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

Fatih Guvenen[*] Fatih Karahan[†] Serdar Ozkan[‡] Jae Song[§]

February 15, 2013

Abstract

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.

The first part of 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. In the second part, we examine if these documented features can be explained in a standard job ladder model with learning about match quality and depreciation of skills during unemployment.

[*] University of Minnesota and NBER; guvenen@umn.edu

[†] Federal Reserve Bank of New York; yfkarahan@gmail.com

[‡] Federal Reserve Board; serdar.ozkan@frb.gov

[§] Social Security Administration; jae.song@ssa.gov

1 Extended Abstract

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.

1.1 The Data Set

This paper uses a 10% random sample of the US male population, 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 can allow 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]

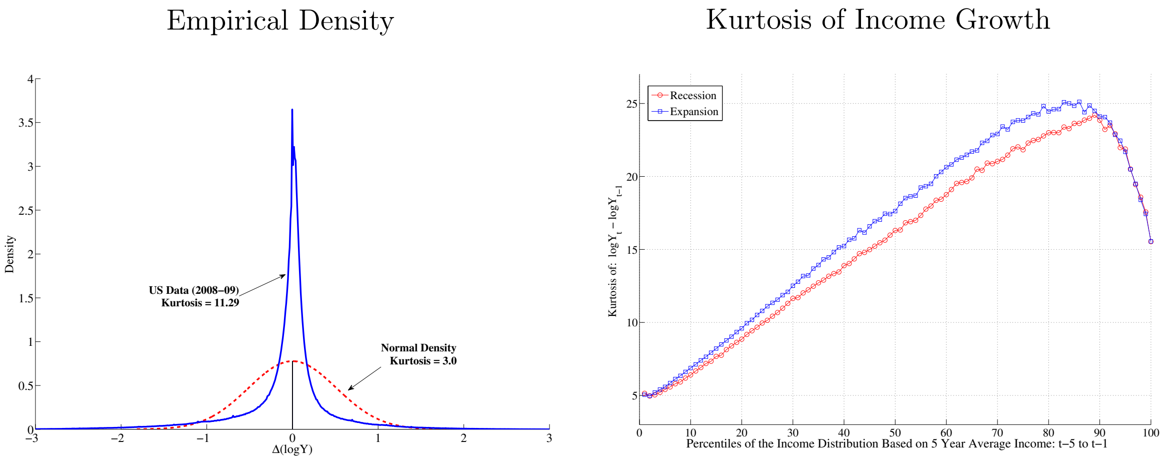
1.2 Going Beyond the Covariance Matrix: An Indirect Inference Approach

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.

[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 Altonji et al. (2010), Guvenen and Smith (2009), and Browning et al. (2010).


Empirical Density

Kurtosis of Income Growth

Figure 1: Kurtosis of Income Change Distribution

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.

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 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 control for past income. The right panel of Figure 1 plots the kurtosis of $(y_{t+1} - y_t)$ for individuals grouped by their past 5-year average income. Notice first that the kurtosis increases monotonically with past income up to the 90th

[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%.

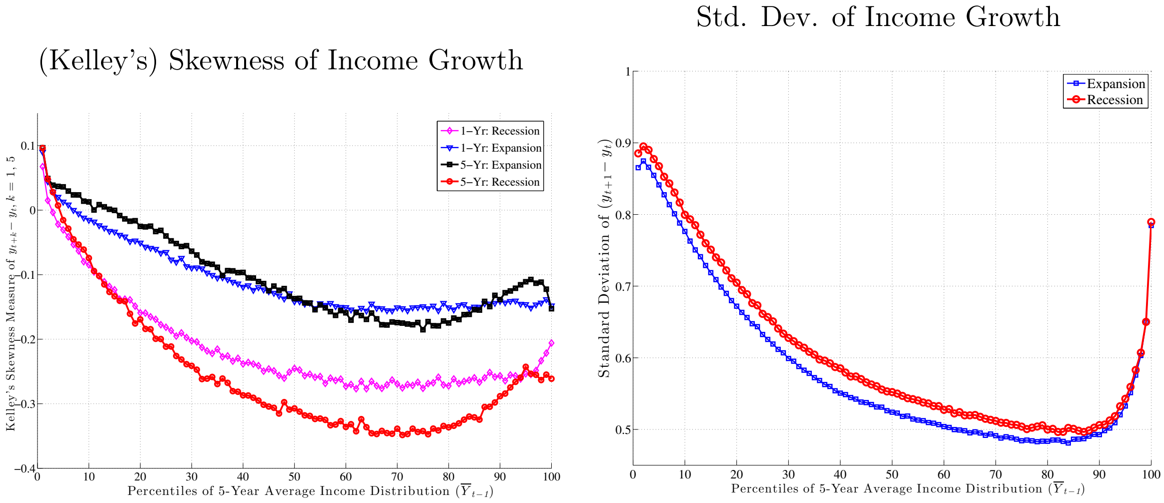
percentile, to reach a level of 25! That is, high-income individuals experience *even smaller* income changes of either sign, with few experiencing very large changes. This is a substantial deviation from the log-normality assumption and raises serious concerns about the current focus in the literature 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 infrequent jumps experienced by few.

