American Economic Review 2019, 109(5): 1623–1647
https://doi.org/10.1257/aer.20151684

Wealth Distribution and Social Mobility in the US: A Quantitative Approach†

By JESS BENHABIB, ALBERTO BISIN, AND MI LUO*

We quantitatively identify the factors that drive wealth dynamics in the United States and are consistent with its skewed cross-sectional distribution and with social mobility. We concentrate on three critical factors: (i) skewed earnings, (ii) differential saving rates across wealth levels, and (iii) stochastic idiosyncratic returns to wealth. All of these are fundamental for matching both distribution and mobility. The stochastic process for returns which best fits the cross-sectional distribution of wealth and social mobility in the United States shares several statistical properties with those of the returns to wealth uncovered by Fagereng et al. (2017) from tax records in Norway. (*JEL* D31, E13, E21, E25)

Wealth in the United States is unequally distributed, with a Gini coefficient of 0.82. It is skewed to the right, and displays a thick, right tail: the top 1 percent of the richest households in the United States hold over 33.6 percent of wealth.[1] At the same time, the United States is characterized by a nonnegligible social mobility, with an intergenerational Shorrocks mobility index 0.88.[2] This paper attempts to quantitatively identify the factors that drive wealth dynamics in the United States and are consistent with the observed cross-sectional distribution of wealth and with the observed social mobility.

To this end, we first develop a macroeconomic model displaying various distinct wealth accumulation factors. Once we allow for an explicit demographic structure, the model delivers implications for social mobility as well as for the cross-sectional

*Benhabib: Department of Economics, New York University, 19 W 4th Street, FL 6, New York, NY 10012, and the National Bureau of Economic Research (email: jess.benhabib@nyu.edu); Bisin: Department of Economics, New York University, 19 W 4th Street, FL 6, New York, NY 10012, and the National Bureau of Economic Research (email: alberto.bisin@nyu.edu); Luo: Department of Economics, Emory University, 1602 Fishburne Drive, Atlanta, GA 30307 (email: mi.luo@emory.edu). Luigi Pistaferri was the coeditor for this paper. Thanks to seminar audiences at Duke, NYU, Minneapolis Fed, SED-Warsaw, Lake Baikal Summer School, SAET-Cambridge, University College of London, Wharton School, NBER Summer Institute. Special thanks to Alberto Alesina, Fernando Alvarez, Orazio Attanasio, Laurent Calvet, Tim Christensen, Tim Cogley, Mariacristina De Nardi, Pat Kehoe, Dirk Krueger, Per Krusell, Konrad Menzel, Ben Moll, Andrew Newman, Tom Sargent, Ananth Seshadri, Rob Shimer, Kevin Thom, Gianluca Violante, Daniel Xu, Fabrizio Zilibotti. Special thanks to Luigi Guiso, for many illuminating discussions and for spotting a mistake in a previous version, to the editor and the referees for their exceptional work on the paper. Generous financial support from the Washington Center for Equitable Growth is gratefully acknowledged.
†Go to https://doi.org/10.1257/aer.20151684 to visit the article page for additional materials and author disclosure statements.

[1] See Díaz-Giménez, Glover, and Ríos-Rull (2011), Table 6, elaborating data from the 2007 SCF.

[2] See Charles and Hurst (2003), Table 2, from PSID data. By construction, mobility matrices have Shorrocks indices increasing as the transition step gets long (indeed the index converges to 1 as the step goes to ∞).

distribution. We then match the moments generated by the model to several empirical moments of the observed distribution of wealth as well as of the social mobility matrix. While the model is very stylized and parsimonious, it allows us to identify various distinct wealth accumulation factors through their distinct role on inequality and mobility.

