Discussion of "Trends in Individual Earnings Variability and Household Income Variability Over the Past 20 Years"

(Dahl, DeLeire, and Schwabish; draft of Jan 3, 2008)

Austin Nichols

Jan 4, 2008

Broad Comments

- Very useful paper: adds to discussion of potentially important trend in income variability.
- Particularly valuable are comparisons of estimates using only wage and salary income to estimates including self-employment, and estimates using restricted samples (without imputations).
- Estimates differ from others' contributions: a valuable addition would be a step-by-step comparison of methods, so the source of large differences in estimates can be more directly traced to choices made by various authors.

Risk or Variation?

Much of the literature cited concerns "transitory variation" or short-term income risk. The desideratum is usually to estimate individual or family income risk at a point in time. The authors, as do others, reframe this as a problem of measuring variability over time in income, but without a discussion of what this means for conclusions—the estimation of income changes, their causes, and consequences, is another thread in a related literature, e.g. Burkhauser and Duncan (1989).

Can't measure income risk, nor the welfare consequences of risk, nor the welfare consequences of income changes, and these all differ substantially from the incidence of 50% income changes. Authors mention cost of living and work/home production tradeoffs, but there are more fundamental questions about what we would like to measure if we could—and why. It would be helpful to discuss these issues, at least briefly.

Importance of Definitions

General approach of this literature is to compute some measure of variance for an individual case (person or household) at some point in time using lags and/or leads, then compute the central tendency across individual cases; but the distribution of these individual measures is often quite skewed or even bimodal.

Choice of definition for "percentage change" and for "large changes" (authors report the incidence of 50% declines and increases) matters. Definition of percentage change used by these authors is actually twice the coefficient of variation for each individual, using two observations of annual income, with the sign of the change in income:

$$c_i = 2(y1_i - y0_i)/(y1_i + y0_i) = 2[\mathrm{sgn}(y1_i - y0_i)]\widehat{\sigma}_i/\widehat{\mu}_i$$

for two years 0 and 1. For $c = 50\%$ (seemingly a 50% increase in income), y1 must be 1.67 times y0, and for a seeming 50% decrease ($c = -50\%$), income y1 must be 0.60 times y0 (a 40% decrease).

Other Definitions

The coefficient of variation has at least one feature to recommend it: symmetry. A decrease from 5 to 3 gives $c_i = -0.5$ and a subsequent increase from 3 to 5 gives $c_i = 0.5$, which is not the case for percentage changes (-40% followed by $+67\%$ does not average zero).

But it is not percentage change, and it differs from other measures of variability in common use:

- E[sd(Y)]: A case-specific sd is computed over some vector of individual income values Y, e.g. five years of annual data, then one takes the mean (or better, quantiles) over individual cases. Note that sd(Y) for a two-year window is proportional to the absolute value of the change in income ($|\Delta Y|$), and the sd is not invariant to scale (so deflating by different price indices will create different results). Variants: E[sd(ln(Y))] The sd is computed over some vector of logged individual income values ln(Y), to measure changes approximately in percentages; E[sd(e)] computed from residuals from a regression of ln(Y) on e.g. age categories, to measure "unexpected" changes.
- E($|\Delta Y|/Y$): Compute percent/proportional changes e.g. in annual income between a pair of years, then take the mean (or quantiles) over individuals. Often with truncation; e.g. Dynan Elemendorf and Sichel (2007) topcode proportional gains at 1 (doubling income), and assign 1 to increases from zero (where any gain is positive infinity percent).
- Var-Cov: Variance minus covariance, or regression of covariances on lag length. Compute the variance of income across individuals and subtract off covariance of own income, in the simplest formulation. Relies on assumptions not supported by data (lognormality of income, no measurement error, and no serial correlation of errors) for theoretical justification.

Volatility or Inequality

Taking the mean of sd(ln(Y)) and other methods involving ln(Y) founder on what to do with zero or negative income. One cannot simply truncate the income distribution, since the bias from such a truncation could be enormous (and can be estimated by trying various truncation points or assuming in turn various extreme possibilities à la Manski). However, note that sd(ln(Y)) is a measure of inequality, in this case the inequality of a person's (or family's or household's) income realizations over some period of time. Other measures of inequality can serve a similar purpose, e.g. some members of the Generalized Entropy class, the Atkinson class, and the Gini coefficient.

Regression Using Logs

Assuming percentage change is of interest, we would like a symmetric transformation. Consider the class of models that look at changes across pairs of years—we would like a halving and doubling of income to be treated similarly—but reductions in proportional terms are on the interval of length one $[-1, 0)$ and increases on the infinite interval $(0, \infty)$ so means over such quantities treat them very asymmetrically.

