Online Robustness Appendix to "Are Household Surveys Like Tax Forms: Evidence from the Self Employed"

October 2012

Erik Hurst
University of Chicago

Geng Li
Board of Governors of the Federal Reserve System

Benjamin Pugsley
Federal Reserve Bank of New York

Online Robustness Appendix

This document serves as the online robustness appendix to our paper "Are Household Surveys Like Tax Forms: Evidence from Income Underreporting of the Self-Employed". In this document we present details on two additional topics that are referenced within the main text. First, we outline the procedure we used to test for the effects of income underreporting by the self employed on estimates of precautionary savings. Second, we discuss how our estimating methodology relates to that of Pissarides and Weber (1989).

1. Estimating the Effect of Income Mismeasurement By Self Employed on Precautionary Savings Estimates

To estimate the effects of income underreporting of the self employed on estimates of precautionary savings, we draw on the specification from Hurst et al. (2010). The goal of the Hurst et al. paper was to show how the estimates of precautionary savings fall to close to zero when the self employed are excluded from the analysis. The procedure used in the Hurst et al. paper was nearly identical to the procedure used by Carroll and Samwick (1997, 1998) to provide micro data estimates of the importance of precautionary savings for younger households.

The empirical strategy of estimating the size of precautionary balances using micro data is based on the following specification:

$$\ln(W_{it}) = \alpha_0 + \alpha_1 \sigma_{it}^{permy} + \alpha_2 \sigma_{it}^{transy} + \alpha_3 \ln(y_{it}) + Z_{it}\beta + u_{it} \qquad \textbf{(R1)}$$

where $\ln(W_{it})$ is the log of a measure of household i's wealth in period t, $\ln(y_{it})$ is the log of i's permanent income in t, and σ_{it}^{permy} and σ_{it}^{transy} are, respectively, measures of the variance of permanent shocks and transitory shocks to i's income. The Z vector includes additional controls designed to capture potential household differences in preferences and the hump-shaped profile of wealth over the life cycle.

According to the precautionary saving model, wealth is a function not only of permanent income, but also of uninsurable income risk faced by the household. Almost all empirical studies designed to estimate the size of precautionary balances using micro data proxy uninsurable risk with either the variance of income, the variance of consumption, or they exploit actual job loss or expectations of future job loss. For our paper, we follow Carroll and Samwick (1997, 1998) by using panel data to distinguish between the variance of permanent and transitory shocks to income.

To estimate (R1), we use data from the PSID. We examine accumulated household wealth in either 1984 or 1994. This broadens the analysis performed in Carroll and Samwick (1997, 1998), which only analyzed household wealth accumulation within the PSID using 1984 wealth data. The measure of wealth used is total net worth, which is defined as the sum of checking and savings accounts, bonds, stocks and mutual funds (including IRAs), home equity, other real estate, business equity, cars and other vehicles, and other assets, minus the value of all debts. Since we use logs, we exclude households who have negative or zero net worth in our sample, which amount to a little more than five percent of our sample.

Following equation (R1), we regress the log of household wealth in year *t* (either 1984 or 1994) on both permanent income and measures of the variance of income. We construct permanent income for each household by taking the seven-year average of non-capital income around the period for which we are measuring their wealth. Specifically, when explaining 1984 (1994) wealth holdings, we define permanent income as the average of non-capital income between the years of 1981 and 1987 (1991 and 1997). We use panel data from the PSID to

compute the variances of permanent and transitory shocks to income. We follow the same procedure put forth by Carroll and Samwick (1997, 1998).[1]

Since both permanent income and the variances of permanent and transitory income are measured with error, we instrument for these variables using a large instrument set. As suggested by Carroll and Samwick (1997, 1998), we use occupation dummies and these dummies interacted with age and age squared, as well as industry dummies. In addition, we use the unemployment rate in the county of residence during the prior year, the variance in the county unemployment rate over the sample period, and a dummy for whether the head belongs to a union.

When estimating (R1), we also include additional controls (Z) to capture additional reasons why household wealth may differ across households. The Z vector includes the following demographics: age, age squared, race, gender, marital status, and educational attainment. In addition, we exploit the panel dimension of the PSID to control for past income and wealth shocks experienced by households. Specifically, we include year dummies, along with two dummies for whether the household head was unemployed during the year when the wealth data were collected and whether they were unemployed any time during the prior four years (1980–1983 or 1990–1993). Households that are more likely to face high income risk are also more likely to have been hit by past negative income shocks, and this may weaken the estimated relationship between wealth and risk. We also include dummies for past positive shocks, such as having received inheritances or other lump-sum payments. These were the same included when Hurst et al. (2010) estimated their version of (R1).

[1] See the data appendix to Hurst et al. (2010) for a detailed summary of how the income variances were computed.

Lastly, similar to Carroll and Samwick, we restrict our sample to households whose head is between the ages of 26 and 50 in the year in which the wealth is measured. A detailed description of other restrictions we used in constructing our final sample is reported in the data appendix to Hurst et al. (2010). Our final sample includes 2,144 households.

