University of Michigan Retirement Research Center

Working Paper
WP 2013-302

What Do Data on Millions of U.S. Workers Say About Life Cycle Income Risk?

Fatih Guvenen, Fatih Karahan, Serdar Ozkan, and Jae Song

MRRC

Project #: UM13-08

What Do Data on Millions of U.S. Workers Say About Life Cycle Income Risk?

Fatih Guvenen
University of Minnesota

Fatih Karahan
Federal Reserve Bank of New York

Serdar Ozkan
Federal Reserve Board

Jae Song
Social Security Administration

October 2013

Michigan Retirement Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48104
www.mrrc.isr.umich.edu
(734) 615-0422

Acknowledgements
The research reported herein was performed pursuant to a grant from the U.S. Social Security Administration (SSA) funded as part of the Retirement Research Consortium through the Michigan Retirement Research Center (Grant # 5 RRC08098401-05-00). The opinions and conclusions expressed are solely those of the author(s) and do not represent the opinions or policy of SSA or any agency of the Federal Government or the Michigan Retirement Research Center.

What Do Data on Millions of U.S. Workers Say About Life Cycle Income Risk?

Abstract

This paper sheds new light on individual labor income risk using a unique and confidential dataset from the Social Security Administration on individuals' earnings histories. The substantial sample size allows us to cut the data in different and novel ways and document how earnings risk varies over the lifecycle and across individuals that differ in their lifetime income. The main conclusion of our research is that earnings risk varies significantly across the population in the following ways. First, the overall size of income risk becomes smaller with age, from age 25 to 50 and then increases again. Second, as individuals age, the likelihood of getting very small and very large shocks increases relative to the likelihood of middling shocks. Third, income shocks becomes more left skewed with age, meaning that, relative to the average change in income, a large fall becomes more likely than a large rise as individuals get older. Finally, earnings growth rates are dramatically different from for individuals ranked by their lifetime income: individuals with lifetime earnings in the top 5% experience a growth rate from age 25 to 55 that is 10 times larger than individuals with average lifetime earnings. To provide useful input to policy relevant research, we estimate an econometric process that captures these salient features of earnings dynamics to provide a reliable “user's guide” for applied economists.

Citation

Guvenen, Fatih, Fatih Karahan, Serdar Ozkan, and Jae Song (2013). “What Do Data on Millions of U.S. Workers Say About Life Cycle Income Risk?” Ann Arbor MI: University of Michigan Retirement Research Center (MRRC) Working Paper, WP 2013-302. http://www.mrrc.isr.umich.edu/publications/papers/pdf/wp302.pdf

1 Introduction

The importance of idiosyncratic labor income risk for individuals' economic choices and, hence, their welfare is hard to overstate. The literature that relies on incomplete-markets (or heterogeneous-agent) models is continuing to expand at a rapid pace. A crucial ingredient in this research is the precise nature of income risk that researchers feed into their models. For example, predicting individuals' lifecycle consumption-savings behavior, which is at the heart of the discussions on retirement wealth and the role of the Social Security system, requires a sound understanding of how workers perceive their lifetime income risk.

The goal of this paper is to shed new light on idiosyncratic income risk using a unique and confidential dataset from the Social Security Administration on individuals' earnings histories that has three key advantages: (i) a very large sample size (with 5+ million individuals) with a long time span (1978–2011), (ii) minimal measurement error, and (iii) no top-coding. These features of the dataset allow us to relax a number of restrictive assumptions that previous studies were forced to make. The substantial sample size allows us to cut the data in different and novel ways and document some interesting empirical facts. First, earnings changes display extreme leptokurtosis, meaning that compared to a normal distribution (with the same standard deviation), most earnings changes are very close to zero but few changes are extremely large. The resulting distribution looks very different from Gaussian, which is the typical assumption made in the literature. Second, there is enormous dispersion in the variance of earnings shocks across individuals: the top 10% most volatile individuals have an average standard deviation of shocks that is 6 times larger than the least volatile 10%. Third, the lifecycle growth rate of earnings varies strongly with the *level* of lifetime earnings. For example, the individual with the median lifetime earnings experiences an earnings growth of 30% from age 30 to 60, whereas for the the individual in the 95th percentile, this figure is 200%; and for the individual in the 99th percentile, this figure is 1000%. These and other features of individual earnings turn out to be difficult to capture with standard specifications used in the existing literature. This paper estimates a set of stochastic processes with increasing generality to capture these salient features of earnings dynamics to provide a reliable "user's guide" for applied economists.

