Estimating the Hourly Earnings Processes of Top Earners[*]

Jason DeBacker[†] Shanthi Ramnath[‡]

February 2, 2018

Abstract

We describe and apply a methodology to impute hours from survey data (here, the Current Population Survey) to administrative data (here, tax return data from Internal Revenue Service). We use the imputations of hours worked to calculate hourly earnings using earned income reported on individual tax returns. With these hourly earnings in hand, we estimate the dynamics of the hourly earnings processes for tax filing units using a large panel of tax returns. These earnings processes are an important input in calibrating macroeconomic models with endogenous labor supply. Moreover, imputing hours onto administrative data is important for model calibration because these data allow researchers to account for earners in the far right tail of the income distribution. This is in contrast to the typical source of hourly earnings data, which are survey data subject to top-coding.

keywords: earnings, income dynamics
JEL classifications: D31, D91, J69

1 Introduction

A key input to models that study household savings, consumption, and labor supply decisions are the earnings processes households face. As income inequality increases, driven largely by the relative gains of those at the top of the income distribution (see for example, Piketty and Saez (2003) and Bricker, Henriques, Krimmel and Sabelhaus (2016) and Office (2011)), it is increasingly important to calibrate quantitative, heterogeneous agent models using data that allow one to observe these top earners. It is in this context that researchers such as Guvenen, Ozkan and Song (2014), and DeBacker, Heim, Panousi, Ramnath and Vidangos (2013) have been driven to use administrative data to estimate earnings processes. Survey data such as the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID) suffer from top-coding, limiting ones ability to calibrate models in ways that capture the full distribution of incomes. In contrast, administrative data allow researchers to observe the income processes of household inclusive of those at the very top of the income distribution.

[*] The views expressed here are solely those of the authors and do not reflect the views of the Department of the Treasury or the Office of Tax Analysis.

[†] Darla Moore School of Business, University of South Carolina, Department of Economics, DMSB 427B, Columbia, SC 29208, (803) 777-1649, jason.debacker@moore.sc.edu.

[‡] U.S. Department of the Treasury, Office of Tax Analysis, 1500 Pennsylvania Ave., Washington, DC 20220, shanthi.ramnath@treasury.gov.

Earnings processes such as those of DeBacker et al. (2013) and Guvenen et al. (2014) can be used to calibrate many macroeconomic models such as the buffer-stock savings models exemplified by Carroll (1997). However, the income processes these authors estimate are limited in their application due to their measuring earnings under the assumption of exogenous labor supply. Endogenous labor supply is a key component of many macroeconomic models which seek to understand inequality (Castaneda, Diaz-Gimenez and Rios-Rull (2003)), savings (Aiyagari (1994)), income taxation (Guvenen, Kuruscu and Ozkan (2009)), consumption taxation (Daz-Gimnez and Pijoan-Mas (2011)), wealth taxation (Cagetti and Nardi (2009)), political economy (Corbae, D'Erasmo and Kuruscu (2009)), and other topics. However, researchers seeking to calibrate models with endogenous labor supply have not been able to capture the extreme upper tails of the income distribution. This is because in order to calibrate exogenous hourly earnings processes, one needs hourly earnings data - and hours worked are only available from survey data such as the PSID or CPS, which suffer from top-coding. Administrative data such as that from the Social Security Administration (SSA) or the Internal Revenue Service (IRS) do not include any information on hours worked, only total income over the year or quarter.

In this paper, we propose a method to impute hours onto administrative data. We use the CPS to estimate the determinants of hours worked, being careful to use covariates that are represented in the administrative data we use, namely IRS data from individual income tax returns. We then use the models of hours worked from the CPS data to impute hours onto the administrative data. The imputed hours are then used to compute hourly earnings for tax filing units. We then provide estimates of the hourly earnings processes households face.

The results of this paper should be useful in calibrating a wide class of models where earnings dynamics are important. These include computational, general equilibrium tax models such that in Holter, Krueger and Stepanchuk (2014), who study capital and income taxes in an environment with income and wealth inequality. In models such as this where questions of tax progressivity are central, one would like to have estimates of earnings processes that allow for endogenous labor supply and include the tails of the earnings distribution.

The remainder of the paper is structured as follows. Section 2 describes our data sources and how we map the CPS data to tax data. Section 3 presents the models used to estimate hours worked from the CPS data. Section 4 outlines how we use the models of hours worked to impute hours onto tax data and how hourly earnings are calculated. Section 5 estimates models of hourly earnings, presenting the results which will be of interest to researchers looking to calibrate dynamic models with endogenous labor supply. Finally, Section 6 offers concluding remarks.

2 A description of the CPS and tax data

We use two sources of data, a publicly available survey and confidential administrative data. We use the CPS's March Supplement to estimate hours worked as a function of a number of covariates. These are cross-sectional data with weighting variables that allow one to compute nationally representative statistics. Second, we use the IRS's Statistics of Income (SOI) annual cross-sections of individual income tax returns. Specially, we work with a subset of the annual cross-sections produced by SOI. The subset is called the Continuous Work History Sample (CWHS), and represents a panel of tax filing units that is a 1 in 5,000 random sample of tax filers based on the last four digits of the primary filer's social security number.[1] The panel is unbalanced, with some tax units exiting the sample due to death, emigration, or falling below the filing threshold, and others added due to immigration or becoming filers.

