Notes on Estimating Earnings Processes

Christopher Tonetti*

New York University

March 11, 2011

This note describes how to estimate earnings processes commonly used in macro-labor economics. The approach will be to formulate a statistical model that describes earnings and to estimate it using only earnings data. The main focus is on data preparation, standard model specifications, parameter identification, and standard estimation routines. Measurement error, time variation in parameters, popular alternative model specifications, and alternative estimators will also be discussed.

*This document draws heavily from class notes of Gianluca Violante's Advanced Macro course. Some sections rely on material from Nakata and Tonetti (2011).

1 Introduction

Understanding individual income risk is essential to modeling consumer behavior, designing economic policy, and comparing economies over time or across countries. For most individuals, labor earnings are the primary source of income. Hence, an extensive literature has developed estimating various idiosyncratic labor income processes both in the labor and macro fields. As heterogeneous agent incomplete market macroeconomic models continue to grow in popularity, it has become increasingly important for economists to appropriately model the labor income risks agents face. The earnings process, with the assumption of incomplete markets, delivers the heterogeneity in Bewley models and characteristics of the process determine agent behavior, both over calendar time and over the life cycle.

There exists a large and crowded literature in labor and macroeconomics estimating individual labor earnings processes. Dating back to Lillard and Willis (1978), Lillard and Weiss (1979), MaCurdy (1982), and Abowd and Card (1989), there is a history of fitting ARMA models to panel data to understand the labor income risk facing individuals. Many models assume labor income is the sum of a transitory and persistent shock, where often the persistent shock is assumed to follow a random walk.[1] Some models allow for heterogeneity, either in income profiles or more pervasively in shock distributions and ARMA parameters.

1.1 Earnings Process Estimation Strategies

The first major choice in estimating a process for labor earnings is to decide whether to use data on consumption or restrict the estimation to using only earnings data. Models of consumption, whether statistical or structural, have strong predictions on how consumption should respond to earnings shocks. It is therefore feasible to use data on changes in consumption to gain inference on the earnings process. The main advantage of this strategy is to introduce more data, but the main drawback is that the estimation relies heavily on the proposed model of consumption.

An alternative is to only use earnings data. Free from any structural modeling assumptions, specify and estimate a statistical process for earnings. This is the approach covered in these notes.

1.2 Data

For the estimation we require a panel of earnings data. A repeated cross section is not enough. In the U.S., this leads many people to use the Panel Study of Income Dynamics (PSID). Most studies apply exclusion restrictions to the data to remove outliers and achieve a more homogeneous population. Following common practice, we will think of an individual in the sample as a male head of household between the ages of 25 and 60 with non-zero annual labor earnings. Data are annual. Unfortunately, the panel is short, i.e., the PSID has a significantly larger cross sectional dimension than time dimension.

2 Model Specification

First we want to remove the predictable components of labor earnings. Then we specify a process for residual earnings.

2.1 Obtaining Residual Earnings

Assume a competitive model in the labor market, yielding a wage per efficiency unit of labor, w_t. Let i index individual, j age, and t time.

$$Y_{i,j,t} = w_t \exp(f(X_{i,j,t}) + y_{i,j,t})\bar{h} \tag{1}$$

- $Y_{i,j,t}$ - measured annual disposable labor income
- $\bar{h}$ - exogenous number of hours worked[2]

[1] MaCurdy (1982), Abowd and Card (1989), Gottschalk and Moffitt (1994), Meghir and Pistaferri (2004), and Blundell, Pistaferri, and Preston (2008) all assume a unit root in the persistent component.

[2] With an elastic labor supply, estimate a wage process by using earnings divided by hours.

- $\exp(f(X_{i,j,t}))$ - predictable individual labor efficiency
- $X_{i,j,t}$ - demographic observables and predictable variables [age, gender, edu, time dummy, etc.]
- $y_{i,j,t}$ - stochastic component of earnings
- f - time invariant function of observables $X_{i,j,t}$

Note:

$$\ln Y_{i,j,t} = \beta_t + f(X_{i,j,t}) + y_{i,j,t} \tag{2}$$

where β is the price of labor.