The data used in this paper comes from a 10% representative sample of the US males with a Social Security Number (SSN), between the ages 25 and 60 from 1978 to 2011. There are about four million individuals in this sample in 1978, and this number grows to approximately six million individuals by 2011. Furthermore, earnings records are uncapped (no top-coding),

allowing us to study individuals with very high incomes.[1] Second, the substantial sample size allows us to employ flexible methods and rich econometric specifications and still obtain extremely precise estimates. Third, thanks to their records-based nature, the data contain very little measurement error, which is a serious issue with survey-based micro datasets.[2]

2 Empirical Strategy

The existing approaches to estimating income dynamics face two important challenges. First, the bulk of the literature (with very few exceptions[3]) relies on the (often implicit) assumption that income shocks can be approximated reasonably well with a log normal distribution. This assumption, combined with an AR(1) or random walk specification to capture the accumulation of such shocks, made higher order moments irrelevant and allowed researchers to focus their estimation to match the covariance matrix of log income either in levels or in first difference form. Our investigations so far from the SSA data reveal that this assumption is grossly counterfactual, with important implications.

Below, we present a number of statistics that require a very large and clean sample to measure precisely. As such, we are not aware of any previous studies that documented these facts. These moments will form the basis of the more formal econometric analysis that this paper undertakes.

2.1 Dimension 1: First Four Moments

Excess Kurtosis. First, and most importantly, annual income growth displays extremely high kurtosis—ranging from 10 to 12—compared with a normal distribution, whose kurtosis is 3. (A distribution with a kurtosis of 5 or 6 is considered to be highly leptokurtic.) In plain English, this means that most individuals experience income changes that are very small (*relative* to the overall standard deviation), with few individuals experiencing very large changes. This can be seen in the left panel of Figure 1, which plots the empirical density of income changes ($y_{t+1} - y_t$) for the 2008–09 period. Notice how pointy the center is, how narrow the shoulders are, and how

[1] Haider and Solon (2006) and Kopczuk et al. (2010) focus on earlier periods (starting from the 1950s), when labor income was top coded at the SSA contribution limit (until 1978). Because this limit was very low in the 1960s and 1970s, about 2/3 of Haider and Solon (2006)'s observations are top-coded during this period.

[2] One drawback is possible underreporting (due to, e.g., cash earnings), which can be a concern at the lower end of the earnings distribution.

[3] Exceptions include Guvenen and Smith (2009), Browning et al. (2010), and Altonji et al. (2013).

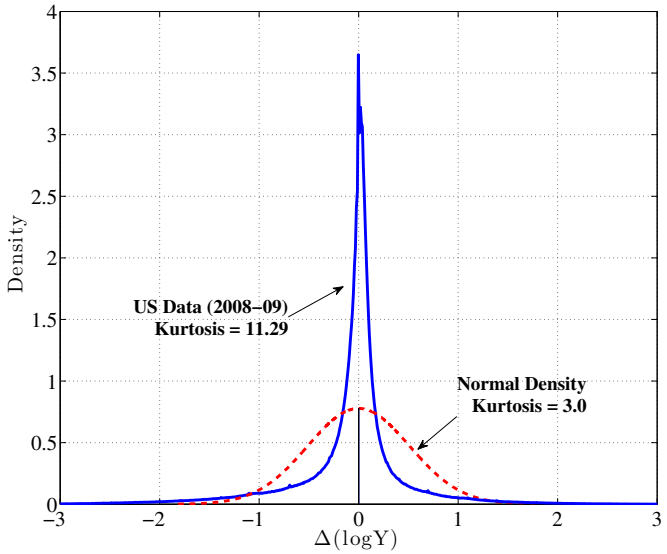


Figure 1: Histogram of Earnings Growth

long the tails are compared with a Normal density chosen to have the same standard deviation of 0.51.[4] Thus, there are far more people with very small income changes in the data compared to what would be predicted by a normal density.

