October, 2011

Using Panel Data to Exactly Estimate Income Under-Reporting by the Self-Employed

Bonggeun Kim*
John Gibson**
Chul Chung***

Abstract

Self-employment income is believed to be understated in economic statistics but there is debate about the extent of under-reporting and the resulting estimates of the underground economy. This paper refines a method developed by Pissarides and Weber (1989) to use discrepancies between food shares and reported incomes to estimate under-reporting by the self-employed. Our panel data approach distinguishes income under-reporting from transitory income fluctuations of the self-employed. Previous studies only provide an interval estimate of under-reporting but the current study provides an exact estimate and also relaxes unlikely assumptions that under-reporting is independent of transitory fluctuations. Panel data from Korea and Russia are used to illustrate the method, and suggest that in both countries approximately 20 percent of the income of self-employed households is not reported.

JEL: D12, H26, H31, O17

Keywords: Engel curve, Measurement error, Self-employment, Underground economy

Acknowledgements:
We are grateful for the financial support from Marsden Fund grant UOW0504. Helpful assistance was received from Steven Stillman. We are also grateful for comments from Martin Browning, Bob Gregory, Steven Stillman and participants in the 2008 ESAM meetings. All remaining errors in this paper are those of the authors.

*Department of Economics, Seoul National University, bgkim07@snu.ac.kr
**Department of Economics, University of Waikato, Private Bag 3105, Hamilton, New Zealand. Fax: (64-7) 838-4035. jkgibson@waikato.ac.nz.
***Korea International Trade Association and Korea Institute for International Economic Policy, cchung@kiep.go.kr

I. Introduction

The income of the self-employed is often assumed to be understated in both economic statistics generated from tax records and in data gathered from surveys. The motive for understating when dealing with tax collectors is clear but there may seem to be less reason for the self-employed to understate when talking to survey data collectors. However, as Pissarides and Weber (1989, p.17) point out: "[d]espite assurances about confidentiality, people may have no incentive to reveal the true extent of their activities to the data collector from fear that they may not be, after all, protected from the law." Nevertheless, it takes a sophisticated cheat to appear consistently poorer throughout all parts of a survey. A respondent may remember to reduce reported income but not expenditure, or to reduce totals of both but not adjust the ratios between expenditure components, such as food shares, in ways that would be consistent with their claimed lower income level.

Consequently, several studies of the underground economy rely on relationships between survey sub-aggregates, such as income or expenditure components.[1] For example, Pissarides and Weber (1989) [henceforth, PW] assume that all survey respondents correctly report food expenditure but only employees correctly report incomes.[2] The relationship between food and income for employees helps back out a range of estimates for true self employment income. That only a range is estimated reflects a weakness of cross-sectional data, which cannot distinguish under-reporting from the greater deviation of current income around permanent income for the self-employed. Despite this weakness, and a reliance on an assumed log-normal distribution to make the estimates tractable, the PW method has been used in several applied studies (Schuetze, 2002; Johansson, 2005). The PW method has also been extended to complete demand systems (Lyssiotou, Pashardes and Stengos, 2004). This is a useful

[1] A larger literature uses macro approaches that measure the underground economy by the gap between recorded activity and proxies for true economic activity like currency or electricity demand (Johnson, Kaufmann and Shleifer, 1997). There is considerable criticism of these macro approaches (Thomas, 1999).

[2] With taxes automatically deducted from wages at source in the countries studied in this paper, there is little reason for employees to understate incomes.

refinement if self-employment income is not spent in the same way as other income. For example, households may allocate volatile self-employment income to 'big ticket' expenses and use steady wages for food and other necessities. Also, self-employment income may be associated with certain expenses like restaurant meals that can be used as business deductions.

In this paper we further refine the PW method to obtain an improved measure of income under-reporting by the self-employed, by using panel data. Our approach can separate the effects of income under-reporting from the effects of transitory income variations. Hence we can form an exact estimate of the degree of under-reporting as opposed to the interval estimates from the original PW method. Also our method avoids having to assume that the degree of under-reporting is independent of the degree of transitory fluctuations. This assumption carries the undesirable implication that, when questioned about their income, the self-employed adopt a proportional rule such as 'always report 70% of true income' rather than a rule based on actual amounts like 'never report more than $50,000 of income' or an under-reporting approach that varies from year to year as their income fluctuates. [3]

While the PW method has previously been used on pooled annual data by Schuetze (2002), there was no refinement of the estimation framework to exploit the advantages of a longitudinal aspect of the data. Thus, only an interval estimate of the under-reporting rate was reported by Schuetze (2002). The examples below show that these intervals may be too wide and too volatile to be of practical value. In contrast, our methodological refinement allows more exact estimates of income under-reporting. This better measurement may matter both for improved GDP statistics and for tax policy. Undeclared economic activities reduce the tax base but raising tax rates to compensate for the loss of public revenue reinforces the incentive to under-report (Lyssiotou, et al, 2004). Hence, having good estimates of the size of the

[3] The assumption that the self-employed adopt a propotional underreporting rule is also not empirically supported in the Canadian data used by Tedds (2010), who provides a refinement of the PW method in a different direction than the current paper, by using a nonparametric reporting function.

underground economy may help the tax authorities decide on their best strategy. Also, correctly measuring self-employment income is important for models of growth and technology that assume identical functional income shares across time and space (Gollin, 2002).

Our study also links to a more recent literature using food Engel curves to estimate CPI bias (Costa, 2001; Hamilton, 2001a). The logic of this method is that Engel curves should not drift over time if preferences are stable and nominal income variables and deflators have no systematic errors. In a related paper, Hamilton (2001b) backs out the true black-white income difference by observing that food budget shares in the U.S. fell substantially more for blacks than whites (over 1974-91) due to uneven CPI biases across race. In our case, the analogous drift in the Engel curve of the self-employed relative to that of employees is attributed to the income under-reporting of the self-employed since there is no reason to believe that deflator bias varies by employment status.