Many recent studies of wealth distribution and inequality focus on the relatively difficult task of explaining the thickness of the upper tail. We shall concentrate mainly on three critical factors previously shown, typically in isolation from each other, to affect the tail of the distribution, empirically and theoretically. First, a skewed and persistent distribution of *stochastic earnings* translates, in principle, into a wealth distribution with similar properties. A large literature in the context of Aiyagari-Bewley economies has taken this route, notably Castañeda et al. (2003) and Kindermann and Krueger (2015).[3] Another factor which could contribute to generating a skewed distribution of wealth is *differential saving rates* across wealth levels, with higher saving and accumulation rates for the rich. In the literature this factor takes the form of non-homogeneous bequests, bequests as a fraction of wealth that are increasing in wealth; see, for example, De Nardi (2004).[4] Stochastic idiosyncratic returns to wealth, or *capital income risk*, also has been shown to induce a skewed distribution of wealth, in Benhabib, Bisin, and Zhu (2011); see also Quadrini (2000) and Cagetti and De Nardi (2006), which focuses on entrepreneurial risk.[5] Finally, allowing rates of return on wealth to be increasing in wealth might also add to the skewness of the distribution. This could be due, e.g., to the existence of economies of scale in wealth management, as in Kacperczyk, Nosal, and Stevens (2015), or to fixed costs of holding high return assets, as in Kaplan, Moll, and Violante (2016). See Saez and Zucman (2016), Fagereng et al. (2016, 2017), and Piketty (2014, p. 447) for evidence about the relationship between returns and wealth.

While all of these factors possibly contribute to produce skewed wealth distributions, their relative importance remains to be ascertained.[6] In our quantitative analysis we find that all of the factors we study (stochastic earnings, differential savings, and capital income risk) have a fundamental role in generating the thick right tail of the wealth distribution and sufficient social mobility in the wealth accumulation process. We also identify a distinct role for these factors. Capital income risk and differential savings both contribute to generating the thick tail. Their effect on social mobility is however more nuanced: both differential savings and capital income risk increase social mobility across the distribution, more pronouncedly at the top in the case of capital income risk, while decreasing the probability of escape from the

[3] Several papers in the literature include a stochastic length of life (typically, "perpetual youth") to complement the effect of skewed earnings on wealth. We do not include this in our model as it has counterfactual demographic implications.

[4] See also Piketty (2014), which directly discusses the saving rates of the rich.

[5] Stochastic discount factors, as introduced by Krusell and Smith (1998), induce a skewed distribution of wealth through a similar mechanism. However, such discount factors are nonmeasurable, while microdata allowing estimates of capital income risk are instead rapidly becoming more available; see, e.g., the tax records for Norway studied by Fagereng et al. (2016, 2017) and the Swedish data studied by Bach, Calvet, and Sodini (2017).

[6] Other possible factors which qualitatively would induce skewed wealth distributions include a precautionary savings motive for wealth accumulation. In fact, the precautionary motive, by increasing the savings rate at low wealth levels under borrowing constraints and random earnings, works in the opposite direction of savings rates increasing in wealth. We do not exploit this channel for simplicity, assuming that life-cycle earnings profiles are random across generations but deterministic within lifetimes.

bottom 20 percent. On the other hand, stochastic earnings have a limited role in filling the tail of the wealth distribution but are fundamental in inducing enough mobility in the wealth process. Finally, a rate of return of wealth increasing in wealth itself is also apparently supported in our estimates, improving the fit of the model across the wealth distribution (though, without directly observing return data, this mechanism is somewhat poorly identified).

The rest of the paper is structured as follows. Section I lays out the theoretical framework. Section II explains our quantitative approach and data sources we use. Section III shows the baseline results with the model fit for both targeted and untargeted moments. The main extensions and robustness exercises we perform are also discussed in this section. Section IV presents several counterfactual exercises, where we re-estimate the model shutting down one factor at a time. Section V introduces an empirical exercise where we relax the stationarity assumption on the wealth distribution and measure the transition speed our model delivers. Section VI concludes.

I. Wealth Dynamics and Stationary Distribution

Most models of the wealth dynamics in the literature focus on deriving skewed distributions with thick tails, e.g., Pareto distributions (power laws).[7] While this is also our aim, we more generally target the whole wealth distribution and its intergenerational mobility properties. To this end we study a simple microfounded model (a standard macroeconomic model in fact) of life-cycle consumption and savings. While very parsimonious, the model exploits the interaction of the factors identified in the Introduction that tend to induce skewed wealth distributions: stochastic earnings, differential saving and bequest rates across wealth levels, and stochastic returns on wealth.

