Idiosyncratic Income Risk Estimated From IRS Administrative Wage Data

Fatih Guvenen, Bradley Haim, Vasia Panousi, Ivan Vidangos, and Gianluca Violante*

January 22, 2010

Extended Abstract

The importance of idiosyncratic labor income risk for individuals' economic choices and, hence, for allocations and welfare is hard to overstate. The literature that relies on incomplete markets/heterogeneous agent models is not only vast, but is also continuing to expand at a rapid pace. Not surprisingly, a crucial ingredient in this research is the precise nature of income risk that researchers feed into their models. This demand for reliable information has led to the development of an active empirical literature that estimates stochastic processes for individual (labor) earnings, going back to the early seminal contributions by Lillard and Willis (1978), MaCurdy (1982), and Abowd and Card (1989), and more recent contributions by Meghir and Pistaferri (2004), Heathcote, Storesletten, and Violante (2008), Guvenen (2009), and Altonji, Smith, and Vidangos (2010). Although these studies have made important contributions to our understanding of idiosyncratic income risk, an important shortcoming is that they rely on very noisy survey data, with small sample sizes—most commonly, the Panel Study of Income Dynamics (PSID).

The goal of this paper is to shed new light on income risk by using—for the first time—a panel dataset on labor earnings obtained from administrative records that has three key advantages: (i) very large sample size with a long average time span (7 to 20 times the size of PSID samples used in the existing literature), (ii) minimal measurement error, (iii) no top-coding.[1]

First, as I elaborate in the data description below, the smallest sample we use has about 188,000 observations on approximately 19,000 individuals (compared to 10,000 to 35,000 observations on 500 to 2000 individuals, in most existing work). Furthermore, what is also critical is that more than half of all individuals in our sample (ie., 10,000+ individuals) are observed for 15 years or more, which is about *10 times larger* than the number of individuals observed for the same duration in PSID. As discussed in Guvenen (2009), a long time dimension is essential for constructing higher-order autocovariances, which are critical for identification between alternative income processes. Unfortunately, due to attrition, in most survey data sets few individuals are observed for 15 to 20 years, which has been an important obstacle for precise empirical identification in existing work.

*Author affiliations: University of Minnesota, US Department of Treasury, Federal Reserve Board, Federal Reserve Board, New York University.

[1] Although some recent studies have used different versions of US administrative data (Haider and Solon (2006); Kopczuk, Saez, and Song (2010)), these studies do not estimate stochastic processes for labor income that can be used to guide parametrizations of economic models. Schulhofer-Wohl and Norets (2010) is a partial exception: they use wage data from the Social Security Administration and provide an interesting approach by estimating a Bayesian econometric model of individual incomes. However, their data source is top-coded and (in their current version) the multi-dimensional distribution for parameters generated by Bayesian estimation does not lend itself easily to calibrating existing incomplete markets models.

To discuss some specific examples of the unresolved empirical issues, consider the following stochastic process for log labor income of individual i with h years of labor market experience:

$$y_h^i = \underbrace{g(\theta_0, X_h)}_{\text{common life-cycle component}} + \underbrace{\left[\alpha^i + \beta^i h\right]}_{\text{profile heterogeneity}} + \underbrace{\left[p_h^i + z_h^i + \varepsilon_h^i\right]}_{\text{stochastic component}} \tag{1}$$

$$p_h^i = p_{h-1}^i + \nu_h^i, \qquad z_h^i = \rho z_{h-1}^i + \eta_h^i, \qquad \varepsilon_h^i, \eta_h^i, \nu_h^i \sim i.i.d$$

This specification is fairly general (in fact, much more general than what is typically used to calibrate economic models) and, thus, provides a good framework for discussion. The first function, g, captures the life-cycle variation in labor income that is *common* to all individuals—notice the lack of i superscript in θ_0—with given observable characteristics X_h (such as age, education, etc.). The second component captures potential *individual-specific* differences in income levels and, more importantly, growth rates. Such differences would be implied, for example, by a human capital model with heterogeneity in learning ability (Huggett, Ventura, and Yaron (2007); Guvenen and Kuruscu (2009)). Finally, the terms in the last bracket represent the stochastic variation in income, which is written here as the sum of a random walk, an AR(1) component and a purely transitory shock.

A vast empirical literature has estimated various versions of (1) in an attempt to quantify the importance of different components, yet no consensus has been reached. For example, studies that allow for the second term (profile heterogeneity) find it to be quantitatively important and also estimate the stochastic component to have low persistence (for example, $\nu_h = 0$ and $\rho \ll 1$). Other studies set $\beta^i \equiv 0$ *a priori*, and, with this restriction, they estimate a nearly unit root shock process (e.g, $\sigma_\nu \gg 0$ and $\rho = 0$ or 1). So, after decades of research, the literature has still not been able to provide an unequivocal answer to a seemingly very simple question: what is the persistence of income shocks? But clearly, the answer is crucial for a range of economic questions. Furthermore, and as can be evident from this discussion, decomposing the fraction of the stochastic component that is a unit root versus a persistent process has proved almost hopeless. Clearly, this lack of conclusive evidence is not a reflection of researchers' diligence or competence; rather it is a direct consequence of the strong limitations imposed by the PSID data set, in terms of sample size, attrition patterns, and significant measurement error that may not be of classical form.