To complete the first step, run a regression on Equation 2 to obtain residuals $y_{i,j,t}$.

2.2 Parameterize Residual Earnings Process

After obtaining residual earnings, we need to specify a statistical model for log earnings. For example, we can choose the commonly used time invariant model from Storesletten, Telmer, and Yaron (2004a).

$$\begin{aligned} y_{i,j} &= \alpha_i + \eta_{i,j} + \epsilon_{i,j} && (3)\\ \eta_{i,j} &= \rho\eta_{i,j-1} + \nu_{i,j} && (4) \end{aligned}$$

where

$$\alpha \sim (0, \sigma_\alpha^2), \quad \epsilon \sim (0, \sigma_\epsilon^2), \quad \nu \sim (0, \sigma_\nu^2), \quad var(\eta_{i,-1}) = 0$$

and

$$\alpha_i \perp \epsilon_{i,j} \perp \nu_{i,j}, \quad \text{i.i.d}$$

Finally, group all parameters to be estimated into $\theta = \{\rho, \sigma_\alpha^2, \sigma_\epsilon^2, \sigma_\nu^2\}$.

3 Identification

Define the cross-sectional moment $m_{j,n}(\theta)$ between agents of age j and n:

$$\begin{aligned} m_{j,n}(\theta) &= \mathbb{E}[y_{i,j} \cdot y_{i,j+n}] \\ &= \mathbb{E}[(\alpha_i + \eta_{i,j} + \epsilon_{i,j}) \cdot (\alpha_i + \eta_{i,j+n} + \epsilon_{i,j+n})] \\ &= \begin{cases} \sigma_\alpha^2 + \sigma_\epsilon^2 + \sigma_\nu^2 & \text{if } j = n = 0 \\ \sigma_\alpha^2 + \rho^n \sigma_\nu^2 & \text{if } j = 0, n > 0 \end{cases} \end{aligned}$$

Formal Identification: The Autocovariance Function

1. The slope identifies ρ:
   $\frac{m_{03}-m_{02}}{m_{02}-m_{01}} = \frac{\sigma_\alpha^2+\rho^3\sigma_\nu^2-\sigma_\alpha^2-\rho^2\sigma_\nu^2}{\sigma_\alpha^2+\rho^2\sigma_\nu^2-\sigma_\alpha^2-\rho\sigma_\nu^2} = \frac{\rho^2(\rho-1)}{\rho(\rho-1)} = \rho$
2. The difference identifies σ_ν^2:
   $m_{02} - m_{01} = \sigma_\nu^2 \rho\,(\rho - 1)$
3. The level of the covariance at $n > 0$ identifies σ_α^2:
   $m_{01} = \sigma_\alpha^2 + \rho\sigma_\nu^2$
4. The variance identifies σ_ϵ^2:
   $m_{00} = \sigma_\alpha^2 + \sigma_\nu^2 + \sigma_\epsilon^2$

We have full identification from the autocovariance function. Obviously, the model is overidentified.

Note:
There exist two prominent identification strategies used to create moments for estimation. Quoting from Heathcote, Perri, and Violante (2010): "The first, common in labor economics (e.g., Abowd and Card (1989), Meghir and Pistaferri (2004), Blundell, Pistaferri, and Preston (2008)), uses moments based on income growth rates (first-differences in log income). The second, more common in macroeconomic applications (e.g., Storesletten, Telmer, and Yaron (2004b), Guvenen (2007), Heathcote, Storesletten, and Violante (2010)), uses moments derived from log income levels. Although either approach can be used to estimate the permanent-transitory model described above, they differ with respect to the set of moments that identify the structural parameters." If the model was properly specified, these two estimators should yield similar results. We can perform a specification test to formally test the model by examining the difference between two consistent estimators.