In the examples below we follow PW and apply our refined approach to a food Engel curve. While our refinement could also be extended to full demand systems like those used by Lyssiotou et al. (2004) we prefer to work with food budget data which we believe to be more accurately measured in the surveys used here. More generally, estimating full demand systems with survey data may result in biases from error-ridden measurements of certain volatile expenses (e.g. non-regular durable goods expenses) contaminating the parameter estimates for more reliably measured regular consumption items like food. These biases may offset the potential advantages of allowing for preference heterogeneity.[4]

The structure of the paper is as follows. Section II discusses the empirical methodology and puts our refinement into the context of the Pissarides and Weber approach. We describe

[4] Of course the full demand system approach is still valuable if reliable data on all budget items are available. For example, Feldman and Slemrod (2007) use a full demand system approach with using unaudited tax return data for US taxpayers, which is potentially more accurate than survey data.

our two data sets and empirical results in section III and the discussion and conclusions are in Section IV.

II. Methodology

1. The Food Engel Curve

We use an Engel curve where the food expenditure share is a linear function of log transformed real permanent income, a relative price of food to non-food, and other household characteristics:

$$w_i = \phi + \gamma\left(\ln P_F - \ln P_N\right) + \beta \ln y_i^P + \mathbf{X}'\theta + \varepsilon_i, \tag{1}$$

w_i is household i's food budget share, $P_{F,}$ and P_N are the price indexes of food and non-food, y_i^P is the permanent income of household i deflated by a consumer price index, X is a vector of other characteristics of household i and ε_i is a pure random error. Although this starts as the same Engel curve used in the CPI bias literature we develop it in a different way.

2. The Pissarides and Weber Method

Pissarides and Weber (1989) note that instead of y_i^P, surveys record income y_{it}^* in year t which has two error components compared to the true permanent income:

$$\begin{aligned} &y_{it} = g_{it} y_i^P \quad , \quad y_{it} = k_{it} y_{it}^* \\ &\Leftrightarrow \ln y_{it}^* = \ln g_{it} + \ln y_i^P - \ln k_{it} \end{aligned} \tag{2}$$

The first component is that even with no under-reporting, the best that can be measured is y_{it} -- the actual income in year t -- which is expected to be sensitive to the business cycle and other fluctuations, with g_{it} degree of transitory income variation around permanent income y_i^P. If g_{it} is greater than one, a household has a good year and has positive transitory income. It is assumed by PW that g_{it} has the same mean for employees and the self-employed but that the variance of g_{it} is higher for the self-employed.

The other error component, k_{it} represents income under-reporting, and is the factor (assumed greater than 1.0 for the self-employed and exactly 1.0 for employees) by which reported income has to be multiplied in order to obtain true current income. To make estimation of income under-reporting by the self-employed feasible, PW and subsequent applications assume that the components g_{it} and k_{it} follow log normal distributions:

$$\begin{aligned} \ln k_{it} &= \mu_k + v_{it} \\ \ln g_{it} &= \mu_g + u_{it} \end{aligned} . \tag{3}$$

Inserting equation (2) and (3) into equation (1):

$$w_i = \phi + \gamma\left(\ln P_F - \ln P_N\right) + \beta \ln y_{it}^* + \beta(\mu_k - \mu_g) + \beta(v_{it} - u_{it}) + \mathbf{X}'\theta + \varepsilon_i . \tag{4}$$

The key part of equation (4) for estimating the degree of income under-reporting by the self-employed is $\beta(\mu_k - \mu_g) + \beta(v_{it} - u_{it})$ which has several unobserved components. If instead, an Engel curve is estimated using only observable variables, including a dummy variable to identify households with self-employment income:

$$w_{it} = \phi + \gamma\left(\ln P_{Ft} - \ln P_{Nt}\right) + \beta \ln y_{it}^* + \delta D_{it} + \mathbf{X}'\theta + \varepsilon_{it}, \tag{5}$$

where $D_{it} = 1$ for households with self-employment income, then the dummy coefficient is:

$$\begin{aligned} \delta &= \beta[(\mu_{kSE} - \mu_{kEE}) - (\mu_{gSE} - \mu_{gEE})] \\ &= \beta[\mu_{kSE} + \frac{1}{2}(\sigma_{uSE}^2 - \sigma_{uEE}^2)] \end{aligned} \tag{6}$$

where the subscripts SE and EE denote the self-employed and employees. The simplification in equation (6) follows from μ_{kEE}=0, under the assumption that k_{it}=1 for employees and from the assumed log-normality of g_{it} which lets the mean be written in terms of the variance.

The mean of the under-reporting component can be derived from the properties of the log-normal distribution for k_{it} and by substituting in from equation (6) for μ_{kSE}:

$$\ln \bar{k} = \mu_{kSE} + \frac{1}{2}\sigma^2_{vSE} = \frac{\delta}{\beta} + \frac{1}{2}[\sigma^2_{vSE} - (\sigma^2_{uSE} - \sigma^2_{uEE})] \qquad (7)$$

However in equation (7) the variances of transitory income of both occupational groups, σ^2_{uSE} and σ^2_{uEE} and the variance of the self-employed income under-reporting rate, σ^2_{vSE} are not known. So, PW turn to another source of information on those variances by using the residual variance from a reduced-form regression for reported income as below:

$$\ln y^*_{it} = Z'\pi + \zeta_{it} \qquad (8)$$

where Z includes a set of proxy variables representing permanent income.[5] The composite error term of equation (8) contains deviations of transitory from permanent income, reporting deviations and random variation in permanent income. The residual variances for SE and EE are related by:

$$\sigma^2_{\hat{\zeta}SE} - \sigma^2_{\hat{\zeta}EE} = \sigma^2_{vSE} + (\sigma^2_{uSE} - \sigma^2_{uEE}) - 2\,\mathrm{cov}(uv)_{SE}\,. \qquad (9)$$

