Supplemental Material

Estimating flexible income processes from subjective expectations data: evidence from India and Colombia

Manuel Arellano* Orazio Attanasio† Sam Crossman‡ Víctor Sancibrián§

July 2026

A Data

A.1 India – sample selection and validation

This section describes sample selection and validation of the subjective expectations data in detail. For the most part, it mirrors the analysis in Attanasio and Augsburg (2016, Section 1). However, we update some of the criteria used and review additional required sample-selection decisions.

The first two columns in Tables A.1 and A.2 can be understood as extended versions of Tables 2 and 3 reported in Attanasio and Augsburg (2016), respectively.

Table A.1 details basic sample selection steps and resulting sample sizes. For instance, we exclude observations with missing income or at least one reported probability. We also report the number of households for which either the elicited subjective lower or upper bound on future income is missing, although we do not exclude these from the final sample.[1] We also exclude from the sample some extreme reports of current household income under the category "implausible income", which are likely to correspond to survey measurement error.[2]

*CEMFI: arellano@cemfi.es
†Yale University and NBER: orazio.attanasio@yale.edu
‡UK Government Economic Service: samuelcrossman@gmail.com
§Bocconi University and IGIER: victor.sancibrian@unibocconi.it

[1] A total of 1,041 households were originally interviewed in the first round. We drop five respondents who were asked about monthly rather than yearly income. In the remaining rows, minor differences with respect to those reported in Attanasio and Augsburg (2016) are due to slightly different/updated criteria.

[2] In particular, these correspond to reports outside the range $[0.5 \times rmin, 2 \times rmax]$, where $rmin$ and $rmax$ are the reported lower and upper bound on subjective income distributions (respectively), and are replaced by r_A and r_C, respectively, if missing. Unlike in the Colombian data, as reported in Appendix A.2, using looser or stricter "cutoffs" does not lead to substantial changes in sample sizes.

The total number of unique households drops to 926 and 873 in the first and second rounds, respectively. Among these households, we study bunching of the reported probabilities at the 0%, 50% and 100% marks in Table A.2, and note substantial bunching at 50% for the midpoint (especially in the first round).

Keeping only households present in both rounds implies a balanced panel with 789 unique households. We further drop households who report at least one probability equal to zero or one or whose elicited subjective *cdf* is not *strictly* monotonic. This further reduces the final sample size to $N = 770$ households. We present robustness checks keeping those households in Appendix D.1.

TABLE A.1. India and Colombia: sample sizes and response rates (percentages)

	India		Colombia	
	Round 1	Round 2	Round 1	Round 2
Total number of obs.	**1036**	**947**	**10743**	**9463**
Missing income	0.19%	0.11%	14.17%	20.56%
Missing either Min or Max	1.06%	1.16%	12.05%	10.12%
Missing at least one prob.	2.12%	1.16%	12.67%	10.19%
Wrong — direction	0.29%	0.53%	0.34%	0.35%
Wrong — violation of monot.	0.87%	2.01%	3.26%	3.08%
Implausible thresholds	6.08%	0.21%	0.22%	0.22%
Implausible income	2.12%	4.33%	5.89%	4.12%
Available observations	**926** (89.38%)	**873** (92.19%)	**7262** (67.60%)	**6295** (66.52%)
Missing covariates	0%	0%	0.35%	0.26%
Balanced panel (robustness)	**789** (76.16%)		**4420** (41.14%)	
At least one prob. is 0 or 1	1.14%	1.39%	45.36%	32.44%
At least two prob. are equal	0.38%	0.25%	19.59%	13.57%
Balanced panel (final)	**770** (74.32%)		**2230** (20.76%)	

Note. "Wrong — direction" refers to households that in a given survey report $\Pr\left(y_{t+1} \geq r_C\right) > \Pr\left(y_{t+1} \geq r_B\right) > \Pr\left(y_{t+1} \geq r_A\right)$, among those with no missing probabilities. "Wrong — violation of monotonicity" refers to weak monotonicity violations. "Implausible thresholds" refers to households for which r_A, r_B or r_C is missing, among those who report no missing probabilities, households for which $r_B < r_A$ or $r_C < r_B$, and households with implausibly large interval differences between $r_B - r_A$ and $r_C - r_B$. "Implausible income" refers to households that report income outside $[0.5 \times rmin, 2 \times rmax]$, where $rmin$ and $rmax$ are replaced by r_A and r_C, respectively, if missing. Percentages are calculated relative to the number of available observations in each survey round and country (in bold).

TABLE A.2. India and Colombia: bunching (percentages)

	India		Colombia	
	Round 1	Round 2	Round 1	Round 2
	Threshold A (Lower)			
0%	0.65%	1.15%	14.33%	16.46%
50%	0.11%	2.29%	9.80%	7.94%
100%	0.22%	0%	1.14%	0.37%
	Threshold B (Midpoint)			
0%	0%	0%	2.77%	2.92%
50%	54.31%	16.27%	11.10%	12.11%
100%	0.22%	0%	3.77%	1.40%
	Threshold C (Higher)			
0%	0%	0%	1.17%	0.57%
50%	4.75%	11.69%	7.56%	8.33%
100%	0.32%	0.11%	15.68%	8.66%
Available observations	**926**	**873**	**7262**	**6295**

Note. The table shows the number of respondents who reported probabilities equal to 0, 0.5 and 1 in each survey round and country. See Table A.1 for a discussion on the number of available observations.

A.2 Colombia – sample selection and validation

We repeat the same analysis as for the Indian data (Appendix A.1) here. The relevant columns in Tables A.1 and A.2 summarize sample selection decisions and validation of the expectations data and bunching, respectively.

As shown in Table A.1, of the original 11,462 households, 10,743 were re-interviewed in 2003 and 9,463 in 2005/06. Household-income information was provided by 9,221 households in the first round and 7,517 in the second. The fraction of households with missing reported probabilities or extreme values of household income is also larger than in India.[3] Relative to the analysis for India, here we include an additional step for "missing covariates", where we exclude a few households for which covariates used in the main analysis (village, nature of income shocks and the proportion of regular income) were missing.