Each agent's life span is finite and deterministic, T years. Every period t, consumers choose consumption c_t and accumulate wealth a_t, subject to a no-borrowing constraint. Consumers leave wealth a_T as a bequest at the end of life T. Each agent's preferences are composed of a per-period utility from consumption, $u(c_t)$, at any period $t = 1, \ldots, T$, and a warm-glow utility from bequests at T, $e(a_T)$. Their functional forms display constant relative risk aversion,

$$u(c_t) = \frac{c_t^{1-\sigma}}{1-\sigma}, \quad e(a_T) = A\frac{a_T^{1-\mu}}{1-\mu}.$$

Wealth accumulates from savings and bequests. Idiosyncratic rates of return r and life-time labor earnings profiles $w = \{w_t\}_{t=1}^{T}$ are drawn from a distribution at birth, possibly correlated with those of the parent, deterministic within each generation.[8]

[7] See Benhabib and Bisin (2018) for an extensive survey of the theoretical and empirical literature on the wealth distribution.

[8] As we noted, assuming deterministic earning profiles amounts to disregarding the role of intragenerational life-cycle uncertainty and hence of precautionary savings. While the assumption is motivated by simplicity, see Keane and Wolpin (1997); Huggett, Ventura, and Yaron (2011); and Cunha, Heckman, and Schennach (2010) for evidence that the life-cycle income patterns tend to be determined early in life.

We emphasize that r and w are stochastic over generations only: agents face no uncertainty within their life span. Lifetime earnings profiles are hump-shaped, with low earnings early in life. Borrowing constraints limit how much agents can smooth lifetime earnings.

Let $\beta < 1$ denote the discount rate. Let $V_t(a_t)$ denote the present discounted utility of an agent with wealth a_t at the beginning of period t. Given initial wealth a_0, earnings profile w, and rate of return r, each agent's maximization problem, written recursively, then is

$$V_t(a) = \max_{c,a'} u(c) + \beta V_{t+1}(a')$$

subject to

$$a' = (1+r)a - c + w,$$

$$0 \leq c \leq a, \quad t = 1,\ldots,T-1,$$

$$V_T(a) = u(c) + e(a').$$

The solution of the recursive problem can be represented by a map,

$$a_T = g(a_0; r, w).$$

Following Benhabib, Bisin, and Zhu (2011), we exploit the map $g(\cdot)$ as the main building block to construct the stochastic wealth process across generations. Adding an apex n to indicate the generation and slightly abusing notation, we denote with $\{r^n, w^n\}_n$ the stochastic process over generations for the rate of return on wealth r and earnings w. We assume it is a finite irreducible Markov chain. We assume also that r^n and w^n are independent, though each is allowed to be serially correlated, with transition $P(r^n \mid r^{n-1})$ and $P(w^n \mid w^{n-1})$. The life-cycle structure of the model implies that the initial wealth of the nth generation coincides with the final wealth of the $(n-1)$th generation: $a^n = a_0^n = a_T^{n-1}$. We can then construct a stochastic difference equation for the initial wealth of dynasties, induced by $\{r^n, w^n\}_n$, mapping a^{n-1} into a^n:

$$a^n = g(a^{n-1}; r^n, w^n).$$

This difference equation in turn induces a stochastic process $\{a^n\}_n$ for initial wealth a.

It can be shown that, under our assumptions, the map $g(\cdot)$ can be characterized as follows:

- If $\mu = \sigma$, then $g(a_0; r, w) = \alpha(r,w)a_0 + \beta(r,w)$;
- If $\mu < \sigma$, then $\frac{\partial^2 g}{\partial a_0^2}(a_0; r, w) > 0$.

In the first case, $\mu = \sigma$, the savings rate is $\alpha(r, w)$ and it is independent of wealth. In this case, the wealth process across generations is represented then by a linear stochastic difference equation in wealth, which has been closely studied in the math literature (see De Saporta 2005). Indeed, if $\mu = \sigma$, under general conditions,[9] the stochastic process $\{a^n\}_n$ has a stationary distribution whose tail is independent of the distribution of earnings and asymptotic to a Pareto law,

$$\Pr(a > \underline{a}) \sim Q\underline{a}^{-\gamma},$$

where $Q \geq 1$ is a constant and $\lim_{N\to\infty} E\left(\prod_{n=0}^{N-1} \left(\alpha(r^{-n}, w^{-n})\right)^{\gamma}\right)^{\frac{1}{N}} = 1$.[10]

If instead, keeping σ constant, $\mu < \sigma$, differential savings rate emerge, increasing with wealth. In this case, a stationary distribution might not exist; but if it does,

$$\Pr(a > \underline{a}) \geq Q\underline{a}^{-\gamma},$$

and hence it displays a thick tail.