Pissarides and Weber then consider both the lower bound case ($\sigma^2_{vSE} = 0$) and the upper bound case ($\sigma^2_{uSE} = \sigma^2_{uEE}$) in equation (7), which gives an interval in which $\bar{k}$ must lie:

$$\ln \bar{k} \in [\frac{\delta}{\beta} - \frac{1}{2}(\sigma^2_{\hat{\zeta}SE} - \sigma^2_{\hat{\zeta}EE}) + \mathrm{cov}(uv)_{\mathrm{SE}}, \frac{\delta}{\beta} + \frac{1}{2}(\sigma^2_{\hat{\zeta}SE} - \sigma^2_{\hat{\zeta}EE}) + \mathrm{cov}(uv)_{\mathrm{SE}}]\,. \qquad (10)$$

However, equation (10) still contains an unobservable, $\mathrm{cov}(uv)_{\mathrm{SE}}$, so PW further assume that $\mathrm{cov}(uv)_{\mathrm{SE}}$ =0. This (unlikely) assumption that the degree of under-reporting is independent of the degree of transitory income variation yields an empirically estimatable interval for k as:

$$\ln \bar{k} \in [\frac{\delta}{\beta} - \frac{1}{2}(\sigma^2_{\hat{\zeta}SE} - \sigma^2_{\hat{\zeta}EE}), \frac{\delta}{\beta} + \frac{1}{2}(\sigma^2_{\hat{\zeta}SE} - \sigma^2_{\hat{\zeta}EE})]\,. \qquad (11)$$

[5] These same variables are used as instruments for endogenous income when equation (5) is estimated by 2SLS as used in the conventional PW method.

3. A More Exact Panel Data Method

With panel data it is possible to make an exact estimate of the degree of income under-reporting by the self-employed, rather than just the interval estimate that comes from applying the PW approach to cross-sectional data. A further advantage of panel data is that the under-reporting estimate can be made with fewer assumptions. In particular, there is no need to assume that the degree of under-reporting is independent of the degree of transitory income variation. This allows for the possibility that the self-employed may increase their under-reporting rate as positive transitory income increases, which is consistent with a rule based on actual amounts like ‘never report more than $50,000 of income’

Specifically, with panel data one can use “between estimation” where the mean value of reported incomes over time for the same household is used as the data in the regression. This use of household-specific means enables the transitory income variations of both self-employed and employee households to be controlled for. The potential comovements of income variations with the degree of income under-reporting by the self-employed can also be controlled for so that there is no need to rely on simply assuming that the under-reporting rate is independent of the degree of transitory income variation.

With between estimation the counterpart to equation (2) is:

$$\overline{\ln y_{it}^{*}} = \overline{\ln y_{it}} - \overline{\ln k_{it}} = \ln y_{i}^{P} + \overline{\ln g_{it}} - \overline{\ln k_{it}} \tag{12}$$

where $\overline{\ln y_{it}^{*}}$ means $\sum_{t=1}^{T} \ln y_{it}^{*} / T$.[6] This household-specific mean allows the positive and negative variations of transitory income over time to cancel each other out, since:

[6] With between estimation, we estimate the food Engel curves and the income equations on the time-averaged values of all variables. Unlike the previous study of Schuetze (2002) which mixes “between” and “within” variation in pooled annual data, our analysis uses only the “between” variation across households. This allows a more revealing comparison to previous studies using cross-sectional data. In the results below we compare “between” estimates with year-to-year application of the PW method. Also, “within” variation may differ

$$p\lim_{T\to\infty}\sigma^2_{\bar{u}_i} = p\lim_{T\to\infty}\frac{\sigma^2_u}{T} = 0\,. \tag{13}$$

In other words, with large enough T, the variations of transitory income go away and the covariance between the degree of under-reporting and the degree of transitory income variation also disappears. This greatly simplifies the estimation task. For example, in comparison with equation (10) the cov(*uv*)$_{\mathrm{SE}}$ term disappears and since the variations due to transitory income have also disappeared it is logically true rather than just an assumption that ($\sigma^2_{uSE} = \sigma^2_{uEE}$).

Allowing $\overline{\ln k_i}\,(=\mu_k + v_i)$ to follow a normal distribution, with the only stochastic contribution coming from the cross-sectional variance of the self-employment income under-reporting rate, σ^2_{vSE} the estimator of interest is:

$$\ln\bar{k} = \mu_{kSE} + \frac{1}{2}\sigma^2_{vSE} = \frac{\delta}{\beta} + \frac{1}{2}(\sigma^2_{\hat{\zeta}SE} - \sigma^2_{\hat{\zeta}EE}) \tag{14}$$

Unlike in the cross-sectional case there is no need to estimate upper and lower bounds and we instead have an exact estimate of the under-reporting rate (albeit subject to sampling error, which also affects the estimated bounds in the original PW approach). Thus with panel data it is possible to remove one source of uncertainty about the extent of income under-reporting, while not resorting to unrealistic assumptions about the independence of under-reporting from transitory income variations. We provide two examples of our refined approach below.

III. Empirical Analysis

1. Data

We use data from two panel surveys, the Korea Labor Income Panel Survey (KLIPS) from 2000-2005 and the Russian Longitudinal Monitoring Survey (RLMS) from 1994-2000.

greatly from "between" variation, when there is potentially an intrinsic tendency for under-reporting income from self-selected self-employment status.

The survey data for each country have been used in a number of other published papers. For example, for studies of CPI bias the KLIPS data were used by Chung, Gibson, and Kim (2010) and the RLMS data by Gibson, Stillman and Le (2008).