The base results in the paper are identical to the ones reported in Hurst et al (2010). To assess the effects of the underreporting of income by the self employed on the precautionary savings estimates, we inflated the income measures of the self employed by 25 percent. Otherwise, the specification was identical to the base specification.

2. Comparison to Pissarides and Weber (1989) Method

Pissarides and Weber (1989), PW hereafter, use a similar Engel curve-based approach to detect income underreporting of the self-employed. The main difference between their identification method and the method in this paper is the treatment of transient income volatility of employees and the self-employed. Our estimates of underreporting decrease only slightly after accounting for differences of income volatility using the PW method.

The Hurst, Li, Pugsley (2012), HLP hereafter, identification method models reported income as

$$\log y_{ikt} = \log \kappa_k + \log y_{ikt}^p + \Omega_k X_{ikt} + \nu_{ikt}$$

with $E[\nu_{ikt} \mid \log y_{ikt}^p, X_{ikt}] = 0$ for $k = W, S$, and $\log \kappa_W = 0$. This embeds two important restrictions: (1) conditional on a level of permanent income $\log y_{ikt}^p$, the expected transient log deviations in income are equal to zero, and (2) that the fraction reported is constant within

groups at 1 and κ for workers and self-employed respectively.[2] Under these assumptions, κ is identified as

$$\kappa = \exp\left(-\frac{\gamma}{\beta}\right)$$

where γ and β are the coefficients from the Engel curve regression.

PW make the following parametric assumptions in their model of reported income

$$\log y_{ikt} = \log \kappa_{ikt} + \log y_{ikt}^{p} + \nu_{ikt} \ ,$$

with

$$\nu_{ikt} = \begin{cases} \mu_{pW} + \sigma_{pW}\zeta_{iWt} & \text{if } k = W \\ \mu_{pS} + \sigma_{pS}\zeta_{iSt} & \text{if } k = S \end{cases},$$

with ζ_{ikt} independent standard normal and $\mu_{pW} + \frac{1}{2}\sigma_{pW}^{2} = \mu_{pS} + \frac{1}{2}\sigma_{pS}^{2}$. The purpose of this reporting assumption is to ensure that $E\left[y_{ikt}^{p}\exp\left(\nu_{ikt}\right) \mid y_{ikt}^{p}, X_{ikt}, k\right]$ does not depend on group k, i.e., conditional on a level of permanent income (and other individual characteristics), expected annual income is equal across groups. HLP instead assume that $E\left[\log y_{ikt}^{p} + \nu_{ikt} \mid y_{ikt}^{p}, X_{ikt}, k\right]$ does not depend on k. The distinction is relevant if there are large differences in income volatility across groups because of the Jensen's inequality correction. Further, PW allow underreporting to vary within group by assuming

$$\log \kappa_{ikt} = \begin{cases} 0 & \text{if } k = W \\ \mu_{iSt} + \sigma_{\kappa}\omega_{iSt} & \text{if } k = S \end{cases}$$

[2] $\log \kappa$ can also represent the average of the log fraction reported, so long as individual deviations from $\log \kappa$ are zero on average.

with ω_{ikt} independent standard normals. [3] Again, the uncertainty is relevant when computing $E[\kappa]$ the expected fraction reported from $\log\kappa$. Under these assumptions, $E[\kappa]$ is identified as

$$E[\kappa] = \exp\left(-\frac{\gamma}{\beta} + \left(\mu_{pW} - \mu_{pS}\right) + \frac{1}{2}\sigma_{\kappa}^{2}\right)$$

where $\mu_{pW} - \mu_{pS}$ is the volatility correction needed to make the additional variance of log self-employed transient income fluctuations a mean preserving spread in levels, and the $\frac{1}{2}\sigma_{\kappa}^{2}$ adjusts for the additional uncertainty in reported income due to variation in underreporting among the self-employed.

The $\left(\mu_{pW} - \mu_{pS}\right) + \frac{1}{2}\sigma_{\kappa}^{2}$ term can be identified off of differences in reported income volatility assuming income underreporting differences are uncorrelated with transient income volatility for the self-employed. Let σ_{yk}^{2} denote the variance of the error from a regression of reported income on all the covariates and exogenous instruments. The error includes the unpredictable component of permanent income η_{ik} as well as reporting and transient income shocks

$$\sigma_{yk}^{2} = \mathrm{Var}[\eta_{ik} + \sigma_{pk}\zeta_{ik} + \sigma_{\kappa}\omega_{ik}]$$

with $\sigma_{\kappa W} = 0$. Because permanent income is uncorrelated with reporting and transient income shocks the errors, then assuming that the variance of permanent income shocks is equal for both groups then

$$\sigma_{yS}^{2} - \sigma_{yW}^{2} = \sigma_{pS}^{2} + \sigma_{\kappa}^{2} - 2\rho_{\kappa pS}\sigma_{pS}\sigma_{\kappa} - \sigma_{pW}^{2} .$$

[3] PW are actually interested in $k = 1/\kappa$ the adjustment factor to apply to reported income. Of course, both variables are log normal, only the sign of

