UNIVERSITY OF WAIKATO

Hamilton
New Zealand

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

Bonggeun Kim, John Gibson and Chul Chung

Department of Economics

Working Paper in Economics 15/08

October 2008

Corresponding Author

John Gibson
Department of Economics
University of Waikato,
Private Bag 3105,
Hamilton, New Zealand

Fax: +64 (7) 838 4331
Email: jkgibson@waikato.ac.nz

Bonggeun Kim
Sungkyunkwan University **and** ***University of Waikato***

Chul Chung
Korea Institute for International Economic Policy

Abstract

The income of the self-employed is often assumed to be understated in economic statistics. Debate exists about the extent of under-reporting and the resulting measures of the size of the underground economy. This paper refines a method developed by Pissarides and Weber (1989) and uses discrepancies between food shares and reported incomes to estimate under-reporting by the self-employed. In contrast to previous studies our panel data methodology distinguishes income under-reporting from transitory income fluctuations of the self-employed, and provides an exact estimate of the degree of under-reporting rather than just an interval estimate. Using panel data from Korea and Russia we estimate that 38 percent of the income of self-employed households in Korea and 47 percent of the income of Russian self-employed households is not reported.

Keywords
Engle curve
measurement error
self-employment
underground economy

JEL Classification
D12, H26, H31, O17

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, Trinh Le, Steven Stillman and participants in the 2008 ESAM meetings. All remaining errors in this paper are those of the authors.

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 while only employees correctly report incomes. The relationship between food and income for employees is used to back out a range of estimates for true self employment income. That only a range can be estimated reflects the weakness of cross-sectional data, which cannot distinguish between under-reporting and the likely greater deviations of current income from 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) which is a useful refinement if self-employment income is not spent in the same way as other income, since preference heterogeneity may be confused with income under-reporting.[2]

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

1 A much larger literature relies on macroeconomic 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 macroeconomic approaches (Thomas, 1999).

2 For example, households may reserve self-employment income for 'big ticket' items and use wages for food and other regular expenses. A drawback of full demand systems is that they will include certain expenditure items that may qualify as business expenses and there could be measurement error in these for the self-employed. Such errors do not affect approaches that rely on reported food expenditures.

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.

These methodological refinements may be important since accurate measurement of income underreporting by the self-employed matters both to correct measurement of GDP and to 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 underground economy may help the tax authorities decide on their best strategy. Also, correctly measuring self-employment income is important for many models of growth and aggregate technology that assume that functional income shares should be identical 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; Beatty and Larson, 2005). 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.

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 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:

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

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 variations 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 the degree of income under-reporting, and it is the factor (assumed to be 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:

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

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}:

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

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:

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

where Z is a set of proxy variables representing the permanent income. The composite error term 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:

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

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:

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

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:

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

3. A More Exact Panel Data Method

With panel data it is possible to make an exact point estimate of the degree of income under-reporting by the self-employed. This exact estimate contrasts with the interval estimate from the Pissarides and Weber approach on 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 $\sum_{t=1}^{T} \ln y_{it}^* / T$. This household-specific mean allows the positive and negative variations of transitory income over time to cancel each other out, since:

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

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

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_{vSE}^2 the estimator of interest is:

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

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.

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. The survey data for each country have been used in a number of other published papers. For example, the KLIPS data were used in Chung, Kim and Park (2007) and the RLMS data in a study of CPI bias 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 has difficulty in capturing consumption from own production, which is likely to be more important in rural areas. We also restrict attention to households that have 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. Descriptive statistics for the variables used in the analysis are in Appendix Table 1 and 2. Full details on the surveys and the construction of the variables are reported in the Appendix. Control variables include relative food price changes, demographic and educational characteristics, hours of work, and the expenditure share for food out of home.[3]

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.

Table 1. Trend of main variables over time (KLIPS, 2000-05), obs.=6593

Variable	*Employees*			*Self-employed*		
	2000	2003	2005	2000	2003	2005
w (Food Expenditure Share at Home)	.303	.218	.184	.292	.208	.185
X_{res} (Food Expenditure Share out of Home)	.040	.033	.034	.033	.030	.030
$\ln(Y/P)$ (Log Transformed Real Household Income)	16.87	17.12	17.19	16.78	17.08	17.05

[3] This form of consumption 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 and we use the budget share for restaurant meals as an explanatory variable in place of the required price.

This fall in the food share is large relative to the measured growth in real income, which is consistent with the existence of a substantial CPI bias in Korea, as found by Chung, Kim and Park (2007). The implied CPI bias appears even more substantial for Russia since the first row of Table 2 shows that the average food-at-home share fell by about 10 percentage points during the sample period, but the average real household income apparently decreases. Indeed, Gibson, Stillman and Le (2008) report a large CPI bias for Russia from these same data. However these potential CPI biases should not affect the results reported below, since the same CPI is used for both self-employed and employee households. Moreover, our main aim in the empirical section is to demonstrate how the use of panel data may give a more exact estimate of income under-reporting than is possible with the original PW approach rather than to justify a particular value for the under-reporting estimates in these two countries.