In each case we restrict attention to urban households, since measured food shares for rural households may be distorted if the survey imperfectly captures consumption from own production, which is more prevalent in rural areas. We also restrict attention to households with two adults, with or without children, since more precise estimates of the under-reporting parameter may be obtained by focusing on a fairly homogeneous group. The samples are further restricted to those households whose food-at-home shares are in the 0.01-0.99 interval and where both the household head and their spouse are aged between 20-65 years.

To show how our main variables like food shares and household incomes have changed over time, the beginning, middle and end-period averages of those variables are reported in Table 1 and 2. The first row of Table 1 for KLIPS shows that the average food-at-home share in Korea fell by about 12 percentage points from 30 percent in 2000 to 18 percent in 2005. Over the same period, nominal household income grew by 63 percent and its real value adjusted by the CPI grew about 40 percent. For Russia, the average food-at-home share fell by about 10 percentage points during the sample period while the average real household income appears to decrease slightly. While this declining food share is also consistent with CPI bias, as found for both countries by Gibson et al. (2008) and Chung et al. (2010), these potential biases should not affect the results reported below, since the same CPI deflator is used for both self-employed and employee households.

Table 1 also shows that in Korea the average reported income is higher for the employees than for the self-employed, but the food-at-home shares imply the opposite pattern. Assuming that survey respondents correctly report their consumption expenditures, the apparent violation of Engel's Law between the two occupational groups suggests that there may be a substantial degree of income under-reporting by the self-employed. For Russia there is a somewhat similar pattern (Table 2). Even though the average reported income is slightly higher for the self-employed the average food share is substantially lower. It would take an implausibly large income elasticity of demand for food in order for measured income to account for the gap in the food shares between the two employment groups.

Hence it seems likely that in both countries there is a downward shift in the food Engel curve for the self-employed. We attribute this downward shift to unmeasured real income of the self-employed, which results from their under-reporting of nominal income.[7]

2. Estimation Methods and Model Specification

Equation (5) is a linear model and can be estimated separately for each year, either with OLS or if there are concerns about endogenous income then with 2SLS. This approach, of treating the panel as a set of annual cross-sections, would be consistent with the PW method and would yield a separate interval estimate for $\bar{k}$ in each year. But since the data for each country are actually a panel we also can use the method described in Section II.3, relying on between estimation. In other words, we estimate equation (12) using six-year average values to control for variations in transitory income, and using 2SLS to control for the endogeneity of income. The resulting estimate of $\bar{k}$ will be a single value, since there is no need to make an interval estimate and since the year-by-year fluctuations also disappear.

To provide examples of our suggested approach we use data from two panel surveys; the Korean Labor Income Panel Study (KLIPS) and the Russian Longitudinal Monitoring Survey (RLMS). Since the data manipulations and empirical specifications are similar for both we discuss our procedures applied to KLIPS in detail and give a briefer description for RLMS in the Appendix. We use six rounds of KLIPS data from 2000 to 2005, and combine these with

[7] An alternative explanation, of differential CPI bias, can be ruled out. Unlike Hamilton (2001b), who found differential bias for Blacks and Whites due to geographic segregation in the U.S., there is no similar segregation by employment status in either Korea or Russia or more generally. In addition, we also directly checked this alternative explanation by replacing the national CPI with the regional CPI in the results for Korea and there was no change in the main findings.

the annual national CPI and the regional CPI that is calculated for each of the 16 regions of Korea. The estimation sample is households with two adults aged between 20-65 years old, who experienced no changes in their composition during the sample period.[8] The resulting sample size is 10,675 observations (2932 unbalanced panel households) or 4476 observations in the balanced panel of 746 households present in all six years.

The dependent variable is the budget share for food consumed at home, while control variables include real total income (deflated by the national CPI with a 2000 average base), relative food price changes, regional and time dummies, and demographic, educational and employment characteristics (Appendix Tables 1 and 2 have descriptive statistics). The instruments for income (*Z* in equation (8)) are educational levels and age of the head (averaging 12 years and 45 years) and their spouse and the number of children below age 15. The control variables (*X* in equation (5)) include regional dummies, the annual work hours of the head and the wife, household size, and demographic characteristics. The self-employment indicator is based on whether self-employment is the main job of the household head.[9]

The budget share for food out of the home is also included as a control variable. This form of consumption, with an average budget share of 3.2 percent, is not part of the dependent variable because it is assumed that restaurant meals are not perfect substitutes for food-at-home. Ideally, the substitution possibilities between restaurants and home cooking would be captured by including the relative price of restaurant meals but this is not available. Therefore, we follow the practice in the literature that uses Engel curves to measure CPI bias

[8] This removes effects of food budget share changes due to newly added members or exits of original members. Moreover, we also include time dummies to control for the potential changes in compliance behavior driven by the tax authorities' policy changes, and any trending macro variables and shocks during the sample period.

[9] The self-employment indicator also can be defined as the ratio of self-employment income to household total income above a certain threshold (e.g. >.25). The different definitions did not alter the results in Pissarides and Weber (1989). With a panel, in addition to the time-averaged value of the self-employed indicator during the sample period, we can consider as an alternative whether the household is ever defined as self-employed during the sample period or at the other extreme whether the household has maintained its self-employment status for the entire sample period. In between estimation results, there will be no difference in the estimated degree of under-reporting between the time-averaged one and the first alternative since both can count the under-reporting of transient self-employment. However, the latter one would understate the degree of under-reporting of the self-employment income since the under-reporting by the transient self-employed is not counted.

and we use the budget share for restaurant meals as an explanatory variable in place of the required price. Moreover, controlling for this form of food consumption may capture potential preference heterogeneity whereby self-employment income is more likely to be spent on food out of the home since this spending may be eligible for business deductions.