Finally, the model is straightforwardly extended to allow for the Markov states of the stochastic process for r to depend on the initial wealth of the agent a. In this case, the intergenerational wealth dynamics have properties similar to the $\mu < \sigma$ case: a stationary distribution might not exist; but if it does, it displays a thick tail.

II. Quantitative Analysis

The objective of this paper, as we discussed in the introduction, consists in measuring the relative importance of various factors which determine the wealth distribution and the social mobility matrix in the United States. The three factors are stochastic earnings, differential saving and bequest rates across wealth levels, and stochastic returns on wealth. These are represented in the model by the properties of the dynamic process and the distribution of (r^n, w^n) and by the parameters μ and σ, which imply differential savings (the rich saving more) when $\mu < \sigma$.

A. *Methodology*

We estimate the parameters of the model described in the previous section using a method of simulated moments (MSM) estimator: (i) we fix (or externally calibrate) several parameters of the model; (ii) we select some relevant moments of the wealth process as target in the estimation; and (iii) we estimate the remaining parameters by matching the targeted moments generated by the stationary distribution induced by the model and those in the data. The quantitative exercise is predicated then on

[9] More precisely, the tail of earnings must be not too thick and furthermore $\alpha(r^n, w^n)$ and $\beta(r^n, w^n)$ must satisfy the restrictions of a *reflective process*. See Grey (1994); Hay, Rastegar, and Roitershtein (2011); and Benhabib, Bisin, and Zhu (2011) for a related application.

[10] While a denotes initial wealth, it can be shown that when the distribution of initial wealth has a thick tail, the distribution of wealth also does. See Benhabib, Bisin, and Zhu (2011) for the formal result.

the assumption that the wealth and social mobility observed in the data are generated by a stationary distribution.[11]

More formally, let θ denote the vector of the parameters to be estimated. Let m_h, for $h = 1, \ldots, H$, denote a generic empirical moment; and let $d_h(\theta)$ the corresponding moment generated by the model for a given parameter vector θ. We minimize the deviation between each targeted moment and the corresponding simulated moment. For each moment h, define $F_h(\theta) = d_h(\theta) - m_h$. The MSM estimator is

$$\hat{\theta} = \arg\min_{\theta} \mathbf{F}(\theta)' W \mathbf{F}(\theta)$$

where $\mathbf{F}(\theta)$ is a column vector in which all moment conditions are stacked, i.e., $\mathbf{F}(\theta) = [F_1(\theta), \ldots, F_H(\theta)]^T$. The weighting matrix W in the baseline is a diagonal matrix with identical weights for all but the last moment of both the wealth distribution and the mobility moments, which are overweighted (ten times), according to the prior that matching the tail of the distribution is a fundamental objective of our exercise.[12] This is also a reasonable approximation to optimal weighting: an efficient two-step estimation with the optimal weighting matrix produces no relevant changes on estimated parameters nor on fit; see online Appendix C.4 for details.

The model is solved with the *collocation method* by Miranda and Fackler (2004): see online Appendix A.1. The objective function is highly nonlinear in general and therefore, following Guvenen (2016), we employ a global optimization routine for the MSM estimation: see online Appendix A.2.

In our quantitative exercise we proceed as follows.

(i) We fix $\sigma = 2$, $T = 36$, $\beta = 0.97$ per annum. We feed the model with a stochastic process for individual earnings profiles, w^n, and its transition across generations, $P(w^n \mid w^{n-1})$. Both the earning process and its transition are taken from data; respectively from the PSID and the federal income tax records studied by Chetty et al. (2014).

(ii) We target as moments:
- the bottom 20 percent, 20–40 percent, 40–60 percent, 60–80 percent, 80–90 percent, 90–95 percent, 95–99 percent, and the top 1 percent wealth shares; and
- the diagonal of the (age-independent) social mobility Markov chain transition matrix defined over quintiles.