The CPS surveys households and gathers detailed information on the individuals in that household. Thus one can construct household-year or person-year observations from the CPS. Much of our tax data are at the level of the tax filing unit. The exceptions are some demographic information and wages and salary income, which we observe for all individuals on the tax return. Tax filing units do not necessarily map directly to households. To make the data as consistent as possible, we use tax filing status as reported in the CPS to create tax filing units in the CPS. We start this process by only keeping adults that are relatively permeant residents in the household (i.e., we exclude boarders and non-relatives who are not roommates or partners of the head of household). Next, if an individual reports their tax filing status as "married, filing jointly", we create a tax unit that is composed of the individual, their spouse, and their children. If an individual reports as filing as being a single filer (including married, filing singly and head of household status), we create a tax filing unit that is composed of that individual and his or her children.

We make some additional sample restrictions on both the CPS and SOI data. In particular, we drop filing units whose heads are less than 21 years old or more than 80 years old. We drop filing units that report total earned income below $1,250 in absolute value (using constant 2005 dollars).[2] We drop filing units from our CPS data that report working less than 250 hours in the prior year.[3] We keep only data from the years 1992 (when the CPS started recording tax filing status) to 2010. Note that we match year $t+1$ CPS data to year t tax data. The reason is that the CPS questions regarding income and hours worked refer to the prior year. Hence, the CPS in year $t+1$ is asking questions about economic activities in year t. Table 1 summarize the variables from the CPS that we use in our analysis. Table 2 summarizes the same set of variables (with the exception of hours worked) for the the CWHS.[4] Both samples are nationally representative and although the SOI data are representative only of the filing population. The sample means are roughly comparable across the CPS and CWHS, but the range of values and the standard deviations of income variables are much larger in the CWHS as it is able to capture the tails of the income distribution without top-coding.

[1] Note that changes in family circumstances can result in taxpayers being dropped from or added to the sample. For example, if a woman who has a sampled SSN four-digit ending marries a man who does not, and he is listed as the primary filer on the couple's joint return, then the woman will be dropped from the sample. In addition, if a couple divorces, only the primary filer with the four-digit SSN ending will be followed after the divorce. Conversely, if a single man with a SSN four-digit ending gets married, and if that man is listed as the primary filer on the couple's joint return, his wife will be added to the sample. For married filing jointly returns, which account for the vast majority of earned income, the primary filer is overwhelmingly the husband. However, in our panel, we have taken care to track households whose composition did not change but the primary filer did, as well as households who misreported the primary filer's SSN.

[2] Note that in our use of an absolute income value, we keep tax filling units who have significant negative values for earned income, which may result from business income losses.

[3] This restriction drops those who work less than one-eighth of a full-year, full-time job.

[4] CWHS max and min values are averages over the top or bottom 10 filing units in order to protect taxpayer confidentiality.

Table 1: Summary Statistics, Current Population Survey, 1992-2010

Variable	Mean	Std. Dev.	Min	Max	Obs
Household Wages	54,131.16	54,934.55	1	1,135,249	954,479
Male Wages	49,565.15	50,639.63	1	713,263	672,047
Female Wages	29,739.36	29,925.77	1	748,263	637,924
Household Self-employment Income	30,701.89	50,906.77	-33,619	897,756	118,341
Male SE Income	35,327.81	54,182.32	-24,756	593,057	77,689
Female SE Income	17,052.54	32,166.15	-24,756	584,539	51,288
Hours	2,743.68	1,225.07	250	10,296	991,166
Male Hours	2,147.66	646.46	1	5,148	726,759
Female Hours	1,783.13	687.93	1	5,148	672,296
Filing Jointly	0.55	0.50	0	1	991,166
Age	42.60	12.64	21	80	991,166
Male Age	43.42	12.52	3	90	760,885
Female Age	42.38	12.23	5	90	803,367
Number of children	0.90	1.15	0	9	991,166

Table 2: Summary Statistics, Continuous Work History Sample, 1991-2009

Variable	Mean	Std. Dev.	Min	Max	Obs
Household Wages	52,754.40	101,420.80	65	43,800,000	333,386
Male Wages	47,317.04	116,858.80	1	43,700,000	221,677
Female Wages	28,440.31	26,364.50	1	868,610	206,428
Household Self-employment Income	25,024.98	87,170.24	-420,475	3,689,110	60,775
Male SE Income	-	-	-	-	-
Female SE Income	-	-	-	-	-
Hours (Predicted)	2,505.64	1,118.97	100	10,296	344,402
Male Hours	-	-	-	-	-
Female Hours	-	-	-	-	-
Filing Jointly	0.45	0.50	0	1	344,402
Age	41.25	13.14	21	80	344,402
Male Age	42.13	13.06	5	92	245,490
Female Age	41.67	12.63	5	108	252,257
Number of children	0.77	1.06	0	10	344,402

3 Estimating hours worked from the CPS

The CPS asks individuals how many hours they worked for income, including both income from employment and self-employment. Thus hours worked considers hours worked towards all sources of earned income. We take the product of the responses to questions about usual hours worked per week and weeks worked last year to construct a measure of hours worked in the last year, the period over which earned income is measured. For those who report filing jointly, we compute hours worked for the filing unit by adding the hours worked by the head of household to those of his or her spouse.