3. Estimation Results

The results of treating each year of the data as a separate cross-section, and then applying equation (11) to get the upper and lower bounds for $\bar{k}$ in each year are illustrated in Figure 1 for Korea and Figure 2 for Russia.[10] This follows the traditional PW method, but applying it in multiple years rather than to a single cross-section. Two problems with the traditional PW method are highlighted by these figures. First, there is considerable year-to-year variation in the position of the interval within which $\bar{k}$ is meant to lie. Over just a six year period for Korea the upper bound could be as high as 1.39 or as low as low as 1.22. Similarly the lower bound appears to vary between 0.95 and 1.11. There is even greater volatility in the position of the intervals in Russia. Hence, two researchers who both used the PW method on the same survey but each worked with data from a different year might reach substantially different conclusions about the severity of income under-reporting by the self-employed.

The second problem is the large gap between the upper and lower bounds that $\bar{k}$ is estimated to lie within. For Korea, the interval varies from 0.19 (in 2001) to 0.31 (in 2004), with an average interval over the six years of 0.25.[11] Similarly, in the Russian data the interval ranges from 0.06 to 0.69, with an average value of 0.25. Since the upper and lower

[10] The regression results for the year-by-year Engel curve estimates that the bounds for $\bar{k}$ are derived from are not reported, to save space. Similarly, the results for Russia that are referred to in the text are not reported. Both sets of results are available from the authors.

[11] If instead of estimating year-by-year 2SLS the data are pooled and equation (5) is estimated with year dummy variables included, and then equation (11) applied, the lower bound is estimated to be 1.13 and the upper bound to be 1.24. Hence the estimated interval of 0.11 from this pooled approach is smaller than the average interval of 0.25 from year-by-year 2SLS.

bounds are themselves stochastic, due to sampling error,[12] there is likely to be a great deal of uncertainty about the actual extent of under-reporting when using the traditional PW method.

Simply taking the mid-point of the intervals and averaging over the median for each year can give the appearance of exactness in estimating $\bar{k}$ but is unlikely to provide correct estimates. Such an approach would be consistent with several applications of the PW method, which use the median of the interval as their best estimate of $\bar{k}$, the under-reporting parameter.[13] Following this approach, the mean of the medians is 1.185 (1.60 for Russia) while the median of the medians is 1.199 (1.49 for Russia). As will be shown below, these estimates are quite different from those that result from applying equation (14) after between estimation on the panel.

If instead of following the original PW method we use the more exact panel data method outlined in Section II.3 we get substantially different results. The first step is to estimate the food Engel curves on the time-averaged values, using between estimation (reported in the first column of Table 3 for Korea and Table 4 for Russia). According to these estimates the food-at-home share in Korea is 1.9 percentage points lower for self-employed households who otherwise have the same reported income and same demographic characteristics as employee households. For Russia the gap is slightly larger, at 1.7 percentage points. The other key parameter readily apparent from Tables 3 and 4 is β, which is -0.111 in Korea and -0.089 in Russia. This negative and significant coefficient on the log transformed real income indicates that food shares fall as households become richer, which is precisely why food is used as the indicator good here. The ratio of δ, the coefficient on the dummy variable

[12] The standard errors for the upper and lower bounds that are calculated with the delta method range from 0.042 to 0.067.

[13] There is no necessary reason for choosing the mid-point of the interval as the best point estimate since the two sets of assumptions needed to derive the upper bound and lower bound are not necessarily equally realistic in any given setting.

for self employed households, to β, the coefficient on real income, provides part of the calculation for the extent of under-reporting.

When the Engel curve results from Table 3 and 4 are used in equation (14), the estimates of the under-reporting parameter $\bar{k}$ are higher than are any of the averages of midpoints (or the midpoints when the panel data are pooled) from the original PW approach reported above. Specifically, the results, which are reported in Table 5, show that for Korea $\bar{k} = 1.251$ (with a standard error of 0.021) and for Russia $\bar{k} = 1.230$ (standard error of 0.125). These estimates are from 5.6 percent (4.3 percent) higher than the mean (median) of the midpoints in Figure 1. If these estimates are transformed into an under-reporting rate $(=1-1/\bar{k})$ they imply that 20 percent of the income of self-employed households in Korea and 18.7 percent of the income of Russian self-employed households is not reported.[14]

The results in Table 5 appear to be robust to changes in the estimation sample. The first sensitivity check was to drop 56 potential outliers, having food-at-home shares that were either less than 0.05 or more than 0.80. This deletion changed the between estimate of $\bar{k}$ only slightly, from 1.251(±0.021) to 1.247(±0.021) when using the KLIPS data. The year-to-year variation in the position of the interval also reduces slightly, with the standard deviation of the mid-points falling from 0.057 to 0.054. Second, we restricted the analysis to the balanced KLIPS sample, and even though this reduced the number of observations from 10,675 to just 4476, the estimate of $\bar{k}$ changed only slightly, from 1.251(±0.021) to 1.254(±0.030). Third, we dropped 2904 observations from KLIPS where the household received some transfer income, since such income might be spent in a different way than other income and thereby

[14] As expected, the choice of self-employment indicators makes no difference in the estimated degree of under-reporting when the first alternative indicator (if the household is ever defined as the self-employed during the sample period) is used since both this and the time-averaged values count the under-reporting by transient self-employed. Using the most restrictive self-employment indicator (self-employed in all survey waves), the between estimate of $\bar{k}$ decreases, from 1.251 to 1.220 when using the KLIPS data.

change the food shares. This deletion also made only a small difference, changing the estimate of $\bar{k}$ from 1.251(±0.021) to 1.281(±0.026).