The remaining rows in Table A.1 and Table A.2 correspond to the validation exercise on the subjective expectations data, similar to the one performed for India in Appendix A.1 and in Attanasio and Augsburg (2016). The final balanced sample we use in the main analysis has $N = 2,230$ unique households, after excluding those who report probabilities equal to zero or

[3] In the Colombian data, using looser or stricter rules for "implausible income" (described in Table A.1) leads to substantially larger and smaller sample sizes, respectively. For instance, using an interval given by $[0.2 \times rmin, 5 \times rmax]$ allows us to keep approximately 200 additional unique households. Naturally, repeating the analysis with this rather permissive rule tends to exaggerate the features (nonlinear persistence, skewness, etc.) we study.

one or give answers that violate strict monotonicity of (subjective) *cdf*s. We report our main results keeping these observations in Appendix D.2. Finally, Tables A.3 and A.4 report differences in observable characteristics between households in the final dataset and those available but excluded after validation.

Since the subjective expectations data in Colombia have not been used before, we now elaborate on validation:

- *Logical response errors*. Table A.1 shows that in the first survey round 350 households provided answers that violated monotonicity and 37 provided responses that adhered to monotonicity but were "inverted", in that probabilities were non-decreasing, rather than non-increasing.[4] These figures imply a violation rate of around 4% among those who answered the probability questions, somewhat higher than in the Indian context where the logical response error was around 1%, but comparable to other studies. Dominitz and Manski (1997), for example, report violations for almost 5% of their sample, and almost twice that number when including respondents who were prompted (in that their responses would have initially been classified as logical response errors, but changed responses after having been prompted; such prompting was not allowed in either of the contexts considered here).

- *Bunching of percentages*. The relevant columns in Table A.2 report on the extent to which households bunch at the 0%, 50%, or 100% probability marks for the different thresholds. In the first round, there is substantial bunching at 0% for the lowest and 100% for the highest thresholds (around 14% of responses, compared to a negligible amount in India), which suggests some households might not have understood the prompts correctly or that the elicited expected income range is not accurate. There is also apparent bunching at 50% for the midpoint, a common feature with subjective probability data also present in the Indian data. There is a more muted presence of these issues in the second wave data.

Even though these data display a higher degree of logical response errors and bunching than in India, we conclude that the responses provided in the Colombian data conform for the most part to the basic probability laws, and seem to suggest substantial coherence and variability, in that most respondents appear to have understood the instructions and provided thoughtful responses.

[4] Recall that, as described in Figure 1, households were asked to report probabilities of the form $\Pr(y_{t+1} \geq r)$.

TABLE A.3. Colombia: covariate balance (wave 1)

	(0)		(1)		(0)-(1)
Variable	*N*	Mean	*N*	Mean	
Number of adults	3670	2.73	2230	2.69	0.03
Number of female adults	3670	1.39	2230	1.38	0.01
Number of kids	3670	3.11	2230	3.25	-0.15^{***}
Log income	3670	12.74	2230	12.73	0.02
Rural household	3661	0.46	2225	0.45	0.01
Household head:					
Age	3657	44.30	2228	43.60	0.70^{**}
Some primary education	3612	0.43	2213	0.45	−0.02
Some secondary education	3612	0.15	2213	0.15	−0.01
Primary source of income:					
Laborer/employee	3464	0.29	2107	0.27	0.02
Domestic employee	3464	0.06	2107	0.05	0.00
Day laborer	3464	0.21	2107	0.23	−0.02
Self-employment	3464	0.39	2107	0.38	0.00
Partner in farm/plot	3464	0.06	2107	0.07	−0.01
Proportion of regular income	3664	0.79	2230	0.79	0.00
Health shocks	3670	0.13	2230	0.12	0.01
Other shocks	3670	0.14	2230	0.13	0.01

Note. (1) refers to observations in the final sample and (0) to available observations excluded from the final sample (just before "Balanced panel (robustness)" in Table A.1). First wave only. Robust standard errors; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

TABLE A.4. Colombia: covariate balance (wave 2)

	(0)		(1)		(0)-(1)
Variable	*N*	Mean	*N*	Mean	
Number of adults	3118	2.80	2230	2.69	0.11***
Number of female adults	3118	1.43	2230	1.38	0.05**
Number of kids	3118	3.18	2230	3.25	−0.07
Log income	3118	12.73	2230	12.76	−0.02
Rural household	3106	0.47	2225	0.45	0.02
Household head:					
Age	3096	46.42	2216	45.39	1.03***
Some primary education	3027	0.45	2165	0.45	0.01
Some secondary education	3027	0.15	2165	0.18	−0.03***
Primary source of income:					
Laborer/employee	2990	0.31	2162	0.32	−0.01
Domestic employee	2990	0.05	2162	0.04	0.01**
Day laborer	2990	0.26	2162	0.26	0.00
Self-employment	2990	0.32	2162	0.32	0.00
Partner in farm/plot	2990	0.06	2162	0.06	0.00
Proportion of regular income	3115	0.90	2230	0.91	−0.02***
Health shocks	3118	0.15	2230	0.13	0.02*
Other shocks	3118	0.22	2230	0.18	0.04***

Note. (1) refers to observations in the final sample and (0) to available observations excluded from the final sample (just before "Balanced panel (robustness)" in Table A.1). Second wave only. Robust standard errors; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

B Methodological appendix

Recall the flexible model in equation (9), which we reproduce below for convenience:

$$\ell_{jit} = \beta_0(r_{jit}) + \beta_1(r_{jit})\psi(y_{it}) + \beta_2(r_{jit})\eta_i + \varepsilon_{jit}. \tag{B.1}$$

We now provide further details on parameterization (Section B.1), estimation (Sections B.2 and B.3) and implementation (Section B.4).

B.1 Specification details

We view model (B.1) as a sequence of approximating parameter spaces — or sieves — for the nonparametric model in (3)-(4); see Chen (2007) for a technical review of the method of sieves.