An even more interesting picture emerges when we group individuals by age and (past average) income. Figure 1 plots the kurtosis of one-year income change $(y_{t+1} - y_t)$ for individuals grouped by age (25–29, 30–34, 35–39, and 40–54) and by their past 5-year average income (on the x-axis). Notice first that the kurtosis increases monotonically with past income up to the 80th to 90th percentiles for all age groups. That is, high-income individuals experience *even smaller* income changes of either sign, with few experiencing very large changes. Second, kurtosis increases with age, for every level of past income, except perhaps the top 5% of $\overline{Y}$. Furthermore, and most significantly, the peak levels of kurtosis reached ranges from a low of 20 for the youngest group, all the way up to 30 for the middle age group (40–54).

These figures represent dramatic deviations from the log-normality assumption and raises serious concerns about the current focus in the literature on the covariances (second moments) alone. In particular, targeting the covariances only (as currently done) can vastly overestimate the typical income shock received by the average worker and miss out the substantial but

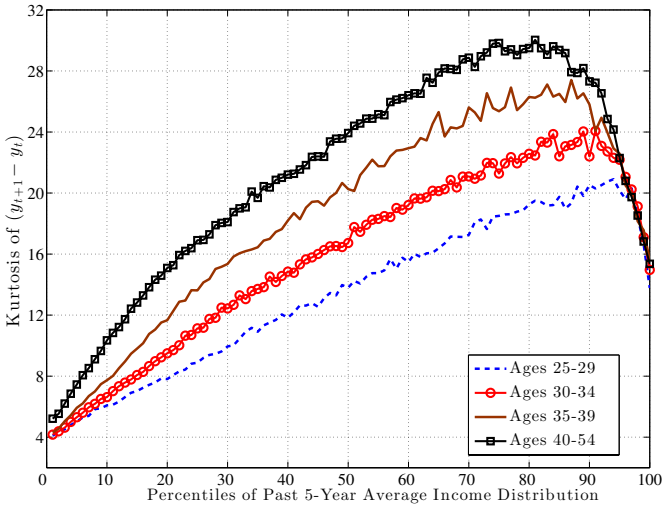
[4]To provide some concrete figures, if income changes were drawn from a normal distribution with a standard deviation of 0.51, only 7.8% of individuals would experience an income change of 5% or less; the corresponding fraction is 28% in the data. Similarly, in the data 45.1% of individuals experience a change of 10% or less (in either direction); under normal density this fraction would have been 15.4%.

Table I: Fraction of Individuals with Selected Ranges of Log Income Change

	$\text{Prob}(\lvert y_{t+1} - y_t \rvert < x)$	
$x \downarrow$	Data	$\mathcal{N}(0, 0.43^2)$
0.05	0.39	0.08
0.10	0.57	0.16
0.20	0.70	0.30
0.50	0.80	0.59
1.00	0.93	0.94

Note: The empirical distributions are all bimodal. The lower mode for the 2007–09 income growth distribution is at –3.91% (–0.47% nominal growth) and the higher one is at 2.09%.

Figure 2: Kurtosis of Annual Income Change, By Age and Past Income



infrequent jumps experienced by few.

There are well-known statistical and economic frameworks that can generate very high kurtosis. One example of a statistical model is one where income shocks follow a Poisson arrival process. Thus, income does not change regularly—most of the time there is no change—and once in a while there is a big up or down move (promotion, job loss, etc.). Alternatively, consider an economic model of job search that can be modeled as a mixture of normals: every period each worker draws a random variable which tells him whether he is going to be changing jobs or not. If he does, he draws a new income realization from a normal distribution with a large variance—and vice versa when he does not change his job. The overall income change distribution can easily be made to have very high kurtosis.

