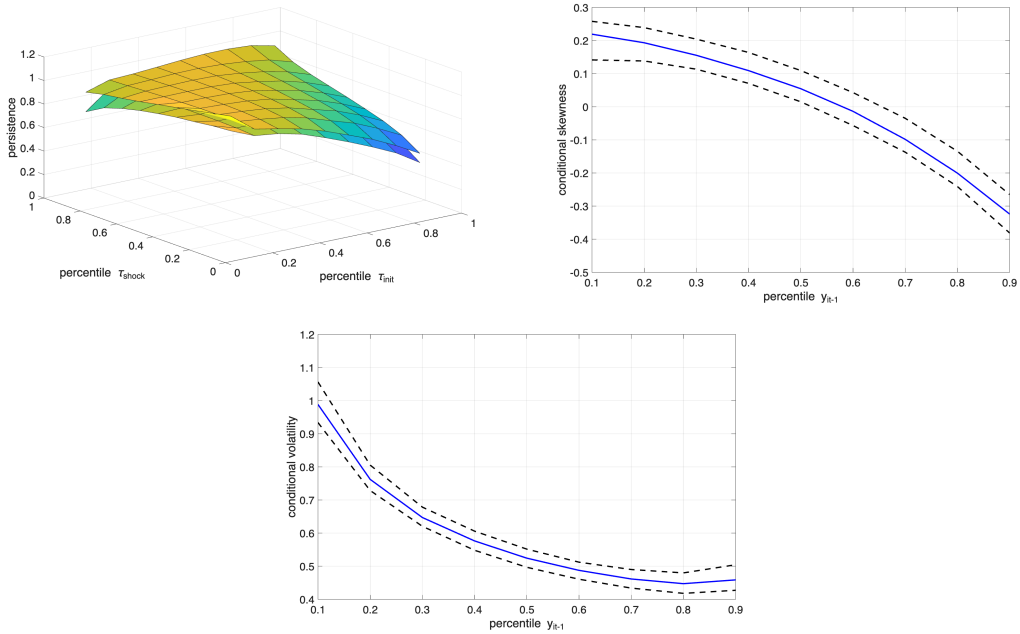
I Additional figures

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



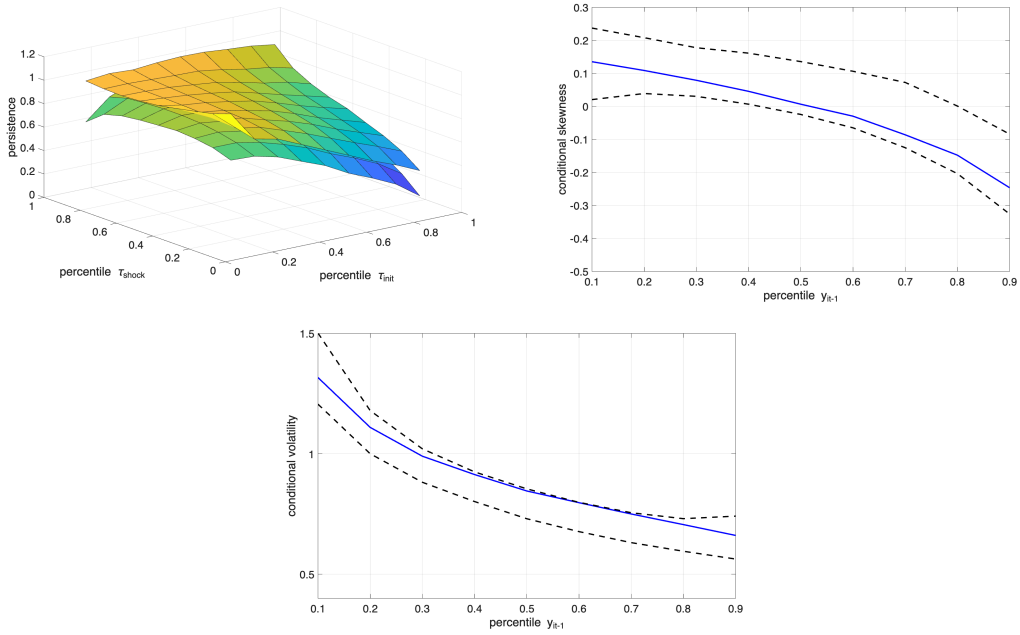
Notes: These plots show the 95% confidence bands associated with the nonlinear persistence measure (top left), conditional skewness (top right), and the conditional volatility (bottom), for the realized income process. These were constructed from 1,000 block bootstrap replications.