In particular, we parameterize model (B.1) as

$$\ell_{jit} = \sum_{\tau=1}^{K_0} \beta_{0,\tau} h_\tau(r_{jit}) + \sum_{\tau=1}^{K_1}\sum_{\kappa=1}^{K_y} \beta_{1,\tau,\kappa} h_\tau(r_{jit}) g_\kappa(y_{it}) + \sum_{\tau=1}^{K_2} \beta_{2,\tau} h_\tau(r_{jit})\eta_i + \varepsilon_{jit}, \tag{B.2}$$

where $g_\kappa(y)$ are Hermite polynomials (we omit the constant term, i.e., g_1 is linear in y) and $h_\tau(r)$ are basis functions of natural cubic splines, with K_s knots ($K_s \geq 2$) for $s = \{0, 1, 2\}$.[5] For ease of notation, we set $K_s = K$ for all three terms in the body of the paper, and explicitly refer to $\beta_0(\cdot)$, $\beta_1(\cdot)$ and/or $\beta_2(\cdot)$. We normalize $\beta_{0,1} = 0$ and $\beta_{2,1} = 1$ to accommodate the level and scale of the fixed effects η_i. This implies there are $K_0 + K_1 K_y + K_2 - 2$ target parameters in model (B.2). The baseline specification we use for nonlinear models in Sections 5.2 and 5.3 sets $K_0 = K_1 = 3$, $K_y = 1$ and $K_2 = 1$ (additive fixed effects) or $K_2 = 2$ (interacted fixed effects).

It is often useful to rewrite (B.2) in vector notation. If $h_{1:K_s}(r)$ is used to indicate the $1 \times K_s$ array obtained by horizontal concatenation of the elements in $\{h_\tau(r)\}_{\tau=1}^{K_S}$, let us define

$$D_{jit} = (h_{2:K_0}(r_{jit}), h_{1:K_1}(r_{jit})g_1(y_{it}), \ldots, h_{1:K_1}(r_{jit})g_{K_y}(y_{it}))'$$

[5] Cubic natural splines are piecewise cubic polynomials that are twice continuously differentiable and restricted to be linear beyond the boundary knots. Differentiability is crucial to compute densities and quantile-based measures of nonlinear persistence, as in (14). In particular, let τ_k for $k \in \{1, \ldots, K_s\}$ index the knots in increasing order, which we place at the $k/(K_s + 1)$th quantiles of the empirical distribution of r_{jit}. The following K_s basis functions can be used to represent the spline model:

$$\left[1, r_{jit}, d_1(r_{jit}) - d_{K_s-1}(r_{jit}), \ldots, d_{K_s-2}(r_{jit}) - d_{K_s-1}(r_{jit})\right],$$

so that $K_s = 2$ corresponds to a linear spline, and for $K_s > 2$ we have

$$d_k(r) = \frac{(r - \tau_k)^3 \mathbb{1}\{r \geq \tau_k\} - \left(r - \tau_{K_s}\right)^3 \mathbb{1}\left\{r \geq \tau_{K_s}\right\}}{\tau_{K_s} - \tau_k}$$

for $k \in \{1, \ldots, K_s - 1\}$.

$$\beta_{0,1} = (\beta_{0,2:K_0}, \beta_{1,1:K_1,1}, \ldots, \beta_{1,1:K_1,K_y})',$$

and similarly $H_{jit} = h_{2:K_2}(r_{jit})'$ and $\beta_2 = \beta'_{2,2:K_2}$. We then stack observations (t, j) vertically for each unit to obtain

$$\ell_i = D_i\beta_{0,1} + \left(1_{TJ} + H_i\beta_2\right)\eta_i + \varepsilon_i, \tag{B.3}$$

where 1_A is a vector of ones of size A and where $\ell_i = (\ell_{1i1}, \ell_{2i1}, \ell_{3i1}, \ell_{1i2}, \ell_{2i2}, \ell_{3i2})'$ and so on.

B.2 Estimation: additive fixed effects

Model (B.2) is then a series regression model on the sequence of parameter sets defined by (K_0, K_1, K_2, K_y), with the additional twist that η_i is unobserved.

When η_i enters additively (set $K_2 = 1$), given the conditional mean assumption $E\left[\varepsilon_i \middle| r_i, y_i, \eta_i\right] = 0$, the model in (B.3) is a static fixed-effects regression that can be estimated using the within-group estimator. In other words, let $\tilde{\ell}_{jit} = \ell_{jit} - (TJ)^{-1}\sum_{(j,t)}\ell_{jit}$ denote variables in deviations with respect to means (recall that $T = 2$ and $J = 3$ here) and use $\tilde{\ell}_i$ for the corresponding vectors. Equation (B.3) then becomes $\tilde{\ell}_i = \tilde{D}_i\beta_{0,1} + \tilde{\varepsilon}_i$ and an estimate of $\beta_{0,1}$ can be obtained via least squares.

Penalization. When considering models where (K_0, K_1, K_y) grow large, regularized estimators might be attractive. We explore implementations that penalize the nonlinear sieve terms and might prove useful in richer setups where even more flexible specifications might be feasible. A simple implementation via a ridge penalty $\lambda > 0$ allows us to maintain the simplicity of the estimation method and an explicit solution that recovers the linear model as $\lambda \to \infty$.

B.3 Estimation: interacted fixed effects

As noted in Section A, treating $\{\eta_i\}_{i=1}^n$ as parameters to be estimated jointly with $\beta_{0,1}$ and β_2 results in an incidental parameters problem that precludes fixed-T consistent estimation. Here we generalize the linear instrumental variables (IV) strategy introduced in the main text, and discuss a method-of-moments approach in greater generality at the end.