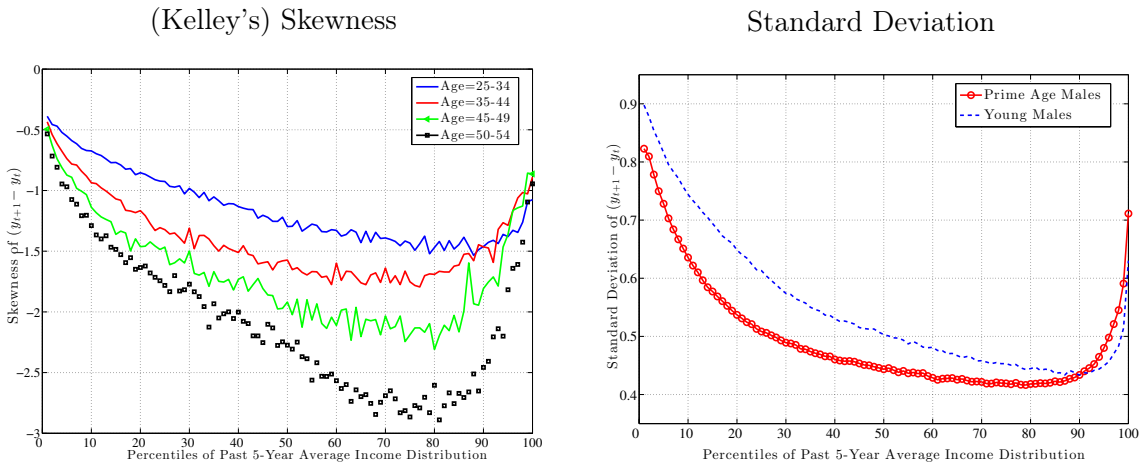


Figure 3: Skewness and Standard Deviation of Earnings Growth

Skewness and Variance. The log-normality assumption also implies that the skewness of income shocks is zero. Figure 3 (left) plots the skewness of income changes both at 1-year and 5-year horizons, conditional on past income as done above. First, notice that income shocks almost always have negative skewness (with the exception of individuals with the lowest past average income). But further, skewness becomes even more negative as we move to the right (higher income levels). Thus, it seems that the higher an individuals' past average income, the more room he has to fall down, and the less room he has left to move up. This is an insight that is modeled in many search models of the labor market, but one that is completely missed with the log-normality assumption made in the income dynamics literature. Furthermore, the magnitude of skewness is substantial.[5]

Finally, the right panel of the same figure plots the variance of income shocks as a function of past income. There is a very pronounced U-shaped pattern of smaller shocks for high income individuals (with the exception of the very top earners). Current specifications of income dynamics do not allow for such dependence and this paper models and estimates such variation.

[5]The "Kelley's measure" of skewness reported in this graph can be used to deduce the following: for the median individual as of time $t-1$ (center of the x-axis), the log 90-50 differential (the right tail) of $y_{t+5} - y_t$ accounts for 35% of the log 90-10 differential, whereas the log 50-10 differential (the left tail) accounts for the remaining 65%. This is very different from a log normal distribution which is symmetric (and therefore both tails contribute 50% of the total).

2.2 Dimension 2: Distribution of Lifetime Income

Another dimension of the data not explicitly targeted in the covariance matrix approach is the distribution of lifetime incomes. Although, this is a crucial statistic in any conceivable life-cycle model of individual behavior, it is very difficult to measure directly using PSID or other survey-based micro panels, given that it would require observing a sufficiently large set of individuals for much of their working life. The SSA dataset allows us to observe tens of thousands of individuals for 33 years, which will be used to compute lifetime incomes and its distribution accurately.[6] Finally, as seen in Figure 5, lifecycle earnings growth, here measured from ages 30 to 55 varies vary strongly by lifetime income level. Although some of this variation could be expected simply due to endogeneity, the magnitude observed here is too large to be accounted for by that channel. For example, a standard persistent-transitory model estimated in the literature (such as in Hubbard et al. (1995); Storesletten et al. (2004)) would predict that individuals in the top 1% of the lifetime income distribution should have earnings growth over the lifecycle that exceeds the median individual by only 5 percentage point. The actual gap in Figure 5 is 235 log points, which corresponds to 1050 percentage points!