To predict hours worked, we regress the log of filing-unit hours worked on a set of controls. Tax data have detailed information on sources of income and deductions, but lack detailed demographic information such as the educational background and race of the filers. Thus, we utilize only a limited set of the information from the CPS. Specifically, we use data on wages, self-employment income, age, gender, and number of children as controls in our hours worked regressions. Note that our definition of self-employment income from the tax data include income from sole proprietorships, partnerships, and subchapter S corporations, which is consistent with the definition of self-employment income from the CPS. Each of these variables are present in the tax data. However, since tax data do not allow us to see partnership and S-corporation business income separately by primary filer and spouse for units that file jointly, we create a variable in our CPS data that is self-employment income at the filing unit level.[5] Filing unit self-employment income in the CPS is equal to the sum of the self-employment income from the head of household and his or her spouse for those units that report filing jointly.

The set of controls we use to predict hours worked include the log of wage income from the head of household and his or her spouse (if filing jointly), the log of filing-unit self-employment income (or loss), a set of age dummy variables for the head and spouse (if filing jointly), and dummy variables for number of children in the filing unit. Further, we estimate each model separately by year and by filer type. By filer type, we mean the filing status of the tax unit and their income sources. That is, we estimate models separately for each combination of tax filing status (filing jointly, single male filer, single female filer), self-employment income (zero, negative, positive), and wage income (zero wage income from both head and spouse, wage income from both head and spouse, wages from the male filer only, wages from the female filer only).[6]

Thus our model of hours worked is described by the following equation:

$$\begin{aligned} ln(hours_{it}) = &\alpha_{0,t} + \alpha_{1,t} ln(wage_{m,i,t}) + \alpha_{2,t} ln(wage_{m,i,t})^2 + \alpha_{3,t} ln(wage_{f,i,t}) + \alpha_{4,t} ln(wage_{f,i,t})^2 \\ &+ \alpha_{5,t} ln(se\_inc_{i,t}) + \alpha_{6,t} ln(se\_inc_{i,t})^2 + \gamma_{m,t} age_{m,i,t} + \gamma_{f,t} age_{f,i,t} + \beta_t n_c hild_{i,t} + \varepsilon_{i,t}, \end{aligned} \tag{3.1}$$

where the subscripts m and f denote male and female variables, i denotes the tax filing unit, and t the year. Note that the covariates change depending upon the filer type. For example, a single filer household will not have both male and female wage and age variables. A household without self-employment income will not have the $ln(se\_inc)$. Households with negative net self-employment income will have the log of the absolute value of their self-employment losses as $ln(se\_inc)$. We include dummy variables for ages 22 to 80 for both spouses (excluding dummies for 21 year old males and 21 year old female) and dummy variables for 0 to 8 children (excluding the dummy variable for nine or more children). We run regressions separately for for each of the 18 years (1993-2010) and 21 possible combinations for filer status, wage income, and self-employment income.

Table 3 provides an example of the regression output. Note that we only present select coefficients from the year 2009 regression for four combinations for filer status, wage income, and self-employment income. Other regression results are similar.[7] As one would expect, higher income correlates with more hours worked, though hours worked grow at a decreasing rate as income increase. The R^2 of the models suggest that they capture a significant amount of the variation in hours worked in the cross-section.

[5] Tax data would allow us to identify sole proprietorship income separately by primary and secondary filer, but we use a broader definition to include labor income derived from services provided to partnerships and S-corporations - as well as to be consistent with the CPS definition of self-employment income.

[6] Note that our sample selection described in Section 2 precludes certain combinations. For example, a filing-unit cannot have both zero wage and zero self-employment income. Nor can a unit be a single filer and have wage income from both the head and spouse.

[7] All coefficients for all models are available from the authors upon request.

Table 3: Select Hours Regressions, 2010 CPS

	Joint Filers, 0 SE Inc Wages from Both	Joint Filers, SE Inc >0, No Wage Income	Single Male, No SE Inc	Single Female, No SE Inc
ln(wage$_{\text{male}}$)	0.427***		2.333***	
	(0.000)		(0.001)	
ln(wage$_{\text{male}}$)^2	-0.016***		-0.100***	
	(0.000)		(0.000)	
ln(wage$_{\text{female}}$)	0.367***			2.388***
	(0.000)			(0.001)
ln(wage$_{\text{female}}$)^2	-0.012***			-0.103***
	(0.000)			(0.000)
ln(self-emp inc)		1.385***		
		(0.006)		
ln(self-emp inc)^2		-0.056***		
		(0.000)		
Constant	3.116***	-0.962***	-5.871***	-6.011***
	(0.003)	(0.034)	(0.007)	(0.006)
Age controls	Yes	Yes	Yes	Yes
Children controls	Yes	Yes	Yes	Yes
R^2	0.438	0.356	0.453	0.500
Observations	119,497	11,048	111,163	113,965

4 Imputing hours worked onto tax data

With the coefficients in hand from the regressions using CPS data, we now turn to imputing hours on the tax data. We apply the coefficients to the tax data separately by year and filer type, as we estimate the models separately by year and filer type in the CPS data.