4 Estimation

The standard estimation strategy in the literature is to use a Minimum Distance Estimator. The goal is to choose the parameters that minimize the distance between empirical and theoretical moments. As discussed in Section 3, we will use the covariance matrix as our moments. Recall from Section 2.1, income data is residual earnings. Let

- $m_{j,n}(\theta) :=$ covariance of earnings between age j and n individuals
- $\widehat{m}_{j,n} :=$ empirical counterpart of $m_{j,n}$
- $\lambda_{i,j,n} := \begin{cases} 1 & \text{if } i \text{ is present at } j \textbf{ and } j+n \\ 0 & \text{o/w} \end{cases}$

then the moment conditions are

$$\mathbb{E}\left[(\lambda_{i,j,n})\left(\widehat{m}_{j,n} - m_{j,n}(\theta)\right)\right] = 0 \tag{5}$$

where

$$\widehat{m}_{j,n} = \frac{1}{I_{jn}} \sum_{i=1}^{I_{jn}} \widehat{y}_{i,j} \cdot \widehat{y}_{i,j+n}$$

The moments can be expressed as a symmetric matrix:

$$\bar{m}(\theta) = \begin{bmatrix} m_{0,0} & m_{0,1} & \cdots & m_{0,n} & \cdots & m_{0,J} \\ m_{1,0} & m_{1,1} & & & & m_{1,J} \\ \vdots & & \ddots & & & \vdots \\ & & & \ddots & & \\ m_{n,0} & & & & & m_{n,J} \\ \vdots & & & & m_{J-1,J-1} & \vdots \\ m_{J,0} & \cdots & \cdots & m_{J,n} & \cdots & m_{J,J} \end{bmatrix}$$

Finally, define $\bar{M} = vech(\bar{m})$, the stacked vector of unique observations, with length $(J+1)(J+2)/2$. The estimated parameters, $\widehat{\theta}$, are the solution to

$$\min_{\theta} \left[\widehat{\bar{M}} - \bar{M}(\theta)\right]' W \left[\widehat{\bar{M}} - \bar{M}(\theta)\right] \tag{6}$$

where W is a weighting matrix.

4.1 Weighting Matrix

To implement the estimator, we need to choose W. Altonji and Segal (1996) show that the Optimal Minimum Distance (OMD) estimator, where W is set to the optimal weighting matrix, introduces significant small sample bias. Many papers in the literature use the Equally Weighted Minimum Distance (EWMD) estimator, where W is the identity matrix, as a result. An alternative, employed by Blundell, Pistaferri, and Preston (2008) is to use Diagonally Weighted Minimum Distance (DWMD), where W is set to the diagonal elements of the optimal weighting matrix with off-diagonal elements set to zero.

4.2 Standard Errors

Chamberlain (1984) shows standard errors can be obtained as

$$\widehat{var(\widehat{\theta})} = (G^{'}WG)^{-1}G^{'}WVWG(G^{'}WG)^{-1} \tag{7}$$

where V^{-1} is the optimal weighting matrix and G is the Jacobian matrix evaluated at the estimated parameters, $\frac{\partial \bar{M}(\theta)}{\partial \theta}|_{\theta=\widehat{\theta}}$. Recall the data were originally obtained as residuals from a first stage regression. See Murphy and Topel (1985) for adjusting second stage standard errors.

Alternatively, standard errors can be computed by bootstrap. Bootstrap samples are drawn (with replacement) at the household level with each sample containing the same number of households as the original sample. Then apply the first stage regression on each sample, estimate the parameters of interest on the residual for each sample, and compute statistics using cross-sample variations. The resulting confidence intervals thus account for arbitrary serial dependence, heteroskedasticity, and additional estimation error induced by the use of residuals from the first stage regressions. Bootstraping is a computationally intensive technique. Run as many samples as is computationally feasible, with a rule of thumb being 500.

5 Transitory Effects

5.1 Measurement Error

Micro data, especially those based on surveys, have measurement error, $\tau_{i,t}$. The typical assumption is that it is i.i.d across agents and over time. With this assumption, measurement error is indistinguishable from ϵ in our specification. Econometricians should thus be aware when interpreting parameter estimates of the transitory component. The transitory component of earnings could be modeled as an MA(q), with q >0, in which case the variance of the transitory component can be estimated separately from classical measurement error.