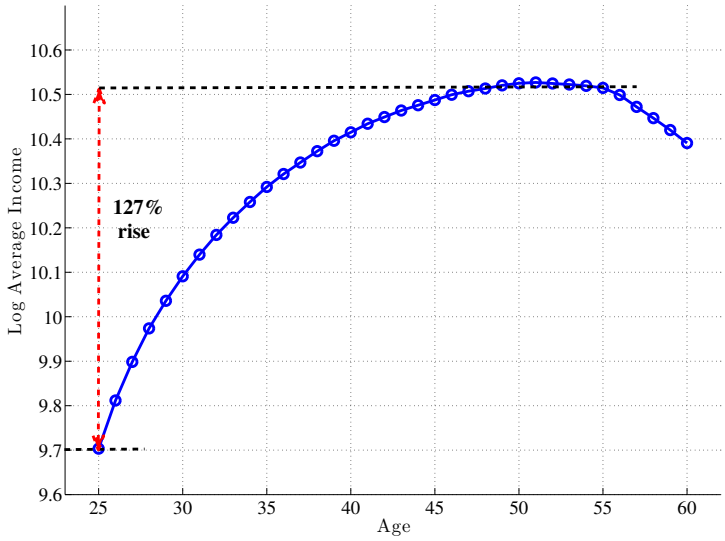
To ensure that the estimated income process captures this heterogeneity, the second set of moment we target are the average income levels at ages 25, 30, 35, ..., 55, for individuals that are in the following percentiles of lifetime income distribution: 1, 2, 3, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, 97, 98, 99, 100.

2.3 Dimension 3: Impulse Response Functions

A third key dimension of life cycle income risk is the persistence of income changes. Typically, this persistence is modeled as an AR(1) process or a low-order ARMA process (typically, ARMA(1,1)), and the persistence parameter is pinned down from the rate of decline of auto-covariances with time. The AR(1) structure, for example, predicts a geometric decline and the rate of decline is directly given by the mean reversion parameter. While this is appropriate in survey data, given the data limitations, it imposes some restrictions on the data that one might be skeptical about, such as the uniformity of mean reversion for positive and negative shocks, for large and small shocks, and so on. Here, the substantial sample size allows us to get a much higher resolution picture of the data, and in particular, characterize persistence without making

[6] In fact, the SSA also maintains the 1% LEED dataset, which covers 1957 to 2004 (used, for example, in Kopczuk et al. (2010)). This dataset can be used to construct even longer time series for each individual and compute full life time earnings.

Figure 4: Life Cycle Profile of Average Labor Income



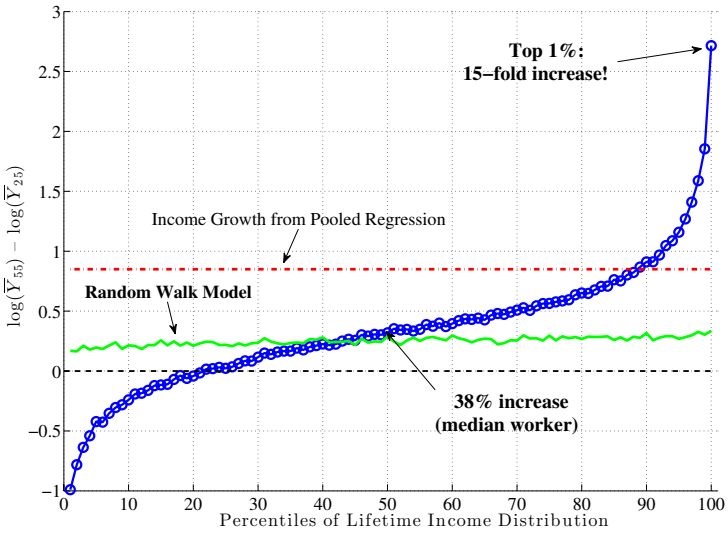


Figure 5: Lifecycle Income Growth Rates by Lifetime Income Percentile

such assumptions. To this end, we rank and group individuals based on their average income from $t-5$ to $t-1$, then within each such group, we rank and group again by the size of the income change between $t-1$ and t. Hence, all individuals within a given group obtained by crossing the two conditions, have the same average income up to time $t-1$ and experienced the same income "shock" from $t-1$ to t. For each such group of individuals, we then compute their average income change from t to $t+k$, for all values of $k = 1, 2, 3, 5, 7, 10$.