Predicted hours are then found by:

$$\begin{aligned} ln(\hat{hours}_{it}) = & \alpha_{0,t} + \alpha_{1,t} ln(wage_{m,i,t}) + \alpha_{2,t} ln(wage_{m,i,t})^2 + \alpha_{3,t} ln(wage_{f,i,t}) + \alpha_{4,t} ln(wage_{f,i,t})^2 \\ & + \alpha_{5,t} ln(se\_inc_{i,t}) + \alpha_{6,t} ln(se\_inc_{i,t})^2 + \gamma_{m,t} age_{m,i,t} + \gamma_{f,t} age_{f,i,t} + \beta_t n_child_{i,t} \end{aligned} \tag{4.1}$$

Because the hours worked models were estimated on top-coded data, we apply the top codes to our IRS data when imputing hours worked (but not when imputing hourly earnings). After we impute hours using the regression models, we apply a maximum hours worked per year that corresponds to the maximal values in the CPS data of 5,148 for singles and 10,296 for couples. Finally, we drop those whose imputed hours are less than 100 hours.. These limitations ensure that our imputed hours are reasonable in the sense that they are physically possible and do not include those with very little labor force attachment.

Table 4 shows what these predicted hours look like by filer type. Panel A shows the summary statistics for the CWHS imputation. Panel B shows summary statistics from the CPS. Comparing hours across those imputed onto the CWHS and those reported on the CPS shows that the hours distributions look quite similar.

To find hourly earnings, we divide earned income by the imputed hours. That is, $e_{i,t} = \frac{wage_{i,t}+se\_inc_{i,t}}{hours_{i,t}}$, where $e_{i,t}$ represent average hourly earnings for all earned income. We then drop from our sample any filing units whose hourly earnings are less than \$5 per hour (in 2005 dollars). Table 5 shows what these hourly earnings look like by filer type.

Table 4: Hours Worked Descriptive Statistics

Panel A: Descriptive Statistics, CWHS Imputed Hours Worked				
Filer Type	Mean Hours	Std. Dev. Hours	Min Hours	Max Hours
All	2,506	1,119	100	10,296
Filing Jointly	3,292	1,187	100	10,296
HH Wage = 0	2,453	952	119	10,296
Male Wage = 0	2,468	1,085	103	10,296
Female Wage = 0	3,772	940	100	10,296
HH SE Income = 0	3,165	1,063	100	10,296
HH SE Income < 0	4,136	2,142	103	10,296
HH SE Income > 0	3,453	1,013	106	10,296
HH Wage > 0 and HH SE Income > 0	3,230	1,059	100	10,296
Single Male	1,926	488	100	5,148
Wage = 0	1,817	488	177	5,148
HH SE Income = 0	1,911	404	100	5,148
HH SE Income < 0	2,464	1,540	102	5,148
HH SE Income > 0	1,936	590	139	5,148
HH Wage > 0 and HH SE Income > 0	1,913	424	100	5,148
Single Female	1,825	466	100	5,148
Wage = 0	1,695	538	238	5,148
HH SE Income = 0	1,817	396	101	5,148
HH SE Income < 0	2,246	1,647	100	5,148
HH SE Income > 0	1,820	603	238	5,148
HH Wage > 0 and HH SE Income > 0	1,817	412	101	5,148

Panel B: Descriptive Statistics, CPS Hours Worked				
Filer Type	Mean Hours	Std. Dev. Hours	Min Hours	Max Hours
All	2,744	1,225	250	10,296
Filing Jointly	3,347	1,234	250	10,296
HH Wage = 0	2,713	1,471	250	10,296
Male Wage = 0	2,367	894	250	10,296
Female Wage = 0	3,909	953	250	10,296
HH SE Income = 0	3,319	1,216	250	10,296
HH SE Income < 0	3,642	1,310	260	10,296
HH SE Income > 0	3,498	1,324	250	10,296
HH Wage > 0 and HH SE Income > 0	3,740	1,175	250	10,296
Single Male	2,043	649	250	6,500
Wage = 0	1,869	853	250	5,148
HH SE Income = 0	2,043	631	250	6,500
HH SE Income < 0	2,173	900	260	5,148
HH SE Income > 0	2,037	789	250	5,148
HH Wage > 0 and HH SE Income > 0	2,102	768	250	5,148
Single Female	1,891	624	250	6,448
Wage = 0	1,869	853	250	5,148
HH SE Income = 0	1,890	610	250	6,448
HH SE Income < 0	2,017	840	260	5,148
HH SE Income > 0	1,911	815	250	5,148
HH Wage > 0 and HH SE Income > 0	1,969	762	250	5,148

Panel A shows the summary statistics for the CWHS imputation. Panel B shows summary statistics from the CPS. Comparing across panels shows some similarities, but the CWHS shows a larger variance and maximum of hourly earnings, which is to be expected as these data are not subject to the top-coding of income variables present in the CPS.