The PSID ran validation studies in 1982 and 1986 where they confirm the earnings and hours data from employer records. They found a small error in earnings (10-20 percent) but larger error in hours worked (20-40 percent). Depending on the question, measurement error may not be that important because so much action comes from the fixed and persistent effects, α_i and η_i.

5.2 Transitory Shocks

Just because earnings dynamics are largely driven by fixed and persistent effects, that does not mean we can omit the transitory shock from our specification.

Assume the true specification is that of Equation (3) but was modeled as

$$y_{i,j} = \alpha_i + \eta_{i,j}$$

then

$$\begin{aligned} m_{j,n}(\theta) &= \mathbb{E}[(\alpha_i + \eta_{i,j}) \cdot (\alpha_i + \eta_{i,j+n})] \\ &= \begin{cases} \sigma_\alpha^2 + \sigma_\nu^2 & \text{if } j = n = 0 \\ \sigma_\alpha^2 + \rho^n \sigma_\nu^2 & \text{if } j = 0, n > 0 \end{cases} \end{aligned}$$

To understand the effect on ρ of omitting ϵ let's analyze $\frac{m_{0,2}-m_{0,1}}{m_{0,1}-m_{0,0}}$. Under the misspecified model:

$$\frac{m_{0,2}-m_{0,1}}{m_{0,1}-m_{0,0}} = \frac{\rho^2\sigma_\nu^2 - \rho\sigma_\nu^2}{\rho\sigma_\nu^2 - \sigma_\nu^2} = \rho$$

however, in the true model:

$$\frac{m_{0,2}-m_{0,1}}{m_{0,1}-m_{0,0}} = \frac{\rho^2\sigma_\nu^2 - \rho\sigma_\nu^2}{\rho\sigma_\nu^2 - \sigma_\nu^2 - \sigma_\epsilon^2} = \rho\left[\frac{1}{1+\frac{\sigma_\epsilon^2}{(1-\rho)\sigma_\nu^2}}\right] < \rho$$

So, under the misspecified model the estimate would be to set ρ equal to the empirical counterpart to $\frac{m_{0,2}-m_{0,1}}{m_{0,1}-m_{0,0}}$. However, we can see that this empirical counterpart is less than ρ in the true model. Thus, omitting the transitory component introduces a downward bias in the estimate of the persistence of earnings.

The intuition for the downward bias is that the transitory variance introduces a big drop in the autocovariance function between lag zero and one, which is misinterpreted as a low autocorellation in the persistent shock. This explains many of the low estimates in the literature, such as Heaton and Lucas (1996) who estimate $\rho = 0.6$ when they specify a process with only an AR(1) component.

6 Time Variation in Parameters

There is extensive evidence that there exists time variation in the variance of persistent and transitory shocks. Authors have estimated both how risk evolves over the business cycle, as well as the long term trends in idiosyncratic earnings risk over the past few decades.

6.1 Cyclicality in Risk

Storesletten, Telmer, and Yaron (2001) allow for the conditional variance of the shocks to be different in times of expansions (σ_E^2) versus contractions (σ_C^2). They find $(\sigma_C^2) > (\sigma_E^2)$, which has asset pricing implications, as well as, implications for the welfare cost of business cycles. See Constantinides and Duffie (1996) for a classic description of how the conditional variance of earnings can affect asset prices. See Heathcote, Storesletten, and Violante (2009) for an extension of the framework suitable for quantitative analysis.

6.2 Long-run Trends in Risk

There have been multiple papers that have analyzed the evolution of the conditional variance of persistent and transitory shocks since the formation of the PSID: 1968-2007. Maintaining the same basic specification, but allowing for heteroskedasticity we can estimate the following system.

$$\begin{aligned} y_{i,t} &= \alpha_i + \eta_{i,t} + \epsilon_{i,t} \\ \eta_{i,t} &= \rho\eta_{i,t-1} + \nu_{i,t} \end{aligned}$$

where

$$\alpha \sim (0, \sigma_\alpha^2), \quad \epsilon \sim (0, \sigma_{\epsilon,t}^2), \quad \nu \sim (0, \sigma_{\nu,t}^2)$$

Identification proceeds in a similar, but more complicated, manner to the homoskedastic case and the same moments and estimator can be used to estimate the variance for each shock at each point in time. Often, ρ is assumed to be unity, but that is not necessary.