3 Estimation

With the few exceptions noted above, the current literature heavily relies on matching the covariance matrix of log income (or of the first difference of log income) in a GMM framework. The evidence outlined above strongly suggests that this approach is likely to miss important aspects of the data and produce a picture of income risk that does not capture salient features of the risks faced by workers. The current paper instead targets moments whose economic significance is more immediate, including the distribution of lifetime income, the kurtosis and skewness of income changes, as well as how these moments vary with rising incomes. These moments will then be used as targets using a method of simulated moments (or more generally, an indirect inference) estimator.

Finally, as we alluded to above, many features of the data discussed here appear to be qualitatively consistent with the outcomes of some of the new generation labor market search models. One goal would be to explore these linkages more thoroughly to see if the new evidence revealed by these rich data can shed light on competing models in this growing literature.

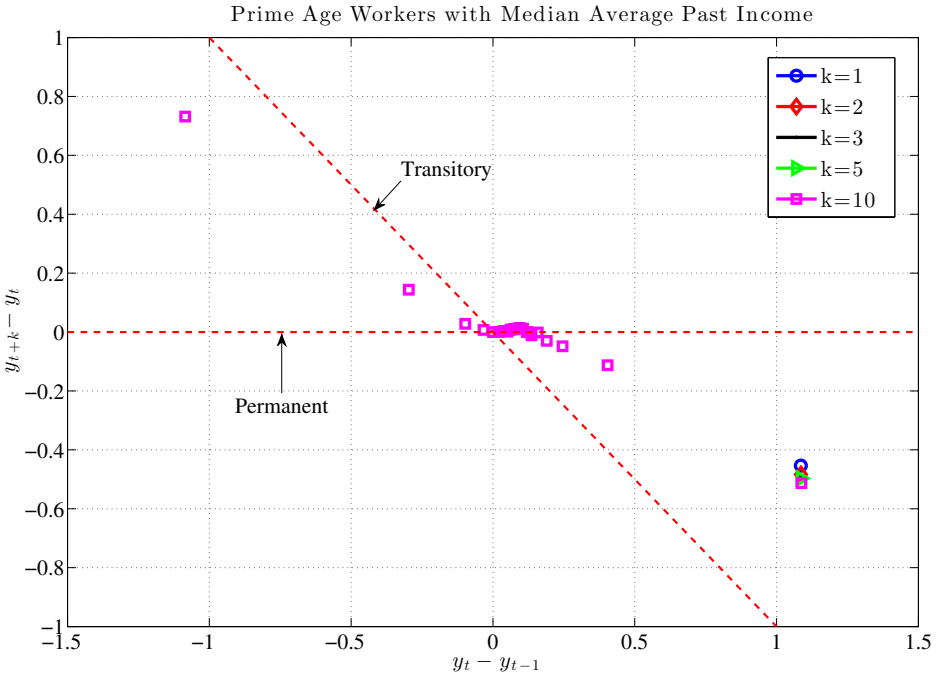
3.1 A Flexible Stochastic Process

Our goal here is to build a sequence of stochastic processes, going from relatively simple to complex, to capture increasingly more of the features outlined above. As we shall we, the most commonly estimated income processes in the literature do a fairly poor job explaining the salient features of the life cycle income data.