(iii) We estimate:
- preference parameters μ, A; and

[11] Very few studies in the literature deal with the transitional dynamics of wealth and its speed of transition along the path, though this issue has been put at the forefront of the debate by Piketty (2014). Notable and very interesting exceptions are Gabaix et al. (2016); Kaymak and Poschke (2016); and Hubmer, Krusell, and Smith (2017). We extend the analysis to possibly nonstationary distributions in Section V as a robustness check. Our preliminary results are encouraging, in the sense that the model seems to be able to capture the transitional dynamics with parameters estimates not too far from those obtained under stationarity.

[12] See Altonji and Segal (1996) for a justification for the adoption of an identity weighting matrix.

Table 1—Life-Cycle Earnings ($thousands) Profiles

	Age range					
Percentile	[25–30]	[31–36]	[37–42]	[43–48]	[49–54]	[55–60]
0–10	9.760	11.55	12.06	12.81	11.74	8.222
10–20	19.95	24.01	25.2	26.42	24.66	19.08
20–30	26.85	32.58	34.96	36.46	33.56	26.78
30–40	33.05	40.33	43.95	45.55	42.23	34.39
40–50	39.02	47.70	52.42	54.37	51.18	42.96
50–60	45.05	54.84	60.70	63.09	60.34	51.91
60–70	51.40	65.10	69.42	72.89	70.63	61.65
70–80	59.16	73.06	80.37	85.09	82.78	74.35
80–90	70.33	87.21	97.51	103.5	101.4	93.42
90–100	100.3	138.1	169.5	182.4	183.4	180.4

Source: Calculated from the cleaned PSID data provided by Heathcote, Perri, and Violante (2010).

- a parameterization of the stochastic process for r defined by 5 states r_i and 5 diagonal transition probabilities, $P(r^n = r_i \mid r^{n-1} = r_i), i = 1, \ldots, 5$, restricting instead the 5×5 transition matrix to display constantly decaying off-diagonal probabilities except for the last row for which we assume constant off-diagonal probabilities.[13]

In total, therefore, the baseline model is exactly identified: we target 12 moments and we estimate 12 parameters.

In Section IIID we modify the stochastic process for r to allow returns to depend on the initial wealth a of the agent. We do this parsimoniously, without increasing the dimensionality of the parameter space. In Section IIID we experiment with an alternative social mobility matrix, defined over the same percentiles of the wealth distribution. This adds three moments to the estimation and the model is hence over-identified.

B. *Data*

Our quantitative exercise requires data for labor earnings, wealth distribution, and social mobility.

Labor Earnings.—We use ten deterministic life-cycle household-level earnings profiles at different deciles, as estimated by Heathcote, Perri, and Violante (2010) from the Panel Study of Income Dynamics (PSID), 1967–2002.[14] We construct the profiles as follows. For each of six age brackets we compute the averages of the earnings deciles, corresponding to the columns of Table 1. The deterministic lifetime profiles are then constructed assuming agents stay in the same decile for their

[13]Formally, $P(r^n = r_i \mid r^{n-1} = r_j) = P(r^n = r_i \mid r^{n-1} = r_i)e^{-\lambda j}$, $i = 1,2,3,4$, $j \neq i$, λ such that $\sum_{j=1}^{5} P(r^n = r_i \mid r^{n-1} = r_j) = 1$; and $P(r^n = r_5 \mid r^{n-1} = r_j) = \frac{1}{4}(1 - P(r^n = r_5 \mid r^{n-1} = r_5))$. We adopt a restricted specification in order to reduce the number of parameters we need to estimate. This particular specification performs better than one with constant off-diagonal probabilities as well as one with decaying off-diagonal probabilities in all rows.

[14]We detrend life-cycle earning profiles by conditioning out year dummies in a log-earnings regression; see online Appendix B.1 for the details of the procedure.