IV. Discussion and Conclusions

In this paper we have presented a refinement of the Pissarides and Weber (1989) method for estimating income under-reporting by the self-employed. Such estimates are important for measuring the size of the underground economy, which is relevant for tax policy. The original Pissarides and Weber method has been applied to household survey data in several countries but has two weaknesses. First, only an interval estimate of the under-reporting parameter $\bar{k}$ is possible. Second, even this interval relies on a troubling assumption that the degree of under-reporting is independent of the degree of transitory income fluctuations. These weaknesses both result from the Pissarides and Weber method typically being applied to cross-sectional data, which cannot distinguish between under-reporting and the likely higher variance of transitory income for the self-employed.

In contrast our panel data method allows us to untangle income under-reporting from transitory income fluctuations. Consequently we can provide an exact estimate of the degree of under-reporting rather than just an interval estimate. Moreover we do not need to assume that the degree of under-reporting is independent of the degree of transitory income variation. This allows for the possibility that the self-employed may increase their under-reporting rate as positive transitory income increases, which seems likely if they adopt a reporting rule based on monetary thresholds rather than proportions of true income.

We illustrate use of our method with panel data from Korea and Russia and estimate the under-reporting parameter $\bar{k}$ in each country. We find that the income under-reporting rates are 20.1 percent in Korea and 18.7 percent in Russia, so that the true incomes are 1.25 and 1.23 times the reported incomes for households with self-employment income. Our estimate of $\bar{k}$ is quite different from the mean (or median) of the midpoints of interval estimates that are derived from the traditional Pissarides and Weber approach estimated on cross-sections. Moreover, these interval estimates from the traditional Pissarides and Weber approach are sufficiently wide that they average 21 percent of the median of the midpoints in Korea and 17 percent in Russia. This wide range of estimates of the extent of under-reporting may be too large to be of practical value for guiding tax policy.

Our method relies on between estimation where the mean value of reported incomes over time for the same household is used as the data in the regression. This use of household-specific means enables transitory income variations to be controlled for. In our illustration we used 6-year averages in both countries to control for the variations in transitory income over time. One outstanding question is whether this is a large enough T to make the variations of transitory income disappear and the covariance between the degree of under-reporting and the degree of transitory income variation disappear. One argument in support of this time period is that in the literature on intergenerational income mobility (Solon, 1992), this same multi-year average has been used extensively to correct for errors-in-variable bias arising from the variations of transitory income. In most cases in this literature the maximum T is five so it may be reasonable to assume that in our illustration a T=6 is sufficient to control for the transitory income variations. A useful task for future research would be to apply our method to longer panels in order to see if the choice of T has any bearing on the resulting estimates of income under-reporting.

References

Chung, C., Gibson, J., and Kim, B. 2010. "CPI mis-measurement and their impacts on economic management in Korea" *Asian Economic Papers* 9(1): 1-15.

Costa, D. 2001. "Estimating real income in the United States from 1888 to 1994: Correcting CPI bias using Engel curves" *Journal of Political Economy* 109(6): 1288-1310.

Gibson, J., Stillman, S., and Le, T. 2008. "CPI bias and real living standards in Russia during the transition" *Journal of Development Economics* 87(1): 140-160.

Gollin, D. 2002. "Getting income shares right" *Journal of Political Economy* 110(2): 458-474.

Feldman, N. and Slemrod, J. 2007. "Estimating tax noncompliance with evidence from unaudited tax returns" *Economic Journal* 117(March): 327-352.

Hamilton, B. 2001a. "Using Engel's Law to estimate CPI bias" *American Economic Review* 91(3): 619-630.

Hamilton, B. 2001b. "Black-White difference in inflation: 1974-1991" *Journal of Urban Economics* 50(1): 77-96.

Johansson, E. 2005. "An estimate of self-employment income underreporting in Finland" *Nordic Journal of Political Economy* 31(1): 99-109.

Johnson, S., Kaufmann, D., and Shleifer, A. 1997. "The unofficial economy in transition." *Brookings Papers on Economic Activity* 2: 159-221.

Lyssiotou, P., Pashardes, P. and Stengos, T. 2004. "Estimates of the black economy based on consumer demand approaches" *Economic Journal* 114(July): 622-640.

Pissarides, C. and Weber, G. 1989. "An expenditure based estimate of Britain's black economy" *Journal of Public Economics* 39(1): 17-32.

Schuetze, H. 2002. "Profiles of tax noncompliance among the self-employed in Canada: 1969-1992" *Canadian Public Policy* 28(2): 219-237.

Solon, G. 1992. “Intergenerational income mobility in the United States” *American Economic Review* 82(3): 393-408.

Tedds, L. 2010. “Estimating the Income Reporting Function for the Self-Employed” *Empirical Economics* 38(3): 669-687.

Thomas, J. 1999. “Quantifying the black economy: measurement without theory yet again” *Economic Journal* 109: F381-387.

Figure 1. Upper and lower bound and interval for under-reporting parameter $\bar{k}$ using the Pissarides and Weber method on KLIPS data, 2000-2005.