3.1.1 Specification 1:

The simplest process we shall consider has the following features:

Impulse Response, by Shock Size



10-Year Mean Reversion

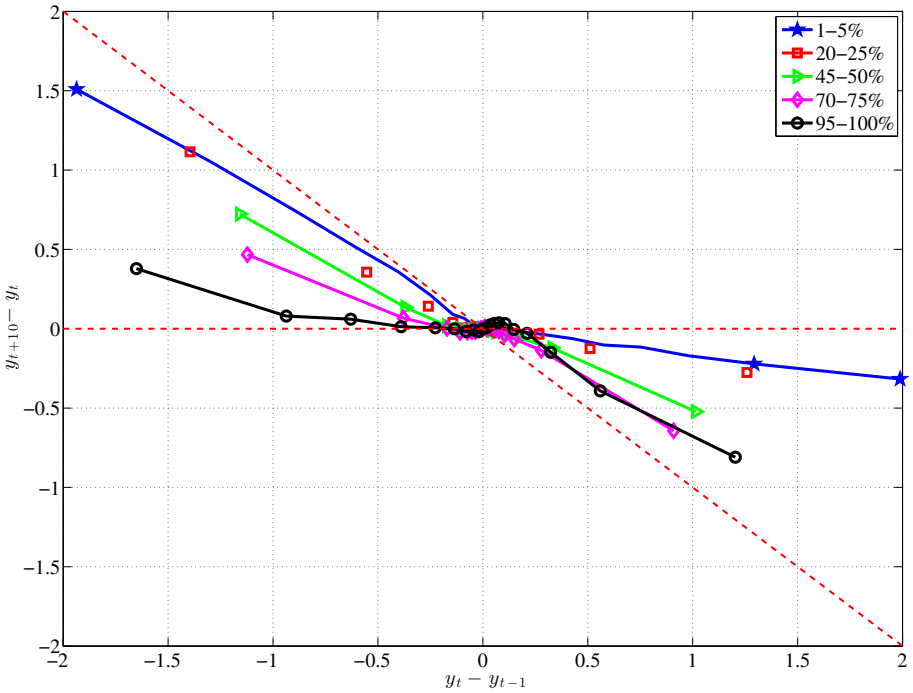


Figure 6: Impulse Responses

- Heterogeneous income profiles that are of quadratic form
- A mixture of two AR(1) processes, denoted with z and x, that vary in their persistence and innovation variance. Each AR(1) process receives a new innovation in a given year with probability $p_j \in [0,1]$ for $j = z, x$.
- An i.i.d transitory shock.

Here is the full specification:

$$\widetilde{y}_t^i = \left(\alpha^i + \beta^i t + \gamma^i t^2\right) + z_t^i + x_t^i + \varepsilon_t^i \tag{1}$$

$$z_t^i = \rho_z z_{t-1}^i + \eta_{zt}^i$$

$$x_t^i = \rho_x x_{t-1}^i + \eta_{xt}^i \tag{2}$$

where for $j = z, x$:

$$\eta_{jt}^i \quad \begin{cases} = 0 & \text{w.p } 1 - p_j \\ \sim \mathcal{N}(0, \sigma_j) & \text{w.p } p_j \end{cases},$$

To avoid indeterminacy, we impose $p_x > p_z$, which is without loss of generality.

3.1.2 Specification 2

We generalize the first specification to allow for n types of workers, who differ in (potentially) all the parameters of their income processes, except for $\overline{\gamma}$ (the average value of curvature). We have estimated specifications with $n = 2$ and $n = 3$, the latter of which worked quite well. We also estimate the fraction of each group of worker. To obtain identification we impose $\overline{\beta}^1 < \overline{\beta}^2 < \overline{\beta}^3$.

3.1.3 Specification 3

An alternative—and simpler—version of the previous specification is obtained as follows. We allow the parameters of the HIP component to vary across groups of individuals but the stochastic process determining the dynamics of income are the same for all individuals. Moreover, the groups for the different HIP process are identified by lifetime income. In particular, we choose the first group to correspond to the bottom 10% of the lifetime income distribution, the second group to correspond to the middle 80 percentiles, and the third group to correspond to the top 10 percentiles by lifetime income. These choices are motivated by the apparent non-linearity in Figure 5.