There are well-known economic frameworks that can generate very high kurtosis. One example is a model 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, we can allow for 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.

Skewness and Variance. The log-normality assumption also implies that the skewness of income shocks is zero. Figure 2 (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).


(Kelley's) Skewness of Income Growth

Std. Dev. of Income Growth

Figure 2: Top and Bottom Ends of Wage Income Change Distribution

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 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 3, lifecycle earnings growth, here measured from age 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 account 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 3 is 235 log points, which corresponds to 1050 percentage points!

1.3 Empirical Strategy

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.

[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.

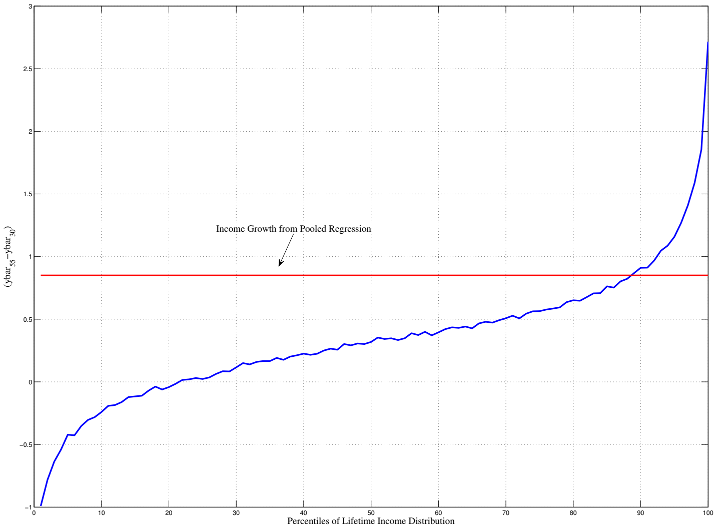


Figure 3: Lifecycle Income Growth Rates by Lifetime Income Percentile

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.

References

Altonji, Joseph, Anthony A Smith, and Ivan Vidangos, "Modeling Earnings Dynamics," Technical Report, Yale University 2010.

Browning, Martin, Mette Ejrnaes, and Javaier Alvarez, "Modelling Income Processes with Lots of Heterogeneity," *Review of Economic Studies*, 2010, *77*, 1353–1381.

Guvenen, Fatih and Anthony A Smith, “Inferring Labor Income Risk from Economic Choices: An Indirect Inference Approach,” Working Paper, University of Minnesota 2009.

Haider, Steven J. and Gary Solon, “"Life-Cycle Variation in the Association between Current and Lifetime Earnings.,” *American Economic Review*, 2006, *96* (4), 1308–1320.

Hubbard, R. Glenn, Jonathan Skinner, and Stephen P. Zeldes, “Precautionary Saving and Social Insurance,” *The Journal of Political Economy*, 1995, *103* (2), 360–399.

Kopczuk, Wojciech, Emmanuel Saez, and Jae Song, “Earnings Inequality and Mobility in the United States: Evidence from Social Security Data Since 1937,” *Quarterly Journal of Economics*, 2010, *125* (1).

Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron, “Cyclical Dynamics in Idiosyncratic Labor Market Risk,” *Journal of Political Economy*, June 2004, *112* (3), 695–717.