Figure I.2: Non-parametric bootstraps, nonlinear income process with realized data, ECV



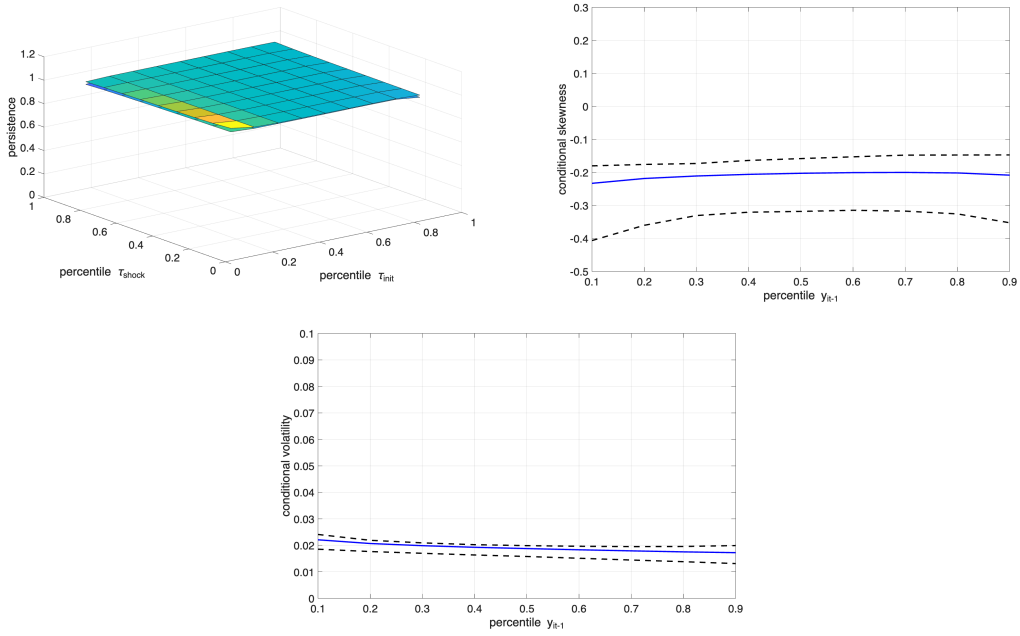
Notes: These plots show the 95% confidence bands associated with the nonlinear persistence measure (top left), conditional skewness (top right), and the conditional volatility (bottom), for the realized income process. These were constructed from 1,000 block bootstrap replications.

Figure I.3: Non-parametric bootstraps, nonlinear income process with realized data, ECV, three years



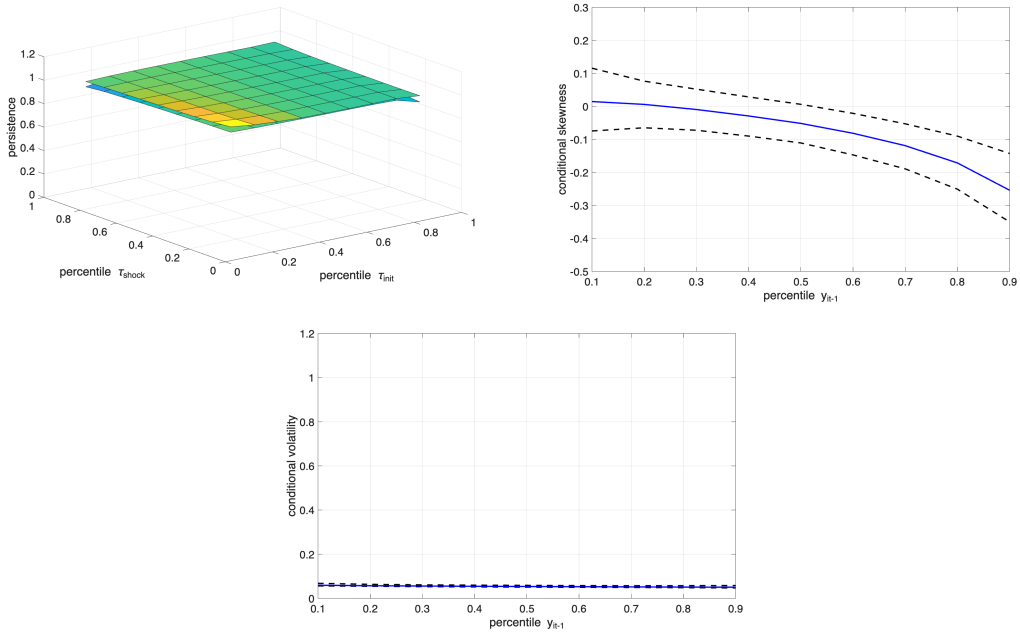
Notes: These plots show the 95% confidence bands associated with the nonlinear persistence measure (top left), conditional skewness (top right), and the conditional volatility (bottom), for the realized income process. These were constructed from 1,000 block bootstrap replications.

Figure I.4: 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 I.5: 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.