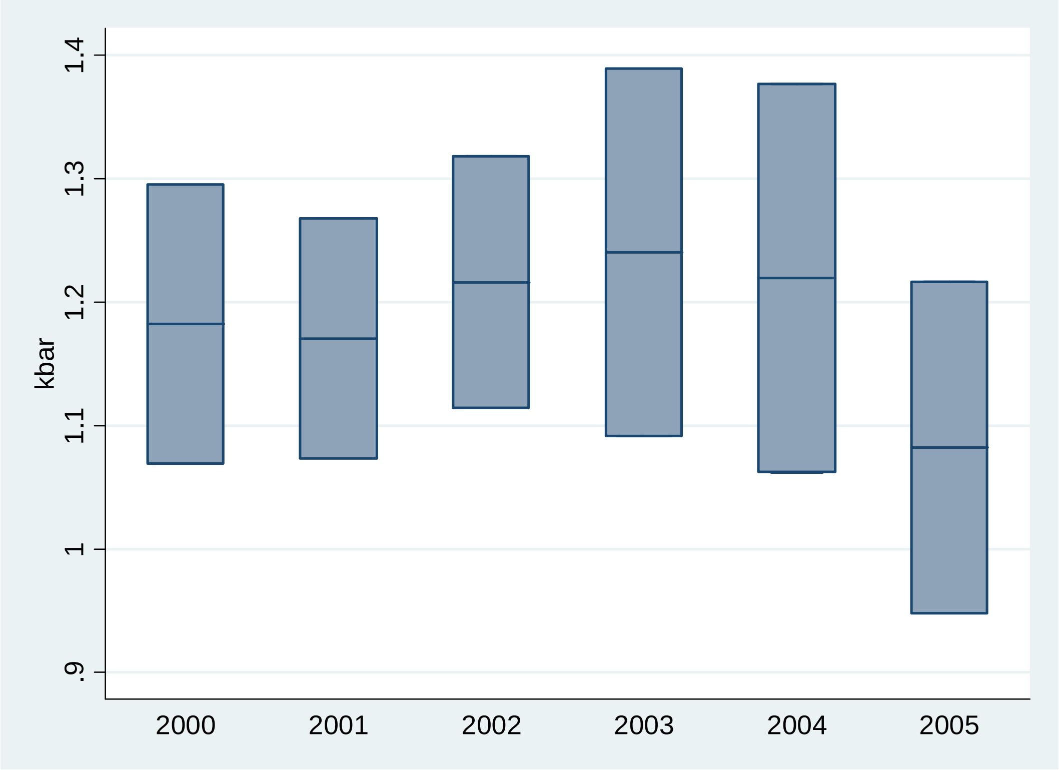

Figure 2. Upper and lower bound and interval for under-reporting parameter $\bar{k}$ using the Pissarides and Weber method on RLMS data, Rounds 5-10.

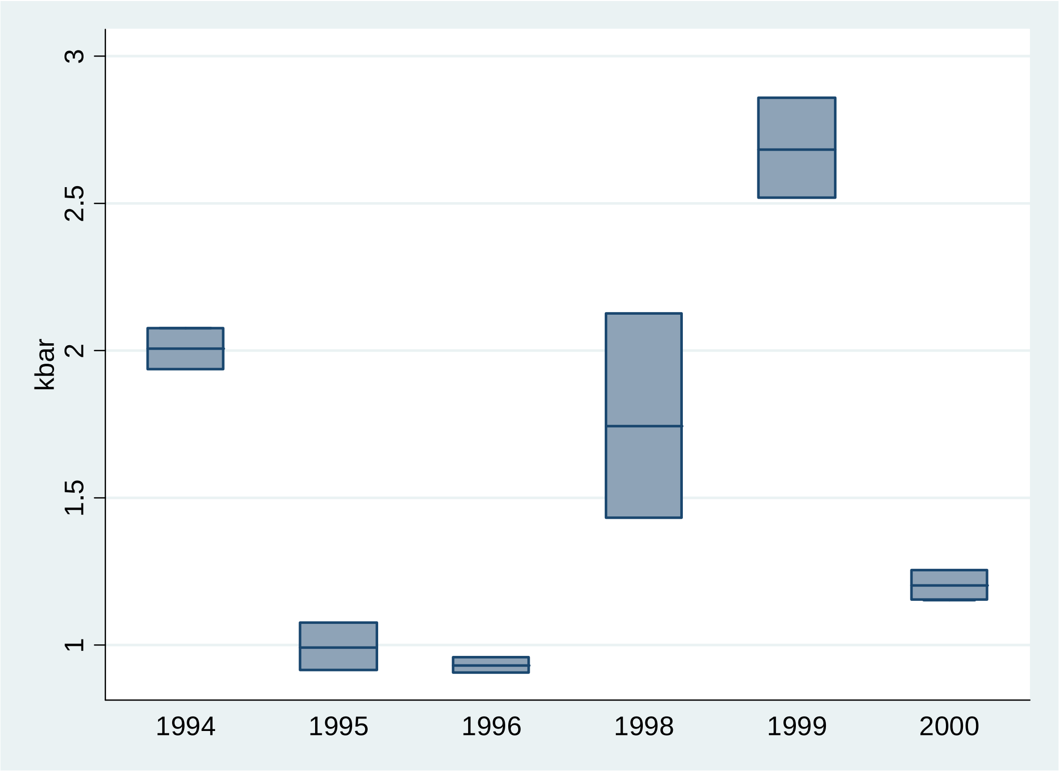

Table 1. Trend of main variables over time (KLIPS, 2000-05), obs.=10675

Variable	*Employees*			*Self-employed*		
	2000	2003	2005	2000	2003	2005
w (Food Expenditure Share at Home)	.301	.217	.181	.296	.203	.185
X_{res} (Food Expenditure Share out of Home)	.036	.031	.033	.030	.027	.026
$\ln(Y/P)$ (Log Transformed Real Household Income)	16.92	17.21	17.28	16.83	17.09	17.14

Table 2. Trend of main variables over time (RLMS, 1994-00), obs.=6680

Variable	*Employees*			*Self-employed*		
	Round 5 (1994)	Round 7 (1996)	Round 10 (2000)	Round 5 (1994)	Round 7 (1996)	Round 10 (2000)
w (Food Expenditure Share at Home)	.564	.540	.465	.498	.473	.408
X_{res} (Food Expenditure Share out of Home)	.046	.039	.042	.056	.046	.049
$\ln(Y/P)$ (Log Transformed Real Household Income)	12.83	12.60	12.78	13.25	13.07	13.08

Table 3. Key Parameter Estimates and Upper and lower bound Estimates for $\bar{k}$ (KLIPS, 2000-05)