Table 5: Hourly Earnings Descriptive Statistics

Panel A: Descriptive Statistics, CWHS Imputed Hourly Earnings				
Filer Type	Mean Earnings	Std. Dev. Earnings	Min Earnings	Max Earnings
All	21.90	39.93	5.00	10,505.63
Filing Jointly	26.78	312.33	5.00	106,246.41
HH Wage = 0	30.07	207.10	5.00	14,351.31
Male Wage = 0	34.78	291.59	5.00	51,685.46
Female Wage = 0	24.82	347.83	5.00	106,246.41
HH SE Income = 0	23.11	321.51	5.00	106,246.41
HH SE Income < 0	46.95	539.45	5.00	51,685.46
HH SE Income > 0	32.78	124.11	5.00	14,351.31
HH Wage > 0 and HH SE Income > 0	25.30	288.82	5.00	106,246.41
Single Male	20.45	51.48	5.00	11,921.42
Wage = 0	16.70	29.92	5.00	669.43
HH SE Income = 0	19.14	26.51	5.00	4,511.44
HH SE Income < 0	47.58	282.61	5.00	11,921.42
HH SE Income > 0	26.68	37.42	5.00	682.97
HH Wage > 0 and HH SE Income > 0	19.80	27.72	5.00	4,511.44
Single Female	19.05	23.74	5.00	2,615.12
Wage = 0	17.84	64.07	5.00	2,615.12
HH SE Income = 0	18.17	18.37	5.00	1,369.89
HH SE Income < 0	49.88	82.13	5.01	1,083.19
HH SE Income > 0	23.31	42.96	5.00	2,615.12
HH Wage > 0 and HH SE Income > 0	18.51	20.90	5.00	2,615.12
Panel B: Descriptive Statistics, CPS Hourly Earnings				
Filer Type	Mean Earnings	Std. Dev. Earnings	Min Earnings	Max Earnings
All	19.99	20.48	0.13	1,381.82
Filing Jointly	22.47	23.11	-52.94	3,679.72
HH Wage = 0	22.64	38.44	-45.90	1,407.05
Male Wage = 0	27.26	32.41	-12.76	3,679.72
Female Wage = 0	21.38	16.67	-10.63	633.39
HH SE Income = 0	22.33	21.28	0.18	1,211.58
HH SE Income < 0	15.20	18.53	-52.94	348.09
HH SE Income > 0	23.84	32.29	0.13	3,679.72
HH Wage > 0 and HH SE Income > 0	24.03	30.04	0.13	3,679.72
Single Male	19.19	22.21	-35.48	3,123.22
Wage = 0	16.03	25.49	-37.46	612.94
HH SE Income = 0	18.86	21.05	0.30	3,123.22
HH SE Income < 0	10.11	19.87	-35.48	228.33
HH SE Income > 0	23.03	31.42	0.48	805.47
HH Wage > 0 and HH SE Income > 0	27.57	32.72	0.53	612.32
Single Female	16.05	17.23	-37.46	1,969.87
Wage = 0	16.03	25.49	-37.46	612.94
HH SE Income = 0	15.91	15.99	0.36	1,969.87
HH SE Income < 0	9.40	15.69	-37.46	123.50
HH SE Income > 0	18.90	31.66	0.42	1,103.15
HH Wage > 0 and HH SE Income > 0	22.00	38.22	0.53	1,103.15

5 Estimating the hourly earnings processes

The panel dimension of the CWHS allows us to estimate models of earnings dynamics. Stochastic income processes are important in a wide class of models, such as those studying consumption behavior over the lifecycle and those considering the lifetime burden of income taxation. Our goal, then, is to produce estimates of the hourly earnings process that reflect the the full extent of income distribution, rely on a minimal parametric assumptions, and are of use for those seeking to calibrate life-cycle models. Thus we want a simple measure of earnings dynamics that captures the persistence and variance of hourly earnings.

We start with $e_{i,t}$, which are the average hourly earnings of household i in year t. We observe several characteristics of these households, but of these we will only condition on age (determined in the tax data by the age of the primary filer). The reason being that most dynamic models include few state variables that would relate to other household characteristics (such as number of children), but all life-cycle models have age as a state variable. Let $e_{i,s,t}$ be an observations of household i of age s in year t.

5.1 Heterogeneous permanent component model

Assuming that hourly earnings are comprised of an age profile, a permanent component, and a transitory component, we estimate the fixed effect regression model:[8]