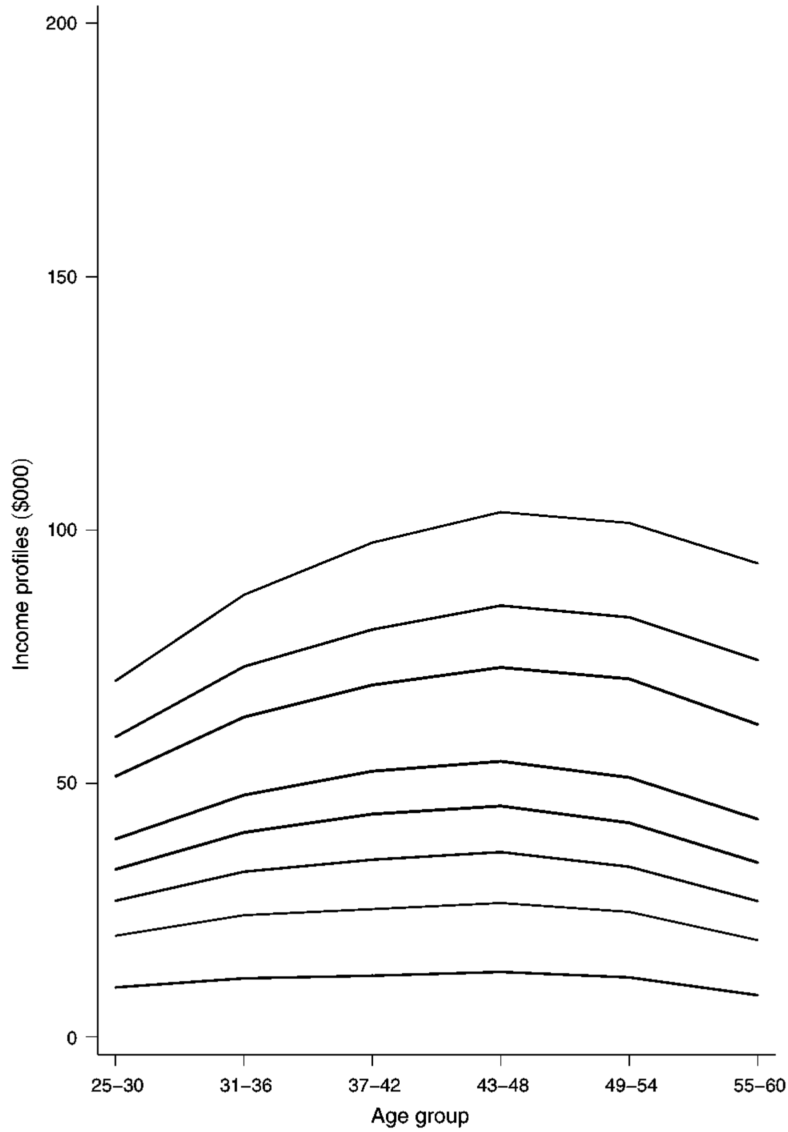

Figure 1. Life-Cycle Earnings Profiles by Deciles

Source: The data source is the same as in Table 1.

whole lifetime, corresponding to the ten rows of Table 1. Agents randomly draw one of these earnings profiles at the beginning of life according to an intergenerational transition matrix. These profiles are drawn in Figure 1.[15]

[15]The panel data on earnings from the US Social Security Administration (SSA) are not yet generally available. However, the crucial aspect of earnings data, for our purposes, is that they are far from skewed enough to account by themselves for the skewness of the wealth distribution. This is in fact confirmed on SSA data directly by Guvenen et al. (2016, Section 7.2.II) and by De Nardi, Fella, and Paz-Pardo (2016). See also Hubmer, Krusell, and Smith (2017).

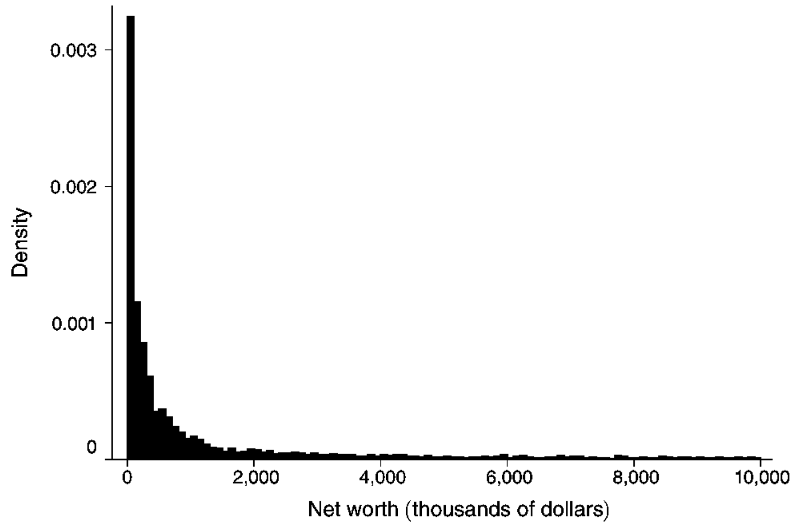

FIGURE 2. WEALTH DISTRIBUTION IN THE SCF 2007 (WEIGHTED)

Source: Net wealth, from 2007 SCF, truncated at 0 on the left and, for the purpose of the figure only, truncated at 10 million on the right.