Recall that we use $\tilde{\ell}_{jit} = \ell_{jit} - \bar{\ell}_i$ and so on as notation for variables in deviations with respect to means:

$$\begin{aligned}\tilde{\ell}_{jit} &= \tilde{D}_{jit}\beta_{0,1} + \tilde{H}_{jit}\eta_i\beta_2 + \tilde{\varepsilon}_{jit},\\ \bar{\ell}_i &= \bar{D}_i\beta_{0,1} + (1 + \bar{H}_i\beta_2)\eta_i + \bar{\varepsilon}_i.\end{aligned}$$

We consider the case $K_2 = 2$, so that effectively $H_{jit} = r_{jit}$. We look for linear transformations of the model that do not depend on η_i but still allow us to estimate β_2. Note that

$$H_{jit}\bar{\ell}_i - \bar{H}_i\ell_{jit} = H_{jit}\left(\bar{D}_i\beta_{0,1} + (1 + \bar{H}_i\beta_2)\eta_i + \bar{\varepsilon}_i\right) - \bar{H}_i\left(D_{jit}\beta_{0,1} + (1 + H_{jit}\beta_2)\eta_i + \varepsilon_{jit}\right)$$

$$= \left(H_{jit}\bar{D}_i - \bar{H}_i D_{jit}\right)\beta_{0,1} + \tilde{H}_{jit}\eta_i + H_{jit}\bar{\varepsilon}_i - \bar{H}_i\varepsilon_{jit},$$

where note the first element in $H_{jit}\bar{D}_i - \bar{H}_i D_{jit}$ is zero. We solve for the term involving η_i and plug it back into the model in deviations,

$$\tilde{\ell}_{jit} = \tilde{D}_{jit}\beta_{0,1} + \left(H_{jit}\bar{\ell}_i - \bar{H}_i\ell_{jit}\right)\beta_2 + \left(H_{jit}\bar{D}_i - \bar{H}_i D_{jit}\right)\gamma + \xi_{jit} \tag{B.4}$$

where $\xi_{jit} = \tilde{\varepsilon}_{jit} - H_{jit}\beta_2\bar{\varepsilon}_i + \bar{H}_i\beta_2\varepsilon_{jit}$ and $\gamma = -\beta_{0,1}\beta_2$, a generalization of the simple model considered in equation (21) in Section A.2. We need at least one instrument for $H_{jit}\bar{\ell}_i - \bar{H}_i\ell_{jit}$. Note that

$$E\left[H_{jit}\bar{\ell}_i - \bar{H}_i\ell_{jit}\middle| r_i, y_i\right] = \left(H_{jit}\bar{D}_i - \bar{H}_i D_{jit}\right)\beta_{0,1} + \tilde{H}_{jit}E\left[\eta_i\middle| r_i, y_i\right],$$

and thus any predictor of η_i (conditional on the included regressors) is a valid instrument under the assumptions of our model. We thus consider a set of $K_0 + K_1K_y - 1$ instruments given by $Z_{jit} = \tilde{H}_{jit}\bar{D}_i$; this corresponds to five instruments in the baseline specification used in Section 5.3. We then propose to estimate (B.4) by TSLS.

Note that the restriction $\gamma = -\beta_{0,1}\beta_2$ does not need to be imposed for consistent estimation. If one is willing to impose it, it can be done ex post via minimum distance estimation or ex ante via nonlinear GMM. We briefly explored the latter, which is exposed to numerical/convergence problems similar to those encountered by the more general nonlinear method-of-moments estimator described below.

A method-of-moments estimator. The IV estimator developed here retains the simplicity and linearity of the within-group estimator even as we move on to more flexible models, and we have found it to be a reliable approach. A general method-of-moments approach for the parameters in equation (B.2) is as follows.

Let $B_i(\beta_2)$ and $Q_i(\beta_2)$ denote the generalized between- and within-group transformations, respectively, defined as

$$B_i(\beta_2) = \left(\left(1_{TJ} + H_i\beta_2\right)'\left(1_{TJ} + H_i\beta_2\right)\right)^{-1}\left(1_{TJ} + H_i\beta_2\right)',$$
$$Q_i(\beta_2) = I_{TJ} - \left(1_{TJ} + H_i\beta_2\right)B_i(\beta_2),$$

where I_A is the identity matrix of size $A \times A$. Back to equation (B.2), we note that the generalized within-group residuals are mean independent of the regressors:

$$E\left[Q_i(\beta_2)\left(\ell_i - D_i\beta_{0,1}\right)\middle| r_i, y_i\right] = 0.$$

A nonlinear GMM estimator inspired in Chamberlain (1992) and Arellano and Bonhomme (2012) can then be constructed by exploiting these conditional moment restrictions. In some sense, the IV

strategy proposed above is a transparent way to find such informative restrictions for the interacted fixed-effects term.

B.4 Additional details on implementation

Here we discuss how we compute the objects of interest after estimation of the model parameters in equation (9) (reproduced here as equation (B.1)) and shed light on some additional details beyond those discussed in Section A.

Standardizing the data. When estimating flexible models, we first standardize the data as follows:

$$\check{r}_{jit} = \frac{r_{jit} - \bar{r}}{\bar{\sigma}_r}$$
$$\check{y}_{it} = \frac{y_{it} - \bar{y}}{\bar{\sigma}_y},$$

where $\bar{x}$ and $\bar{\sigma}_x$ are measures of location and scale for variable x; we use the median and the IQR, respectively, in our implementation. This helps standardize the range of variation of the data across units. Importantly, we need to undo these transformations before reporting the final output: letting $\check{q}_{it}(\tau)$ denote the τth conditional quantile on the standardized data, we need to calculate

$$q_{it}(\tau) = \bar{r} + \bar{\sigma}_r \check{q}_{it}(\tau),$$

and, similarly, $\rho_{it}(\tau) = (\bar{\sigma}_r/\bar{\sigma}_y)\check{\rho}_{it}(\tau)$ for the measure of nonlinear persistence in (15).

Estimation in growth rates. In Section A, we note that redefining $s_{jit} = r_{jit} - y_{it}$ is equivalent to estimating predictive distributions for growth rates and argue that this is a convenient transformation. Note that we are still interested in the range of values of r (or $\check{r}$): this entails careful adjustment of the support grid of conditional distribution functions in implementation. On a related note, the fact that the argument of the conditional distribution now depends on y has to be taken into account when computing numerical derivatives below.

Details on computing quantile-based measures of dispersion, skewness and persistence. Given estimates $(\hat{\beta}'_{0,1}, \hat{\beta}'_2)'$, the target summaries in Section 3.3 can be computed in three steps:

1. Obtain predicted probabilities.

   Given reference conditioning values $(\bar{y}, \bar{\eta})$ (usually a quantile of interest) and for r in a given grid r_{grid}, we calculate fitted probabilities $\hat{p} = \hat{F}(r, \bar{y}, \bar{\eta})$, which we collect in $\hat{p}_{\text{rgrid}}$. When non-monotonic, we follow Chernozhukov, Fernández-Val, and Galichon (2010) in sorting the original estimated curve into a monotone rearranged one.