If a halving to 50% of y and a doubling to 200% of y are treated symmetrically, then one possible symmetric function of the proportional change $d = \Delta Y/Y$ is in the family ln(1+d). Clearly doubling and halving are symmetric since ln(.5) = -ln(2). But ln(1+d) is equivalent to the log of the ratio of the two values, e.g. $\ln(y_t/y_{t-1})$. Using panel data to compute year-specific values of the mean of this measure across individuals is equivalent to regressing $\ln(y_t/y_{t-1})$ on year dummies. This is equivalent to the regression model

$$\ln(y_{it}) = \gamma \ln(y_{i(t-1)}) + X_i\beta + e_{it}$$

where X_i is a vector of time dummies with $X_{is} = 1$ for $s = t$ and 0 otherwise, and we constrain the coefficient on the lagged dependent variable to be one, i.e. $\gamma = 1$. Here, it is clear that we are measuring mean income gains year by year, which may have little to do with income volatility at a point in time.

Measured Variation

On the whole, the coefficient of variation has much to recommend it, but it is very sensitive to mismeasurement or changes in the mean that is its denominator.

More importantly, it is unclear why we want the standard deviation of the individuals' coefficients of variation as our omnibus measure of variability. The mean or median of the absolute value is more in line with other measures. But the distribution is skewed, asymmetric, and otherwise not amenable to a single statistic characterizing the distribution of changes. A comparison of density estimates seems preferable, at least in a technical appendix justifying the approach used.

Data Trims

Effects of trims: "Households in the bottom or top 2 percent of the income distribution in either the first or second year were dropped from the sample." These cases are not necessarily those with the largest swings in income, but they tend to be, and the trim will automatically bias the estimation of variability. Could leave in the extremes of income in each year, and report median of the change/volatility measure (c_i, sd, etc.) instead. Or even better, several measures of the distribution.

The authors also drop cases with zero in both years—but any other case with income unchanged has $c_i = 0$. Would like to see that this choice makes no difference, as the more natural avenue assigns $c_i = 0$ to cases with zero income in both years.

Types of Income Included

Covered earnings only for individual income: "In 1985, 93 percent of wage and salary work was in the covered sector; by 2002, that percentage had rise[n] to 96 percent" but the relevant number is the proportion transitioning between covered and noncovered employment. This proportion can be estimated using matched SIPP-SSA data, where reports of covered earnings can be compared to total earnings, and variation over time in this ratio can be computed for individuals.

Taxes also could be included with relatively little effort, and the transition from AFDC to EITC seems important to the volatility story (especially given the timing of large increases in volatility in the early 1990's in many estimates).

Matched Data

Choice of data for HH income: "Using the matched SIPP-SSA data, however, is somewhat limited in that these data are available at only four points during the 1984 to 2003 period: 1984/1985, 1993/1994, 1997/1998, and 2001/2002." Actually, matched SIPP data is available for all but a few months of 2000, using the 1984 to 1993, 1996, and 2001 panels. More detail on the data used and selection rules would be helpful.

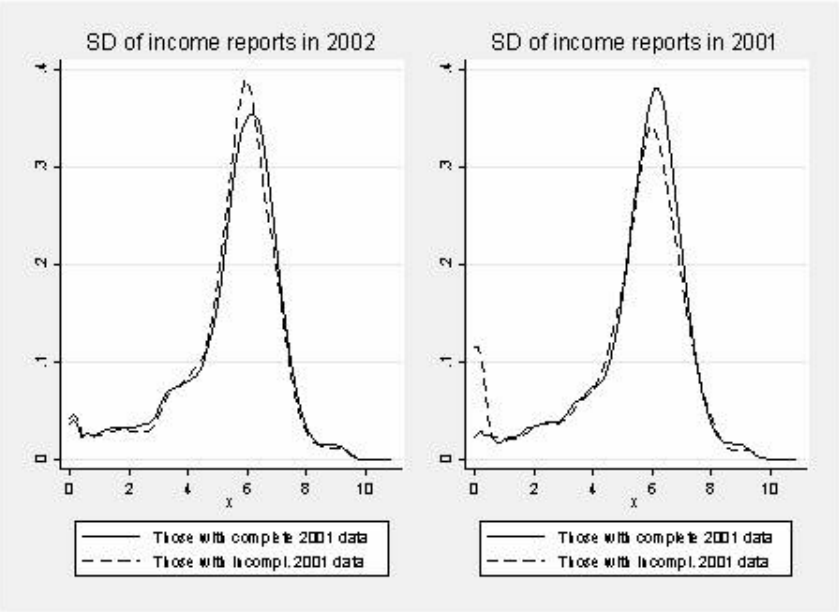
Missing Data

Match rates for SIPP: "The match rate for each SIPP panel was 85 percent in 1984, 81 percent in 1993, 78 percent in 1996, and 57 percent in 2001." Using such a sample restriction could produce nearly any trend or absence thereof.