Para-	2000	2001	2002	2003	2004	2005	Pooled	BE
meters	n=1779	n=1803	n=1848	n=1779	n=1779	n=1687	n=10675	n=10675
$\hat{\delta}$	-.018	-.019	-.022	-.024	-.019	-.007	-.019	-.019
	(.006)	(.005)	(.005)	(.005)	(.005)	(.004)	(.002)	(.002)
$\hat{\beta}$	-.107	-.126	-.116	-.116	-.097	-.100	-.112	-.111
	(.013)	(.012)	(.011)	(.010)	(.010)	(.008)	(.005)	(.004)
$\bar{k}_{LB}$	1.069	1.073	1.114	1.091	1.062	0.947	1.128	
$\bar{k}_{UB}$	1.295	1.267	1.318	1.389	1.376	1.216	1.239	

Table 4. Key Parameter Estimates and Upper and lower bound Estimates for $\bar{k}$ (RLMS, 1994-00)

Para- meters	Round5 n=1334	Round6 n=1174	Round7 n=1105	Round8 n=1075	Round9 n=1066	Round10 n=1126	Pooled n=6880	BE n=6880
$\hat{\delta}$	-.036 (.017)	.001 (.017)	.011 (.021)	-.027 (.019)	-.030 (.017)	-.019 (.007)	-.017 (.007)	-.017 (.007)
$\hat{\beta}$	-.052 (.027)	-.129 (.026)	-.160 (.030)	-.049 (.027)	-.030 (.031)	-.102 (.030)	-.089 (.012)	-.089 (.012)
$\bar{k}_{LB}$	1.9377	0.9115	0.959	1.431	2.858	1.151	1.185	
$\bar{k}_{UB}$	2.0761	1.0753	0.903	2.125	2.518	1.254	1.230	

Table 5. Exact Estimates of Income Under-Reporting by the Self-Employed

	(1) Korea (KLIPS, 2000-05)	*(2) Russia* (RLMS, 1994-2000)
Under-reporting parameter, $\bar{k}$	1.251 (.021)	1.230 (.124)
Under-reporting rate $(= 1 - 1/\bar{k})$	0.201	0.187

Note: The estimates are calculated using equation (14) in the text, and based on the between estimates of the Engel curve results in the first columns of Tables 3 and 4. Standard errors in ( ) are from the delta method.

Appendix: Description of the RLMS Dataset

The Russian Longitudinal Monitoring Survey (RLMS) is also an on-going nationally representative longitudinal household survey, designed and implemented by the Carolina Population Center, University of North Carolina, in collaboration with the Russian Academy of Sciences and the Russian Institute of Nutrition. RMLS collects data on an exhaustive list of individual and household characteristics including detailed expenditure data. We use six waves of data from Phase II, which began in 1994 and collects data annually or bi-annually from approximately 4,000 households.[15] The sampling is based on a division of Russia into 38 strata, with one primary sampling unit (PSU) chosen from each stratum.

The dependent variable is the budget share for food consumed at home, while control variables include real total income (deflated by the CPI with a November 1994 base), relative food price changes, demographic, educational and employment characteristics, indicators of dwelling characteristics, an indicator for whether the household head or spouse is self-employed and the budget share for food out of the home. The self-employment variable is based on whether the household head or their spouse is either an owner or co-owner of the enterprise where they work.

A description of the dependent and explanatory variables is shown in Appendix Table 2. The expenditure share of food consumption at home averages 51.5 percent for the sample period. The household head averages 43.4 years old and 27.7 percent of household heads have tertiary education. Spouses are about three years younger in age and 28.7 percent have tertiary education. The share of self-employed households averages 21.6 percent for the sample period.

[15] Surveys were conducted in late autumn of 1994, 1995, 1996, 1998, 2000, and 2001 with fieldwork typically centered on the month of November.

Appendix Table 1. Descriptive Statistics of the KLIPS data, obs.=10675

Variable	*Mean*	*S.D.*	*Min*	*Max*
w (Food Expenditure Share at Home)	.228	.106	.013	.9
X_{res} (Food Expenditure Share our of Home)	.032	.036	0	.4
$\ln(Y / P)$ (Log Transformed Household Real Income)	17.10	.655	10.66	20.29
Age of Householder	45.18	8.41	21	65
Age of Spouse	42.03	8.14	20	65
Education Years of Householder	12.77	3.29	0	27
Education Years of Spouse	11.08	3.10	0	27
Yearly Hours of Work of Householder	2738.02	1091.04	0	8400
Yearly Hours of Work of Spouse	1310.96	1465.99	0	8400
Dummy: Self-Employed	.375	.484	0	1
Household Size	3.959	.855	2	9
Number of children under 15 years old in the household	1.357	.8952	0	4

Appendix Table 2. Descriptive Statistics of the RMLS data, obs.=6880

Variable	*Mean*	*S.D.*	*Min*	*Max*
w (Food Expenditure Share at Home)	.515	.219	.0133	.989
X_{res} (Food Expenditure Share our of Home)	.044	.087	0	.830
$\ln(Y/P)$ (Log Transformed Household Real Income)	12.70	.944	7.16	16.54
Age of Householder	43.38	10.32	21	65
Age of Spouse	40.90	11.17	21	65
Dummy: Tertiary Education for Head	.277	.447	0	1
Dummy: Tertiary Education for Spouse	.287	.452	0	1
Yearly Hours of Work of Head	1409.16	1165.59	0	7000
Yearly Hours of Work of Spouse	1343.91	1161.49	0	7000
Dummy: Self-Employed	.216	.411	0	1
Ln (household size)	1.107	.290	.693	2.302
% of household ≤ 2 years old	.021	.077	0	0.5
% of HH 3-14 year old boys	.086	.144	0	0.6
% of HH 3-14 year old girls	.085	.144	0	0.6
% of HH 15-17 year old boys	.024	.080	0	0.5
% of HH 15-17 year old girls	.023	.079	0	0.5
Dummy: detached dwelling	.076	.265	0	1