2. Recover conditional quantiles.

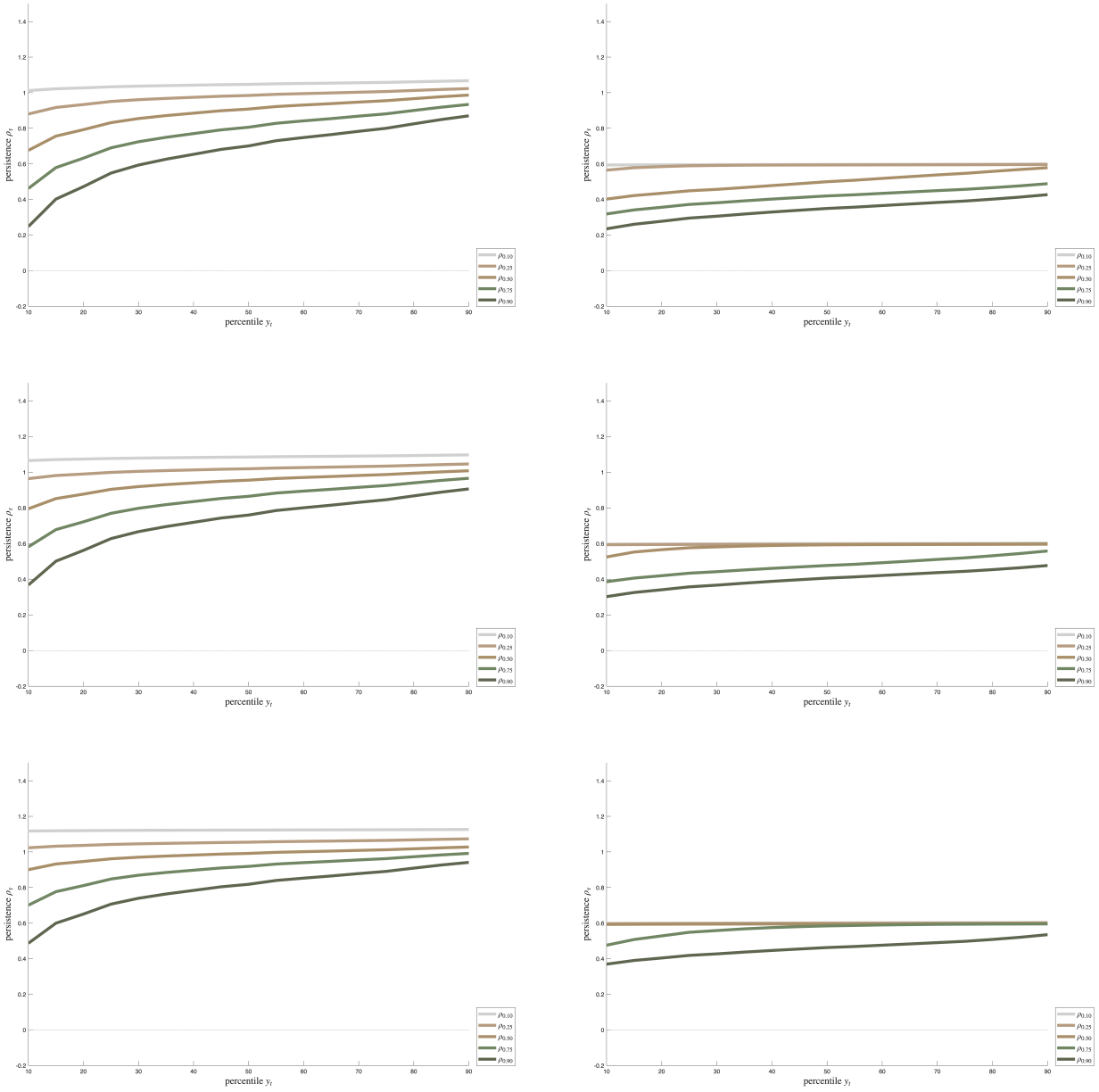
   This involves inverting the fitted conditional distribution function to obtain conditional quantiles for a given τ, which we denote $\hat{r}(\tau)$; see also equation (10) in the main text. We resort to interpolation within $\hat{p}_{\text{rgrid}}$.[6]

3. Compute target summaries.

   Quantile-based measures of dispersion and skewness as in equations (12) and (13) are then readily available (note the comment above on standardizing the data). Regarding nonlinear persistence, we calculate the derivatives in equation (15) numerically. Note that this requires recalculating predicted probabilities along the lines of step 1, so as to condition on $\hat{r}(\tau)$.

[6] Alternative methods are bracketing or root-finding algorithms, which solve for $\hat{r}(\tau)$ in $\tau = \hat{F}(\hat{r}(\tau), \bar{y}, \bar{\eta})$. These are model-based approaches that impose the (estimated) logit structure, which is problematic when it is non-monotonic. In fact, these algorithms impose implicit rearrangement methods that are starting-value dependent.