The present paper aims to make significant progress on these (and related) questions, thanks to the substantial sample size as well as high signal-to-noise ratio offered by the administrative data we have available. The goal is to estimate several variants of the specification given in (1) as well as some richer variants that have been considered in past work but have then been discarded due to the difficulty of obtaining precise estimates with the small PSID samples. We hope that this paper can provide a reliable "user guide" for quantitative economists who can find precisely estimated parameters of income specifications most suitable for their model, ranging from the very simple to the very rich ones.

A second important advantage of our data set is that it has no top-coding, so every income level is recorded no matter how high they may be. In contrast, for example, Haider and Solon (2006) and Kopczuk, Saez, and Song (2010) focus on earlier periods (starting from the 1950s) and obtain their data from the Social Security Administration, which only records incomes up to the social security contribution limit. 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. Kopczuk, Saez, and Song (2010) have been able to obtain non-top-coded data for one quarter each year. In any case, these studies focus on different questions than those this paper aims to address.

Finally, there is strong reason to suspect that the answers this paper will obtain from the administrative data could be substantively different from those obtained from PSID. For example, a large number of studies have documented from the PSID that since the late 1970s income inequality rose because of a rise in permanent inequality in the 1980s, followed by a rise in transitory inequality in the 1990s. In contrast, Kopczuk, Saez, and Song (2010) conclude from administrative records data that earnings mobility has remained virtually unchanged during the same period, which is hard to square with the PSID results that would imply a change in mobility during this period.

Data Description

The main data source for this paper is the IRS Continuous Work History Subsample (CWHS) and includes detailed information about labor earnings as well as non-labor earnings at the *individual level*, as well as demographic information on age and gender. It should be stressed that these data are not publicly available. We are able to access this dataset through our collaboration with a researcher in the Treasury Department, Bradley Heim.

We plan to use a twenty-year panel of tax returns that spans the period 1987-2006. This panel is created by merging tax returns from an existing panel, known as the 1987-96 Family Panel, with returns from cross-sectional files from 1997-2006. The 1987-96 Family Panel was collected by the Statistics of Income (SOI) division of the IRS, and started with a stratified random sample of taxpayers who filed in 1987. Over the following nine years, any return filed that reported any panel member as a primary or secondary taxpayer were included in the sample, including tax returns filed by panel members who were dependents of another taxpayer. In addition to information from each taxpayer's Form 1040, the dataset includes information on age and gender of the primary and secondary filers, information on wages and contributions to employer-based retirement plans from W-2 forms, and information on contributions to tax-preferred savings accounts from form 5498.

The 1997-2006 data come from yearly cross-sections that are also collected by the SOI. Like the 1987-1996 sample described above, in each of these years a stratified random sample was collected consisting of a strictly random sample based on the last four digits of the primary filer's SSN and a high income oversample. Over these years, the size of the strictly random sample grew—consisting of two four-digit endings in 1997, five endings from 1998-2005, and ten endings starting in 2006. Each cross-section contains information from the taxpayer's Form 1040 and a number of other forms and schedules. To these data, we merged in information on age and gender of the primary and secondary filers, information on wages and contributions to employer-based retirement plans from W-2 forms, and information on contributions to tax-preferred savings accounts from Form 5498. The merged dataset represents a one-in-5,000 random sample of taxpayers followed over 1987-2006. The panel is not balanced, and so some taxpayers drop out of the sample due to death, emigration, or falling below the tax filing thresholds, while others enter because of immigration or becoming filers.

After the sample has been cleaned up, the primary sample has about 188,000 observations over 20 years. This is substantially larger than the typical PSID samples. For example, MaCurdy (1982)'s analysis was based on about 10,000 observations; Abowd and Card (1989) had about 15,000; even more recent studies by Guvenen (2009) and Heathcote, Storesletten, and Violante (2008) have about 35,000 observations. Moreover, these PSID samples are marred by pervasive measurement error, which is a much minor problem in the IRS administrative data.

References

Abowd, J. M., and D. Card (1989): "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57(2), 411–45.

Altonji, J., A. A. Smith, and I. Vidangos (2010): "Modeling Earnings Dynamics," Discussion paper, Yale University.

Guvenen, F. (2009): "An Empirical Investigation of Labor Income Processes," *Review of Economic Dynamics*, 12(1), 58–79.

Guvenen, F., and B. Kuruscu (2009): "A Quantitative Analysis of the Evolution of the U.S. Wage Distribution: 1970-2000," *NBER Macroeconomics Annual*.

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.

Heathcote, J., K. Storesletten, and G. L. Violante (2008): "The Macroeconomic Implications of Rising Wage Inequality in the United States," NBER Working Papers 14052, National Bureau of Economic Research, Inc.

Huggett, M., G. Ventura, and A. Yaron (2007): "Sources of Lifetime Inequality," NBER Working Paper 13224.

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).

Lillard, L. A., and R. J. Willis (1978): "Dynamic Aspetcs of Earnings Mobility," *Econometrica*, 46(5), 985–1012.

MaCurdy, T. E. (1982): "The use of time series processes to model the error structure of earnings in a longitudinal data analysis," *Journal of Econometrics*, 18(1), 83–114.

Meghir, C., and L. Pistaferri (2004): "Income Variance Dynamics and Heterogeneity," *Econometrica*, 72(1), 1–32.

Schulhofer-Wohl, S., and A. Norets (2010): "Heterogeneity in Income Processes," Discussion paper, Princeton University.