$$ln(e_{i,s,t}) = \alpha_i + \beta_s + z_{i,t}, \tag{5.1}$$

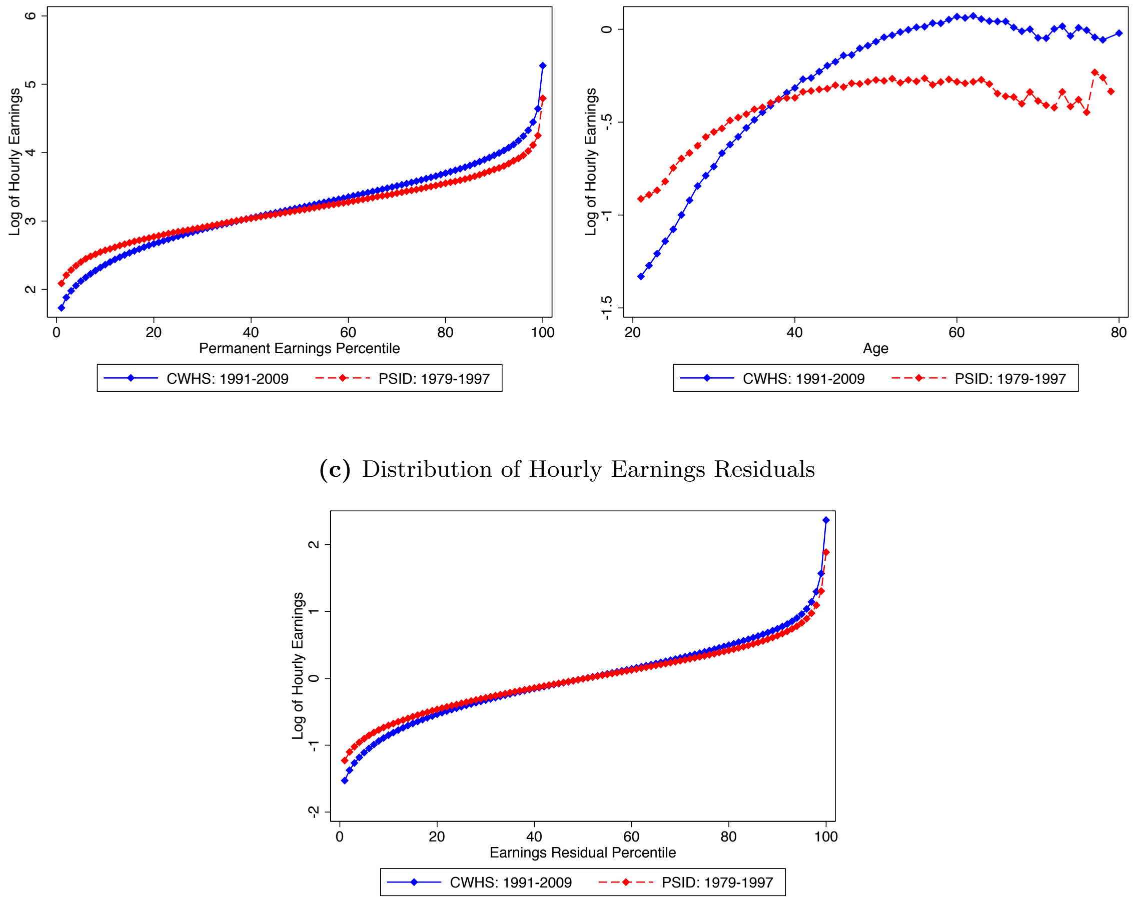
where α_i is the household fixed effect. Note that the vectors of coefficients β_s are the coefficients on dummy variables for each year of age of the primary filer in the filing unit. The household fixed effect is the mean hourly earnings over the life-cycle of the household after controlling for age. This can be interpreted as the permanent component to earnings (e.g., as in Friedman (1957)). The coefficient vector β_s provide estimates of mean hourly earnings by age after controlling for the permanent component of income. Thus β_s give us the life-cycle earnings profile. The $z_{i,t}$ are the stochastic portion of hourly earnings, the residual after conditioning on the filing unit's permanent earnings ability and age. We assume that this transitory component of income follows a First-order Markov Process described by the transition matrix Π_z.

Next we take the $z_{i,t}$ and divide them into percentiles. For each percentile, we calculate the mean. Let $\bar{z}_p$ be a vector containing the means of $z_{i,t}$ for each percentile p. We then calculate the transition matrix for $z_{i,t}$ by finding the fraction that transition between each percentile from one year to the next, call this Π_z. Figure 1a relates the mean of the permanent component, α_i of earnings for each percentile in the distribution of α_i. Call these means for each percentile, $\bar{\alpha}_p$. This figure, and the following figures, plot with these percentiles from the CWHS sample the corresponding values computed from the Panel Study of Income Dynamics (PSID) with the same restrictions we impose on the CWHS data, with the exception of they time frame we consider. The PSID is a panel of survey data from 1968 to the present. We use the PSID from 1979-1997, which covers the same number of years as our CWHS panel, but not the same years. After 1997, the PSID has only been conducted biennially. Given the different frequency of these survey data in later years, we look at the PSID only until 1997. That said, these data still serve as a useful benchmark for the comparison of hourly earnings processes estimated from survey versus administrative data. For example, on can see from Figure 1a that the survey data have the most trouble matching the long right tail of the distribution. For example, the differences between the permanent component of income for the top 1% of the distribution of this component is especially large.The lifecycle component of hourly earnings, β_s, is shown in

[8] Note that we use the log of hourly earnings in our specification, another reason for the restriction noted above that filing units must have an hourly earnings rate in excess of \$5 per hour to be included in the sample.

Figure 1b. Figure 1b plots the mean of the transitory component of earnings by centile, z_p. The distribution of this transitory component again shows how administrative data better capture the top 1% of the distribution.

Figure 1: Hourly Earnings, Fixed Effect Model Results



(a) Distribution of Hourly Earnings Fixed Effects **(b)** Life-cycle Profile Component of Hourly Earnings

(c) Distribution of Hourly Earnings Residuals

Table 6 aggregates Π_z into transitions across deciles for presentational purposes.[9] The probabilities in these transition matrices are consistent with what others have found for mobility in the annual earnings distribution. For example, the probability of staying in the top 1% of the income distribution from one year to the next found by Auten, Gee and Turner (2013) is about 60%. The probability remaining in the top 1% of the residual earnings shock from one year to the next is 54% in our model.

The outputs of estimation, β_s, $\bar{z}_p$, Π_z, and $\bar{\alpha}_p$ are inputs that researchers may use to calibrate a dynamic, structural model with endogenous labor supply. Researchers can use all 100 percentiles or aggregate up from these if they wish to use a smaller state space. Note that with the 100x100 transition matrices, there are some cells with zero or very few observations in them. In order to comply with IRS disclosure restrictions, we move observations from any cell with fewer than 10 observations towards the diagonal and set

[9] The full 100x100 matrix is available in an online appendix at http://jasondebacker.com/papers/EarningsProcess_AppendixTables.xlsx.