C Documenting heterogeneity

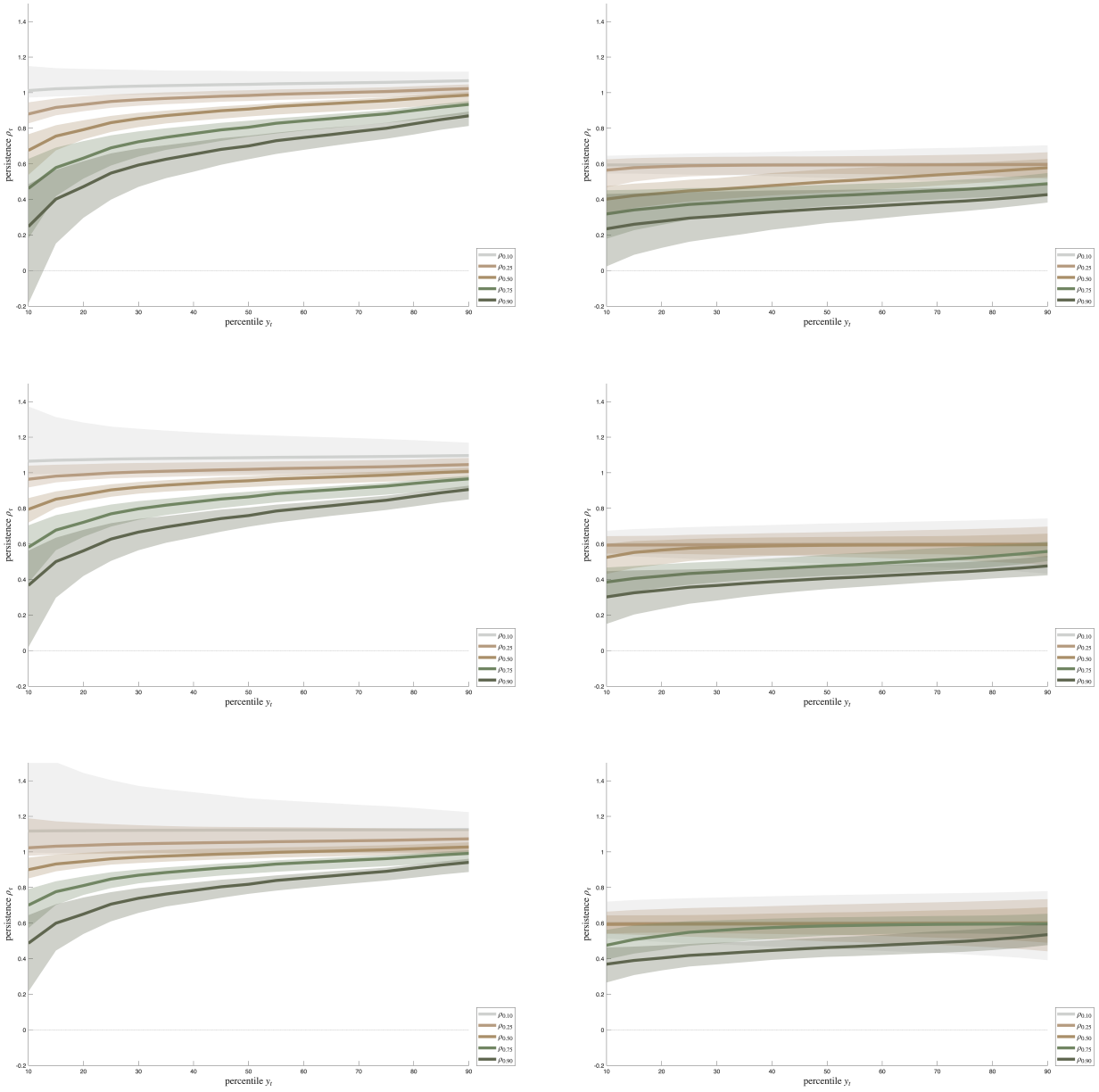


(a) India (b) Colombia

Notes: Panel (a): $N = 770$ households (India). Panel (b): $N = 2230$ households (Colombia). The figure reports estimates of nonlinear persistence from the flexible model with additive fixed effects in equation (16). Upper, middle, and lower subpanels correspond to the 10th, 50th, and 90th percentiles of η, respectively. Pointwise confidence bands are reported in Figure C.2.

FIGURE C.1. Nonlinear persistence (flexible model) at different values of η_i

C.1 Confidence bands



(a) India (b) Colombia

Notes: Panel (a): $N = 770$ households (India). Panel (b): $N = 2230$ households (Colombia). The figure reports estimates of nonlinear persistence from the flexible model with additive fixed effects in equation (16). Upper, middle, and lower subpanels correspond to the 10th, 50th, and 90th percentiles of η, respectively. Shaded areas denote 90% pointwise confidence bands based on block bootstrap.

FIGURE C.2. Nonlinear persistence (flexible model, with confidence bands) at different values of η_i

D Robustness: sample selection and modeling choices

In Sections D.1 and D.2, we report counterparts to (a subset of) the empirical results reported in the main text when including elicited probabilities that equal zero or one (applying the transformation below) and observations that violate strict monotonicity (two reported cumulative probabilities being equal for the same household). We find very similar results in both qualitative and quantitative terms for our target summary objects in both datasets. This is worth stressing in the case of Colombia, where this essentially implies doubling the sample size (see Table A.1) and introducing substantial additional elicitation error (as measured by residual variances).

Keeping zero/one probabilities. The logit transformation in equation (2) in Section 3 restricts observed, elicited probabilities p_{jit} to lie strictly between zero and one. We suggest here an alternative transformation — which still maps these probabilities to the real line — that allows us to keep these observations:

$$\ell_{jit} = \text{logit}\left(\check{p}_{jit}\right), \quad \check{p}_{jit} = \frac{p_{jit} + \frac{1}{2m}}{1 + \frac{J}{2m}}. \tag{D.1}$$

This is a generalization of the modified logit transformation of Cox and Snell (1989, p. 32) for binary data. In a context where p_{jit} are noisy measurements due to possible rounding and randomness in the elicitation process, the adjustment m can be interpreted as a measure of the accuracy of the elicitation such that $m = O\left(1/\sigma_\varepsilon^2\right)$. In particular, elicitation errors ε_{jit} in (3) can be seen as capturing sampling uncertainty from a hypothetical random sample of size m; see Arellano, Bonhomme, De Vera, Hospido, and Wei (2022, Online Appendix F) for additional details in the context of subjective expectations data and a Bayesian interpretation of (D.1).[7]

[7] Below, we set the regularization parameter m in equation (D.1) to $m = 10$ in both cases. Setting $m = 5$ or $m = 20$ leads to very similar results.