Alternatively, the variances can be modeled as evolving according to a process. See Meghir and Pistaferri (2004) for evidence of a GARCH component in variance terms.

7 Alternative Specifications: HIP vs. RIP

Lillard and Weiss (1979) present a model in which the life cycle earnings process is no longer stochastic, but rather deterministic and individual specific. Although the deterministic model was largely abandoned in favor of the stochastic models discussed above, recently the idea of heterogeneous income profiles has been revived.[3] In particular, Guvenen (2009) develops a hybrid model, where there is an individual specific age profile **and** stochastic shocks to income.

Let log labor income deviations from a common age profile be

$$y_{i,j} = \alpha_i + \beta_i j + \eta_{i,j} + \epsilon_{i,j} \quad (8)$$
$$\eta_{i,j} = \rho\eta_{i,j-1} + \nu_{i,j} \quad (9)$$

Note this model provides variation around a common age profile for 3 reasons. The α_i and β_i are a deterministic individual specific intercept and slope. $\eta_{i,j}$ is a stochastic persistent component and $\epsilon_{i,j}$ is a transitory component.

Instead of estimating a separate α_i and β_i for each individual, estimate (σ^2_α, σ^2_β, $\sigma^2_{\alpha,\beta}$), where

$$(\alpha_i, \beta_i) \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_\alpha & \sigma^2_{\alpha,\beta} \\ \sigma^2_{\alpha,\beta} & \sigma^2_\beta \end{bmatrix} \right).$$

Often $\sigma^2_{\alpha,\beta}$ is set to 0.

The HIP and RIP models have very different implications for the labor income risk individuals face over the life cycle. The idiosyncratic age profile generates persistent deviations from the common trend, without introducing any risk from the perspective of the agent. From the perspective of the econometrician, ignoring this variable would lead to an upward bias in the estimation of the autocovariance parameter in the persistent shock. Guvenen (2009) estimates $\rho = 0.99$ when σ^2_β is restricted to 0, while $\rho = 0.82$ when σ^2_β is unrestricted. Guvenen (2007) explores the case where agents have to learn their profile over time. Learning occurs slowly, as the agent, just like the econometrician, has difficulty distinguishing between the income profile slope and persistant shocks. This allows the HIP model to produce a rise in consumption inequality over the lifecycle.

MaCurdy (1982) proposed a test (and Abowd and Card (1989) performed a variant of this test), for HIP models based on the sign of the implied autocovariance of income growth. Both are often cited as supporting RIP models to the exclusion of HIP. However, Guvenen (2009) shows this test has low power, especially because it relies on many period ($\sim$10) lags of covariances which are noisy in the data.

To test the HIP vs. RIP model it is imperative to compare the model implications to the data. HIP and RIP have very different predictions for earnings variance and covariances as a function of age. See Guvenen (2009) for details.

[3] In somewhat charged language, those who use models with ex-ante heterogeneous income profiles (HIP models) sometimes call the stochastic process with common expected income profiles "restricted income profiles" (RIP) models.