Imputation rates also increased due to refused interviews and attrition during the SIPP ramping up across panels. The authors find that "Among all households (including those with imputed data), the variability in household income (as measured by the standard deviation of the percentage change in household income) increased by 21 percent over the 1994 to 2003 period." After dropping any household with any imputed earnings, they find "no change in total household income volatility over the 1984 to 2003 period." Their "analysis of the matched SIPP-SSA data, in contrast to [their] analysis of the survey data alone, did not show a differential trend in variability among households with imputed income data and households with non-imputed income data." Hard to know what is due to what source of missing data.

Monthly Variation

Here is the log of one plus the SD of monthly income reports in the 2001 SIPP for two years; the bulk of the distribution lies between 4 and 8 log dollars, or 50 and 3000 dollars, but the distribution is clearly not normal. A similar summary of the coefficient of variation for pairs of years would be a valuable addition and provide a link from bar graphs showing percent with less than a 25% change to more detailed statistics.



Incomplete Data

A similar pattern is true if you condition on complete income data for 2002 and look at 2001/2002 estimates of variability in month-to-month income reports. Those who have complete or incomplete data look similar at another time, but in the year they have incomplete data are much more likely to have no variability in reported income (sd=0), which may reflect merely the possibility that they report 1 month of data, or 4 months of the same income value. Might be useful to examine the data available within a year more closely, rather than throwing out any data with any imputations in any months.

One cannot condition on other observed income to impute the missing values; plausible assumptions are that all those who are unmatched or suffer attrition from the panel have either zero variation or more than the most extreme observed. Putting each of these assumptions, in turn, into practice for all missing values leads to credible bounds on estimates, particularly for measures of the central tendency of variability that are less sensitive to outliers.

Data Errors Abound

Administrative data and survey data both error-prone. Cannot drop unmatched or imputed cases without introducing bias. Missings not MCAR, not MAR. Folks who experience an income change are more likely to have missing data.

- Imputation and survey nonresponse: imputations typically designed to improve estimation of conditional means, not higher moments.
- Improper matching: of records over time, or of admin records to survey records. Often admin-survey matches contain individuals who have very different age and sex values in the two sources, or are dead in one data source and working in another.
- Incompatible answers: within one data source, some data is incompatible (and is often then imputed somewhat arbitrarily), and across data sources, data is incompatible-impossible to know which is the more correct entry in many cases (though the presumption is usually in favor of admin sources, and more closely audited sources in that family, evidence is mixed).
- Incomplete coverage 1: Self employment, informal labor markets, inter vivos transfers
- Incomplete coverage 2: Hard-to-survey populations may not be accurately represented in admin records either.

Conclusions, part one

Different kinds of measurement errors produce different kinds of bias—would be nice to say something about trends in various types of errors in the various data sources. Estimates of transition rates into and out of self-employment, informal labor market, etc. could generate plausible bounds on coverage bias. Imputing extremes for unmatched or otherwise missing data could generate credible bounds for imputation/matching bias.

In general, one would like to see step-by-step comparison of different results, e.g. Somebody (2007) finds increasing volatility using the same data, and their analysis differs in these five ways, decomposed into these steps, four of which contribute roughly 10% of the difference and one roughly 60% of the difference in measured trends. Then we can see what the important differences are, and judge whether that choice *should* be substantively important and how we think it should matter.

Summary

- Measurement: Volatility is not observed, and measurements relate poorly to target
- Interpretation: Income volatility/change have no direct welfare consequence, but both clearly important
- Robustness within data: Some findings very sensitive to method, some less so: individual results shown to be less robust than family/household income
- Robustness across data: If one takes the estimate of the trend in income volatility by Dahl, DeLeire, and Schwabish of zero (or close to it) as a lower bound and Hacker's estimate of a tripling as an upper bound, most of the literature lies much closer to DDS. But almost all reject zero, and find an increase in volatility. Even if we take the range of DDS estimates as a credible interval, it runs from a small reduction to a 21% increase in the volatility of family income.