D.1 India: larger samples

	No FE	FE
ρ	0.96	0.93
	(0.94, 0.99)	(0.90, 0.95)
σ	1.04	0.60
	(0.96, 1.12)	(0.56, 0.63)
$IQR_{0.75}$	1.26	0.72
	(1.17, 1.36)	(0.67, 0.77)
$IQR_{0.90}$	2.53	1.44
	(2.33, 2.72)	(1.35, 1.53)
σ^2_η		0.18
		(0.15, 0.22)
σ^2_η village		0.12
		(0.11, 0.16)
σ^2_ε	1.05	0.98
	(1.03, 1.08)	(0.95, 1.02)

(a) India

	No FE	FE
ρ	0.72	0.51
	(0.70, 0.74)	(0.48, 0.54)
σ	1.49	1.02
	(1.45, 1.54)	(1.00, 1.05)
$IQR_{0.75}$	1.80	1.24
	(1.75, 1.86)	(1.21, 1.27)
$IQR_{0.90}$	3.61	2.48
	(3.50, 3.72)	(2.41, 2.54)
σ^2_η		0.54
		(0.51, 0.58)
σ^2_η village		0.10
		(0.10, 0.13)
σ^2_ε	1.75	1.34
	(1.72, 1.78)	(1.30, 1.37)

(b) Colombia

Notes: Panel (a): $N = 789$ households (India). Panel (b): $N = 4420$ households (Colombia). Results from the linear model (7), with and without fixed effects (FE), estimated on the robustness sample that includes reported zero/one probabilities and observations violating strict monotonicity (see Section D). This table is the counterpart to Table 3 in the main text. Year (survey round) dummies are included in all specifications; month (interview) dummies are also included for Colombia. 90% block bootstrap confidence intervals in parentheses.

TABLE D.1. Results from the linear model, robustness sample

	y_{p10}	y_{p50}	y_{p90}
$IQR_{0.75}$	0.83	0.63	0.54
	(0.75, 0.93)	(0.57, 0.66)	(0.48, 0.58)
$IQR_{0.90}$	1.69	1.29	1.10
	(1.56, 1.91)	(1.21, 1.37)	(0.99, 1.21)
$SK_{0.90}$	−0.04	−0.11	−0.15
	(−0.16, 0.04)	(−0.21, −0.05)	(−0.29, −0.05)
$\rho_{\tau 0.25}$	0.97	1.02	1.04
	(0.92, 1.04)	(0.98, 1.06)	(0.99, 1.08)
$\rho_{\tau 0.50}$	0.81	0.95	1.00
	(0.74, 0.87)	(0.92, 0.98)	(0.97, 1.02)
$\rho_{\tau 0.75}$	0.59	0.86	0.95
	(0.42, 0.72)	(0.82, 0.89)	(0.91, 0.98)
σ^2_η		0.19	
		(0.16, 0.23)	
σ^2_η village		0.11	
		(0.11, 0.16)	
σ^2_ε		0.95	
		(0.91, 0.99)	

Notes: $N = 789$ households. Results for India from the flexible model with additive fixed effects in equation (16). Estimates are based on the robustness sample that includes reported zero/one probabilities and observations violating strict monotonicity (see Section D). Year (survey round) dummies included; reference group is the first year. 90% block bootstrap confidence intervals in parentheses.

TABLE D.2. Results from the flexible model with additive fixed effects, India (robustness sample)

D.2 Colombia: larger samples

	y_{p10}	y_{p50}	y_{p90}
$IQR_{0.75}$	1.36	1.18	1.10
	(1.31, 1.42)	(1.14, 1.21)	(1.06, 1.14)
$IQR_{0.90}$	2.67	2.37	2.22
	(2.57, 2.78)	(2.31, 2.44)	(2.15, 2.29)
$SK_{0.90}$	0.08	0.05	0.01
	(0.04, 0.13)	(0.01, 0.08)	(−0.02, 0.04)
$\rho_{\tau 0.25}$	0.60	0.63	0.65
	(0.55, 0.64)	(0.59, 0.67)	(0.58, 0.70)
$\rho_{\tau 0.50}$	0.45	0.60	0.63
	(0.41, 0.51)	(0.56, 0.63)	(0.59, 0.66)
$\rho_{\tau 0.75}$	0.34	0.50	0.59
	(0.28, 0.40)	(0.46, 0.53)	(0.55, 0.62)
σ^2_η		0.53	
		(0.50, 0.57)	
σ^2_η village		0.10	
		(0.10, 0.13)	
σ^2_ε		1.33	
		(1.29, 1.36)	

Notes: N = 4420 households. Results for Colombia from the flexible model with additive fixed effects in equation (16). Estimates are based on the robustness sample that includes reported zero/one probabilities and observations violating strict monotonicity (see Section D). Year (survey round) and month (interview) dummies included; reference group is the first year and month. 90% block bootstrap confidence intervals in parentheses.

TABLE D.3. Results from the flexible model with additive fixed effects, Colombia (robustness sample)

E Questionnaires

Figure E.1 reports the original questionnaire for India. The questions on elicitation of subjective expectations follow those on income and income components and correspond to Section 6 of the household survey.[8] Figure E.2 shows the original questionnaire for Colombia (in Spanish).

[8] To be precise, this is the second-round version of the questionnaire. In the first-round version, five households were instead asked about monthly — rather than yearly — income. Importantly, the wording of the questions is unchanged, and the data included an identifier for these households, which are not part of the original sample in Table A.1.

INTERVIEWER: Add all income sources in the shaded column to calculate yearly income of the household.

5.	READ OUT CALCULATED YEARLY INCOME and ask: Is this a typical yearly income for your household?	1. yes 2. no, it is higher than typical 3. no, it is lower	
6.	IF NO: What would be a typical yearly income for your household?	(Rs.)	

IF ONLY INCOME SOURCE IS FROM DAIRY ACTIVITY (7) >> GO TO SECTION 7. ELSE, go on to question 7.

7.	Imagine that you have a very good year, every member of working age in the household managed to have work, and there were no droughts or anything the like. What would be the maximum amount of income your household would receive in such a situation in one year?	Y	(Rs.)	
8.	Now imagine the total opposite: the harvest is bad, animals get sick, finding work is not possible. What would be the yearly income of your household in such a situation?	X	(Rs.)	

INTERVIEWER: Calculate the following values:

Expected Income (threshold B):	B = (X+Y)/2	
Threshold A:	A = (B+X)/2	
Threshold C:	C = (B+Y)/2	

INTERVIEWER: Explain the rainfall question to the respondent (See extra Sheet)

R.1	So, what do you think how likely it is that it will rain *tomorrow*?	
R.2	So, what do you think how likely it is that it will rain within the *coming week*?	
R.3	So, what do you think how likely it is that it will rain within the *coming month*?	