With $\mu_{pW} - \mu_{pS} = \frac{1}{2}\left(\sigma_{pS}^2 - \sigma_{pW}^2\right)$ by assumption, then

$$\mu_{pW} - \mu_{pS} + \frac{1}{2}\sigma_\kappa^2 = \frac{1}{2}\left(\sigma_{yS}^2 - \sigma_{yW}^2 + 2\rho_{\kappa pS}\sigma_{pS}\sigma_\kappa\right)$$

after cancelling the σ_κ^2 terms. If we assume $\rho_{\kappa pS} = 0$, that underreporting and income volatility are uncorrelated for the self employed, then

$$E[\kappa] = \exp\left(-\frac{\gamma}{\beta} + \frac{1}{2}\left(\sigma_{yS}^2 - \sigma_{yW}^2\right)\right)$$

So under these assumptions, the HLP estimates of underreporting are biased up to the extent that self-employed transient income volatility exceeds that for workers.

We estimate $E[1-\kappa]$ using the estimated coefficients from the Engel curve regression and an adjustment $\frac{1}{2}\left(\sigma_{yS}^2 - \sigma_{yW}^2\right)$ estimated from residuals of reported earnings on the controls and instruments using the 1 year and 3 year averages from the PSID.[4] Table R1 shows the original estimate of $1-\kappa$, and the adjusted value of $E[1-\kappa]$ after correcting for differences in transient income volatility using the PW method. Since income volatility is higher among the self-employed, the adjustment attenuates the estimate of underreporting: the unreported fraction falls from 32 percent to 21 percent instrumenting for total family income in the one year sample. The effect is more modest in the 3 year sample we use a 3 year average of total family income in place of the instrumented annual total family income.

Although the correction under the PW assumptions is small, the effect may be even smaller when differences in permanent income volatility are considered. If the unpredictable

[4] This assumes that permanent log income volatility is constant across groups, to the extent that permanent log income volatility of the self employed is greater than that of workers, estimates of $\frac{1}{2}\left(\sigma_{yS}^2 - \sigma_{yW}^2\right)$ are biased up.

component of permanent income is more volatile for the self employed, i.e., $\sigma_{\eta S}^2 - \sigma_{\eta W}^2$ is positive, this further weakens the PW correction.

We try to estimate this difference using residuals of consumption on the controls and instruments. The error includes unpredictable component of permanent income scaled by the income elasticity, as well as other unobserved independent determinants of consumption. If we assume that the second component of consumption volatility is constant across groups then we can estimate the $\sigma_{\eta S}^2 - \sigma_{\eta W}^2$ from residuals normalized by the β^2. When this quantity is non zero then

$$E[\kappa] = \exp\left(-\frac{\gamma}{\beta} + \frac{1}{2}\left(\sigma_{yS}^2 - \sigma_{yW}^2 - \left(\sigma_{\eta S}^2 - \sigma_{\eta W}^2 \right) \right) \right)$$

The last row of Table R1 shows the corrected estimates of $E[1-\kappa]$ after incorporating the estimate of $\sigma_{\eta S}^2 - \sigma_{\eta W}^2$. Underreporting increases from 21 to 23 percent using the one year sample. With this adjustment, while estimated underreporting is still smaller than the HLP estimates, the differences are small in magnitude.

Overall, the effect of adjusting for the differences in volatility between the groups and explicitly incorporating heterogeneity in underreporting has small quantitative effects documented in Table R1. We have also assumed that the fraction reported is uncorrelated with the transient income shock $\rho_{\kappa pS} = 0$. If underreporting is higher (fraction reported is lower) during good years, this would actually increase estimates of underreporting.

References

Carroll, Christopher., and Andrew Samwick (1997), "The Nature of Precautionary Wealth," *Journal of Monetary Economics*, 40(1), pp. 41—71

Carroll, Christopher and Andrew Samwick (1998). "How Important is Precautionary Saving?" *Review of Economics and Statistics*, 80(3), pp. 410-9.

Hurst, Erik, Annamaria Lusardi, Arthur Kennickell, and Francisco Torralba (2012)."The Importance of Business Owners in Assessing the Size of Precautionary Savings," *Review of Economics and Statistics*, 92(1), pp. 61-9.

Pissarides, Christopher and Guglielmo Weber (1989). "An Expenditure Based Estimate of Britain's Black Economy," *Journal of Public Economics*, 39(1), pp. 17-32.

Table R1: Alternative Estimates of 1-κ with PW Income Volatility Adjustment

	Labor + Business Income		Total Family Income	
Estimate	1 Year	3 Year Averages	1 Year	3 Year Averages
$\hat{\beta}$	0.310	0.313	0.315	0.322
$\hat{\gamma}$	0.146	0.140	0.119	0.112
$\hat{\sigma}^2_{yS} - \hat{\sigma}^2_{yW}$	0.403	0.245	0.288	0.204
$\hat{\sigma}^2_{cS} - \hat{\sigma}^2_{cW}$	0.00650	0.0103	0.00650	0.0103
HLP $1-\kappa$	37.7 %	36.0%	31.5 %	29.4 %
PW $E[1-\kappa]$	28.0 %	29.1 %	20.9 %	21.8 %
PW $E[1-\kappa]$ with permanent income adjustment	30.4 %	32.7 %	23.4 %	25.6 %