Table 6: Transitory Earnings Shocks Transition Matrix, Fixed Effects Model

From/To	1	2	3	4	5	6	7	8	9	10
1	0.64	0.20	0.07	0.03	0.02	0.01	0.01	0.01	0.00	0.01
2	0.18	0.45	0.18	0.08	0.05	0.02	0.01	0.01	0.01	0.01
3	0.08	0.18	0.39	0.17	0.08	0.05	0.03	0.01	0.01	0.01
4	0.04	0.08	0.18	0.34	0.17	0.08	0.05	0.02	0.02	0.01
5	0.02	0.04	0.10	0.19	0.33	0.16	0.07	0.05	0.03	0.01
6	0.02	0.02	0.05	0.10	0.18	0.34	0.16	0.07	0.04	0.02
7	0.01	0.02	0.03	0.05	0.08	0.19	0.35	0.17	0.07	0.03
8	0.01	0.01	0.02	0.03	0.04	0.08	0.20	0.38	0.17	0.06
9	0.00	0.01	0.01	0.02	0.02	0.04	0.07	0.21	0.46	0.17
10	0.01	0.01	0.01	0.02	0.02	0.02	0.04	0.06	0.19	0.63

the number in the cell with fewer than 10 observations to zero. This smearing of the data has the effect of making the income processes seem more persistent, but given the small number of observations involved, the effects are very small.

5.2 Earnings dynamics without heterogeneity in permanent component

For structural models where a smaller state space is necessary, we also estimate the earnings process under the assumption that hourly earnings are composed of an age profile and a transitory component (i.e., without filing-unit fixed effects). To estimate this model, we run the estimate the regression model:

$$ln(e_{i,s,t}) = \beta_s + \eta_{i,t}, \tag{5.2}$$

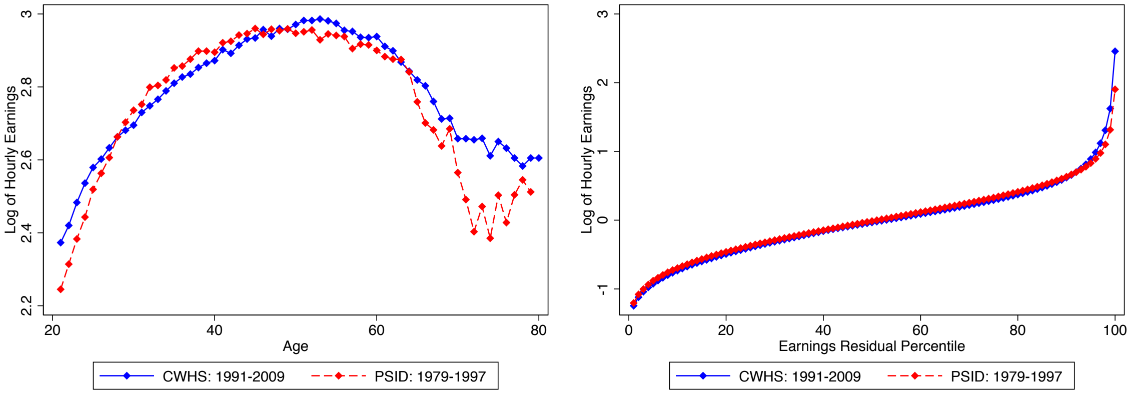
where the coefficient vector β_s provide estimates of mean hourly earnings by age. Thus β_s give us the life-cycle earnings profile.. The $\eta_{i,t}$ are the stochastic portion of hourly earnings, the residual after conditioning on the filing unit's age. We assume that this transitory component of income follows a First-order Markov Process described by the transition matrix Π_η.

Next we take the $\eta_{i,t}$ and divide them into percentiles. For each percentile, we calculate the mean. Let $\bar{\eta}_p$ be a vector containing the means of $\eta_{i,t}$ for each percentile. We then calculate the transition matrix for $\eta_{i,t}$ by finding the fraction that transition between each percentile from one year to the next, call this Π_η. Figure 2 plots mean hourly earnings by percentile, η_p, again comparing the CWHS results to the earnings processes estimated from the PSID. Without fixed effects, a hump-shaped life-cycle profile of hourly earnings becomes more pronounced. We also notice that the CWHS life-cycle profile lies above the PSID profile for middle and older age filers, the points in the age distribution where earnings are most likely to exceed the top-coded values from the PSID. Looking at the distribution of residuals, we again highlight the ability of the administrative data to better capture the right tail of the distribution, as shown by the higher mean value for the top 1% of residual income shocks.

Table 7 aggregates Π_η into transitions across deciles for presentational purposes.[10] Again, these transition matrices are in line with results of income mobility in annual earnings. Of note, the transition matrix from the model without fixed effects has more per-

[10] The full 100x100 matrix is available in an online appendix at http://jasondebacker.com/papers/EarningsProcess_AppendixTables.xlsx.