The intergenerational transition matrix for earnings we use is from Chetty et al. (2014). The data in Chetty et al. (2014) refer to the 1980–1982 US birth cohort and their parental income. We reduce it to a ten-state Markov chain.[16]

Wealth Distribution.—We use wealth distribution data from the Survey of Consumer Finances (SCF) 2007.[17] The wealth variable we use is *net wealth*, the sum of net financial wealth and housing, minus any debts. The distribution is very skewed to the right. We take the shares from the cleaned version in Díaz-Giménez, Glover, and Ríos-Rull (2011). Figure 2 displays the histogram of the wealth distribution.

Table 2 displays the wealth share moments we use.

Social Mobility.—As for wealth transition across generations, we use the mobility matrix calculated by Charles and Hurst (2003), Table 2, from PSID data. This matrix is constructed by means of pairs of simultaneously alive parent and child of different ages. To eliminate age effects, the matrix is obtained by computing transitions from the residuals of the wealth of parents and children after conditioning on age and age squared.

The resulting matrix is shown in Table 3. The matrix shows substantial mobility, with a Shorrocks index of 0.88.[18]

[16] See online Appendix B.2 for details.

[17] As noted, the wealth distribution in our methodology is to be interpreted as stationary. Choosing 2007 avoids the nonstationary changes due to the Great Recession.

[18] Formally, for a square mobility transition matrix A of dimension m, the Shorrocks index given by $s(A) = \frac{m - \sum_j a_{jj}}{m-1} \in (0,1)$, with 0 indicating complete immobility.

TABLE 2—WEALTH DISTRIBUTION MOMENTS

Percentile	0–20	20–40	40–60	60–80	80–90	90–95	95–99	99–100
Wealth share	−0.002	0.001	0.045	0.112	0.120	0.111	0.267	0.336

Source: Calculated by Díaz-Giménez, Glover, and Ríos-Rull (2011) from the 2007 SCF.

TABLE 3—INTERGENERATIONAL SOCIAL MOBILITY TRANSITION MATRIX

	Percentile (child)				
Percentile (parent)	0–20	20–40	40–60	60–80	80–100
0–20	0.36	0.29	0.16	0.12	0.07
20–40	0.26	0.24	0.24	0.15	0.12
40–60	0.16	0.21	0.25	0.24	0.15
60–80	0.15	0.13	0.20	0.26	0.26
80–100	0.11	0.16	0.14	0.24	0.36

Source: From Table 2 in Charles and Hurst (2003). Note that we exchange the row and the column from their version.

In Section IIID we reproduce the estimation exercise in our baseline using an alternative social mobility matrix, using the 2007–2009 SCF panel data, with transitions computed for a synthetic agent over his/her age profile.[19]

III. Estimation Results

The baseline estimation results are reported in Section IIIA, Table 4. The targeted simulated moments of the estimated model are reported and compared to their counterpart in the data in Section IIIB, Table 5. Some independent evidence which bears on the fit of the model is discussed in Section IIIC. Extensions where we re-estimate the model to allow for rates of return dependent on wealth and to match an alternative social mobility matrix constructed using the 2007–2009 SCF panel data are discussed, respectively, in Section IIID.

A. *Parameter Estimates*

The upper part of Table 4 reports the estimates of the preference parameters. The lower part of Table 4 reports the estimated state space and diagonal of the transition matrix of the five-state Markov process for r we postulate. It also reports, to ease the interpretation of the estimates, the implied mean and standard deviation of the process, $E(r)$, $\sigma(r)$; as well as its autocorrelation, $\rho(r)$, computed fitting an AR(1) on simulated data from the estimated process.[20] The standard errors, also reported in the table, are obtained by bootstrapping; details are in online Appendix A.3.

[19] In addition in online Appendix B.3, we also describe another alternative social mobility matrix based on the social mobility matrix of Kennickell and Starr-McCluer (1997) using the SCF panel 1983–1989.

[20] The full transition matrix for r is reported in online Appendix C.1.