9.	How likely do you think it is that your yearly income in the coming year will be higher than _____(A) Rupees?	
10.	How likely do you think it is that your yearly income in the coming year will be higher than _____(B) Rupees?	
12.	How likely do you think it is that your yearly income in the coming year will be higher than _____(C) Rupees?	

FIGURE E.1. India — questionnaire


FORMULARIO TIPO 1 No. MÓDULO 6 4

631 ¿El mes pasado recibió algún ingreso por concepto de trabajo, diferente al de su ocupación u oficio principal? Si 1 → ¿Cuánto recibió? $ No 2

632 ENTREVISTADORA: *Verifique la edad de _____ en **604** y marque de acuerdo con la respuesta registrada.* 10 a 24 años 1 → 635; 25 y más años 2

633 ¿El mes pasado recibió dinero por concepto de pensión de jubilación, sustitución pensional, invalidez o vejez? Si 1 → ¿Cuánto recibió? $ No 2

634 ¿El mes pasado recibió dinero por concepto de arriendos o intereses? Si 1 → ¿Cuánto recibió? $ No 2

635 ¿El mes pasado recibió dinero por otras fuentes diferentes al trabajo? (por ejemplo, venta o empeño de un bien) Si 1 → ¿Cuánto recibió? $ No 2

636 ENTREVISTADORA: *Verifique en el "Reporte de Seguimiento". La persona objeto de este módulo es:* Jefe del núcleo familiar seleccionado 1; Cónyuge del jefe del núcleo familiar seleccionado 2; Otro 3 → E

637 ENTREVISTADORA: *¿La persona objeto de este módulo debe aplicar "expectativas de ingreso"? Tenga en cuenta que esta sección, **aplica sólo a una persona del núcleo familiar seleccionado**.* Si 1; No 2 → E

D. EXPECTATIVAS DE INGRESO

ENTREVISTADORA: *Lea a su entrevistad@ el siguiente texto:*

*"*Ahora vamos a realizar un pequeño juego que consiste en lo siguiente: Aquí tenemos una regla que tiene una escala de 0 a 100. Queremos que la utilice para indicarnos qué tan seguro está Usted, de que alguna situación se va a presentar en el futuro, por ejemplo, si le preguntamos: ¿Qué tan seguro está de que mañana va a llover?.

1. Si Usted esta totalmente seguro que va a llover nos indica el punto 100 de la regla.
2. Si Usted está totalmente seguro de que no va a llover nos indica el punto 0 de la regla.
3. Y si Usted no está seguro de lo que va a ocurrir, pero cree que hay una alta probabilidad de que llueva se colocaría más cerca del 100 que del 0.
4. Y si cree que hay una alta probabilidad que no va a llover se colocaría más cerca del 0 que del 100.

"Ahora muéstreme en la regla qué tan seguro está de que mañana va a llover*" **(Que él indique con un lápiz)***.

638 Ahora suponga que el próximo mes los miembros de su familia que quieren trabajar, consiguen un trabajo bueno. (*Si tiene parcela, decir también: Imagine además que Usted obtiene una buena cosecha*). ¿Cuánto dinero cree que ganaría o le entraría en ese mes al hogar? X $ NS/NR → E

639 Suponga ahora todo lo contrario, que tienen muy poco trabajo el próximo mes (*Si tiene parcela, decir también: Suponga que la cosecha salió mal*), y que sólo viven de eso y de lo que la gente les da, y que la gente les da muy poco. ¿Cuánto dinero cree que recibiría en ese mes el hogar? Y $ NS/NR → E

640 ENTREVISTADORA: *Promedie las dos posibilidades (X y Y), y calcule el ingreso esperado del hogar. Mencione la cifra al entrevistado, diciendo **"entonces el ingreso promedio sería" (Z).*** Z (X+Y)/2 $

641 ENTREVISTADORA: *Calcule el valor de ingreso **M**, a partir del ingreso promedio.* M (Z+X)/2 $

642 ENTREVISTADORA: *Calcule el valor de ingreso **P**, a partir del ingreso promedio.* P (Z+Y)/2 $

643 Ahora vamos a jugar con la regla. Usted debe responder señalándome un punto en la regla, y la pregunta es la siguiente: ¿Qué tan seguro está Usted que el ingreso del hogar va a estar entre $_____ y $_____?

ENTREVISTADORA: *Si no entiende, repítale el ejemplo de la lluvia.*

A	B	C
Entre X y M	Entre X y Z	Entre X y P
%	%	%

ENTREVISTADORA: *Compruebe que la respuesta de C sea mayor que la de B y la de B mayor que la de A. Si no es así vuelva y repítale el ejemplo de la lluvia.*

1256556123 SEI; ARD-234 /DD-06/JUL-03

FIGURE E.2. Colombia — questionnaire

References

Arellano, M. and S. Bonhomme (2012): "Identifying distributional characteristics in random coefficients panel data models." *Review of Economic Studies*, 79, 987–1020.

Arellano, M., S. Bonhomme, M. De Vera, L. Hospido, and S. Wei (2022): "Income Risk Inequality: Evidence from Spanish Administrative Records." *Quantitative Economics*, 13, 1747–1801.

Attanasio, O. and B. Augsburg (2016): "Subjective Expectations and Income Processes in Rural India," *Economica*, 83, 416–442.

Chamberlain, G. (1992): "Efficiency bounds for semiparametric regression." *Econometrica*, 60, 567–596.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models." in *Handbook of Econometrics*, ed. by J. J. Heckman and E. Leamer, Elsevier, vol. 6B, chap. 76.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010): "Quantile and probability curves without crossing." *Econometrica*, 78, 1093–1125.

Cox, D. R. and E. J. Snell (1989): *Analysis of Binary Data*, Cambridge University Press, second ed.

Dominitz, J. and C. F. Manski (1997): "Perceptions of Economic Insecurity: Evidence from the Survey of Economic Expectations." *Public Opinion Quarterly*, 61, 261–287.