Figure 2: Hourly Earnings, Without Fixed Effects Model Results



(a) Life-cycle Profile Component of Hourly Earnings

(b) Distribution of Hourly Earnings Residuals

sistence (as measured by the diagonal elements) than the corresponding transition matrix from the model with fixed effects. At first blush, this seems counter-intuitive, since fixed effects should soak up some of that persistence. But note that the distribution of the residuals differ across these two models, and are larger in the model without fixed effects. Thus, while the staying probabilities appear larger in many deciles in the transition matrix from the model without fixed effects, the range of values defining each percentile is also larger in this distribution. Larger ranges for each quantile, all else equal, would tend to increase staying probabilities.

Table 7: Transitory Earnings Shocks Transition Matrix, Transitory Shocks Only Model

From/To	1	2	3	4	5	6	7	8	9	10
1	0.51	0.22	0.11	0.06	0.03	0.02	0.02	0.01	0.01	0.01
2	0.17	0.36	0.20	0.09	0.07	0.04	0.03	0.02	0.01	0.01
3	0.08	0.16	0.32	0.18	0.10	0.06	0.05	0.03	0.01	0.01
4	0.05	0.08	0.16	0.30	0.19	0.09	0.05	0.04	0.02	0.01
5	0.03	0.05	0.08	0.17	0.32	0.19	0.07	0.04	0.03	0.02
6	0.02	0.03	0.05	0.08	0.16	0.33	0.19	0.07	0.04	0.02
7	0.02	0.02	0.03	0.05	0.07	0.16	0.35	0.20	0.07	0.03
8	0.01	0.02	0.02	0.03	0.04	0.07	0.17	0.38	0.19	0.06
9	0.01	0.01	0.02	0.03	0.02	0.04	0.07	0.17	0.46	0.17
10	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.06	0.16	0.66

6 Conclusion

We have shown and applied a methodology to impute hours from survey data onto administrative data. We then use hours imputations to generate data on hourly earnings that are not subject to top coding using panel data from IRS income tax returns. What we find are hourly earnings distributions with larger right tails than found in the survey data. Using hourly earnings derived from administrative data, we estimate parsimonious models of hourly earnings dynamics that can be used to calibrate a wide class of life-cycle models. We are able to estimate permanent and transitory components of hourly earnings, as well as life-cycle earnings profiles that are important inputs for models with endogenous labor supply that consider the distributional effects of economic policies.

References

Aiyagari, S Rao, "Uninsured Idiosyncratic Risk and Aggregate Saving," *The Quarterly Journal of Economics, MIT Press*, August 1994, *109* (3), 659–84.

Auten, Gerald, Geoffrey Gee, and Nicholas Turner, "Income Inequality, Mobility, and Turnover at the Top in the US, 1987-2010," *American Economic Review*, May 2013, *103* (3), 168–172.

Bricker, Jesse, Alice Henriques, Jacob Krimmel, and John Sabelhaus, "Measuring Income and Wealth at the Top Using Administrative and Survey Data," *Brookings Papers on Economic Activity*, 2016, *47* (1 (Spring), 261–331.

Cagetti, Marco and Mariacristina De Nardi, "Estate Taxation, Entrepreneurship, and Wealth," *American Economic Review*, March 2009, *99* (1), 85–111.

Carroll, Christopher D, "Buffer-Stock Saving and the Life Cycle/Permanent Income Hypothesis," *The Quarterly Journal of Economics, MIT Press*, February 1997, *112* (1), 1–55.

Castaneda, Ana, Javier Diaz-Gimenez, and Jose-Victor Rios-Rull, "Accounting for the U.S. Earnings and Wealth Inequality," *Journal of Political Economy*, August 2003, *111* (4), 818–857.

Corbae, Dean, Pablo D'Erasmo, and Burhanettin Kuruscu, "Politico-economic consequences of rising wage inequality," *Journal of Monetary Economics*, January 2009, *56* (1), 43–61.

DeBacker, Jason, Bradley Heim, Vasia Panousi, Shanthi Ramnath, and Ivan Vidangos, "Rising Inequality: Transitory or Persistent? New Evidence from a Panel of U.S. Tax Returns," *Brookings Papers on Economic Activity*, 2013, *46* (1 (Spring), 67–142.

Daz-Gimnez, Javier and Josep Pijoan-Mas, "Flat Tax Reforms: Investment Expensing and Progressivity," CEPR Discussion Papers 8238, C.E.P.R. Discussion Papers February 2011.

Friedman, Milton, *A Theory of the Consumption Function* number frie57-1. In 'NBER Books.', National Bureau of Economic Research, Inc, January 1957.

Guvenen, Fatih, Burhanettin Kuruscu, and Serdar Ozkan, "Taxation of Human Capital and Wage Inequality: A Cross-Country Analysis," NBER Working Papers 15526, National Bureau of Economic Research, Inc November 2009.

___, Serdar Ozkan, and Jae Song, "The Nature of Countercyclical Income Risk," *Journal of Political Economy, University of Chicago Press*, 2014, *122* (3), 621 – 660.

Holter, Hans A., Dirk Krueger, and Serhiy Stepanchuk, "How Does Tax Progressivity and Household Heterogeneity Affect Laffer Curves?," PIER Working Paper Archive 14-015, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania March 2014.

Office, Congressional Budget, "Trends in the Distribution of Household Income Between 1979 and 2007," Reports 42729, Congressional Budget Office October 2011.

Piketty, Thomas and Emmanuel Saez, "Income Inequality in the United States, 1913-1998," *Quarterly Journal of Economics*, February 2003, *118* (1), 1–39.