8 Likelihood Based Methods

Although it has been traditional to estimate statistical earnings processes with minimum distance estimation, there have been some examples of using likelihood based techniques. One major advantage to certain likelihood based estimators is the ability to estimate more complex models, while a downside is the reliance on distributional assumptions on error terms. To my knowledge, Geweke and Keane (2000) was the first attempt, focusing on jointly estimating earnings process parameters and marital status to analyze the transition probabilities between income quartiles over the life cycle. They used the Gibbs sampler and Bayesian techniques, allowing the error terms to be distributed according to a mixture of Normal distributions for better model fit. More recently, Norets and Schulhofer-Wohl (2010)[4], use Bayesian techniques with hierarchical priors to estimate an earnings process with many degrees of heterogeneity—in shock variances, autoregressive coefficients, individual income profiles, and risk aversion parameters. Nakata and Tonetti (2011) explores the small sample properties of many different estimators of the standard earnings processes. In addition to minimum distance estimation, they examine the Maximum Likelihood estimator built with the Kalman filter, an estimator that uses Metropolis-Hastings, and a Bayesian routine using Gibbs sampling (similar to Norets and Schulhofer-Wohl (2010)). They test the performance of these estimators on different specifications of the income process, including time variation and shocks from mixtures of Normals. See Nakata and Tonetti (2011) for more information on the construction and performance of likelihood based estimators of income processes.

[4]Fun fact: Sargent was Geweke's advisor, who was Norets' advisor.

References

Abowd, J. M., and D. Card (1989): "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57(2), 411–445.

Altonji, J. G., and L. Segal (1996): "Small Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business and Economic Statistics*, 14(3), 353–366.

Blundell, R., L. Pistaferri, and I. Preston (2008): "Consumption Inequality and Partial Insurance," *American Economic Review*, 98:5, 1887–1921.

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. 2, pp. 1247–1318. North-Holland.

Constantinides, G. M., and D. Duffie (1996): "Asset Pricing with Heterogeneous Consumers," *Journal of Political Economy*, 104(2), 219–240.

Geweke, J., and M. Keane (2000): "An Empirical Analysis of Earnings Dynamics Among Men in the PSID: 1968-1989," *Journal of Econometrics*, 96, 293–356.

Gottschalk, P., and R. A. Moffitt (1994): "The Growth of Earnings Instability in the U.S. Labor Market," *Brookings Papers on Economic Activity*, 25(2), 217–272.

Guvenen, F. (2007): "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?," *American Economic Review*, 97(3), 687–712.

——— (2009): "An empirical investigation of labor income processes," *Review of Economic Dynamics*, 12(1), 58–79.

Heathcote, J., F. Perri, and G. L. Violante (2010): "Unequal We Stand: An Empirical Analysis of Economic Inequality in the United States 1967-2006," *Review of Economic Dynamics*, 13(1), 15–51.

Heathcote, J., K. Storesletten, and G. L. Violante (2009): "Consumption Insurance and Labor Supply with Partial Insurance: An Analytical Framework," .

——— (2010): "The Macroeconomic Implications of Rising Wage Inequality in the United States," *Journal of Political Economy*.

Heaton, J., and D. J. Lucas (1996): "Evaluating the Effects of Incomplete Markets on Risk Sharing and Asset Pricing," *The Journal of Political Economy*, 104(3), 443–487.

Lillard, L. A., and Y. Weiss (1979): "Components of Variation in Panel Earnings Data: American Scientists 1960-1970," *Econometrica*, 47(2), 437–454.

Lillard, L. A., and R. J. Willis (1978): "Dynamic Aspects of Earning 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.

Murphy, K. M., and R. Topel (1985): "Estimation and Inference in Two-Step Econometric Models," *Journal of Business and Economic Statistics*, 3(4), 88–97.

Nakata, T., and C. Tonetti (2011): "A Likelihood Approach to Estimating Labor Income Processes," *NYU mimeo*.

Norets, A., and S. Schulhofer-Wohl (2010): "Heterogeneity in Income Processes," *mimeo*.

Storesletten, K., C. I. Telmer, and A. Yaron (2001): "The welfare cost of business cycles revisited: Finite lives and cyclical variation in idiosyncratic risk," *European Economic Review*, 45(7), 1311–1339.

Storesletten, K., C. I. Telmer, and A. Yaron (2004a): "Consumption and risk sharing over the life cycle," *Journal of Monetary Economics*, 51(3), 609–633.

Storesletten, K., C. I. Telmer, and A. Yaron (2004b): "Cyclical Dynamics in Idiosyncratic Labor Market Risk," *Journal of Political Economy*, 112(3), 695–717.