3.1.4 Specification 4:

While the specification in the previous section captured a lot of the properties of shocks in a given cross-section, it totally fails the age and income structure of shocks. There rare a number of ways this could potentially be captured. The actual specification we estimate is

$$_j(z) = a_j + b_j \times z + c_j \times t + d_j \times z \times t.$$

Here, we allow the shock variance to change with the persistent component of earnings, z, as well as with age (t) and the interaction of the two. Again, because this is a specification for variance, we must impose a positivity condition on it. The results with c_j and d_j set equal to zero are shown in Table 2.

We have also considered an alternative specification where the mixing probabilities are functions of income and age:

$$p_j(z) = a_j + b_j \times z + c_j \times t + d_j \times z \times t.$$

In this specification, the probability of each shock being realized will be changing with income. As income changes with age, this could capture both an age structure and a cross-sectional structure. Since this is a probability however, a_j, b_j, c_j must be chosen so as to ensure that p_j stays bounded between 0 and 1. The results with probability only depending on z are shown in Table 3.

Table II: Estimation Results: Specification 3

Parameters	Group 1	Group 2	Group 2
Fractions	0.1480	0.1621	0.6899
mean(α)	0.5691	0.1247	0.7304
mean(β)$\times$100	-0.3512	0	0.1786
$_\alpha$	0.6817	0.4611	0.3768
$_\beta \times 100$	0.4966	0.1280	0.0986
$_{\alpha\beta}$	-0.1831	0.5195	0.3645
p_1		0.1173	
p_2		0.8051	
ρ_1		0.2232	
ρ_2		0.6090	
$_1$		$0.65 + 0.53y_{t-1}$	
$_2$		$0.41 - 0.34y_{t-1}$	
$_\epsilon$		0.0688	

Table III: Estimation Results: Specification 3, linear probability

Parameters	Group 1	Group 2	Group 2
Fractions	0.1626	0.4172	0.4202
mean(α)	-0.0100	-0.5679	-0.2622
mean(β)$\times$100	-0.3200	0.0000	0.1379
$_{\alpha}$	0.4846	0.5698	0.3562
$_{\beta} \times 100$	0.0805	0.2340	0.0712
$_{\alpha\beta}$	0.2201	0.8662	0.0202
p_1		$0.10 + 0.16y_{t-1}$	
p_2		$0.04 - 0.49y_{t-1}$	
ρ_1		0.2895	
ρ_2		0.4607	
$_1$		0.9999	
$_2$		0.6793	
$_{\epsilon}$		0.2032	

Table IV: Estimation Results: Specification 3 (identified by lifetime income)

Parameters	Group 1	Group 2	Group 2
Fractions	0.10	0.80	0.10
mean(α)	-0.8491	0.0000	0.5773
mean(β)$\times$100	-0.4774	0.0000	0.2824
$_{\alpha}$	0.0014	0.3849	0.2172
$_{\beta} \times 100$	0.1566	0.1729	0.0613
$_{\alpha\beta}$	0.2594	-0.5068	0.4187
p_1		0.1163	
p_2		0.8396	
ρ_1		0.2064	
ρ_2		0.6434	
$_1$		$0.97 + 0.58y_{t-1}$	
$_2$		$0.17 - 0.34y_{t-1}$	
$_{\epsilon}$		0.0610	

References

Altonji, J., A. A. Smith, and I. Vidangos (2013). Modeling earnings dynamics. *Econometrica 81*(4), 1395–1454.

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

Guvenen, F. and A. A. Smith (2009). Inferring labor income risk from economic choices: An indirect inference approach. Working paper, University of Minnesota.

Haider, S. J. and G. Solon (2006). "life-cycle variation in the association between current and lifetime earnings. *American Economic Review 96*(4), 1308–1320.

Hubbard, R. G., J. Skinner, and S. P. Zeldes (1995). Precautionary saving and social insurance. *The Journal of Political Economy 103*(2), 360–399.

Kopczuk, W., E. Saez, and J. Song (2010). Earnings inequality and mobility in the united states: Evidence from social security data since 1937. *Quarterly Journal of Economics 125*(1).

Storesletten, K., C. I. Telmer, and A. Yaron (2004, June). Cyclical dynamics in idiosyncratic labor market risk. *Journal of Political Economy 112*(3), 695–717.