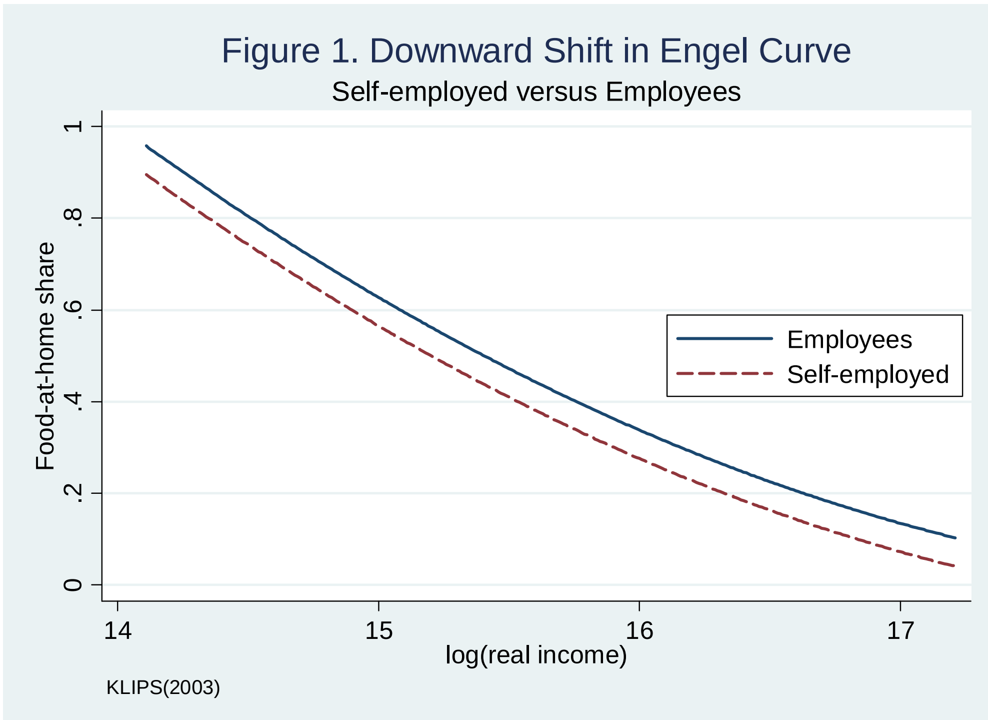
Table 2. Trend of main variables over time (RMLS, 1994-00), obs.=5243

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)	.561	.542	.466	.527	.510	.432
X_{res} (Food Expenditure Share out of Home)	.048	.037	.037	.047	.035	.049
$\ln(Y/P)$ (Log Transformed Real Household Income)	12.83	12.62	12.79	13.11	12.88	12.99

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. Figure 1 illustrates this pattern using the food shares for the self-employed households and employee households in the KLIPS of 2003. We attribute this downward shift to unmeasured real income of the self-employed, which in turn is due to the under-reporting of nominal income by the self-employed.[4]



2. Estimation Methods

Equation (5) is a linear model and can be estimated separately for each year using OLS. In other words, one could treat the panels as six annual cross-sections. Such an approach would be consistent with the PW method and would yield a separate interval estimate for $\bar{k}$ in each year. However, 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. This application of OLS to six-year average values controls for the variations of transitory income, following equation (12). The resulting estimate for $\bar{k}$ will be a single value, since there is no need to make an interval estimate and since the year-by-year fluctuations also disappear.

---

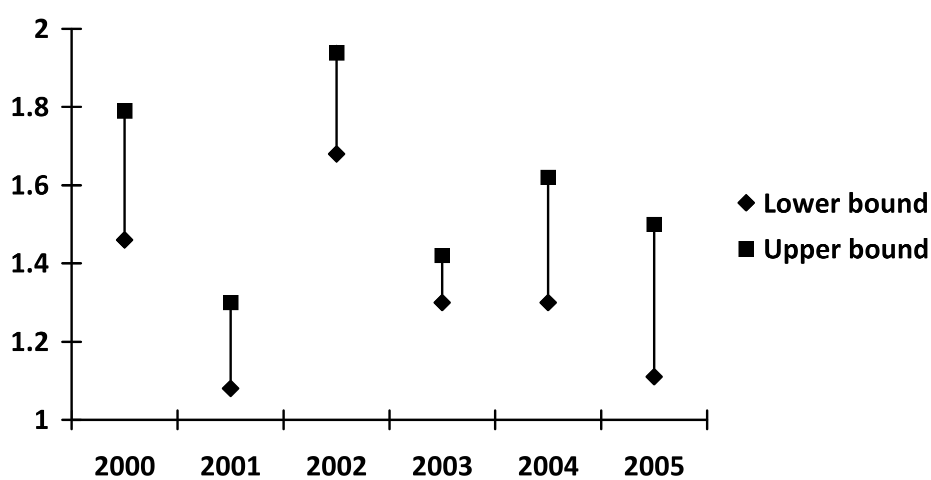
4 An alternative explanation can be considered for the lower food shares of the self-employed. Like Hamilton (2001b), the two occupational groups could face a differential CPI bias and the shift could result from the higher CPI bias for the self-employed than for the employees. However the difference from Hamilton (2001b) is that while Blacks and Whites are geographically segregated in the U.S., there is no similar segregation by employment status in either Korea or Russia or more generally. This lack of geographical segregation rules out one plausible source of differential CPI bias which is that living in different areas could contribute to differential outlet bias, whereby the statistics agency continues surveying prices at base period outlets while households have shifted to shopping at cheaper outlets.

In many settings researchers apply another estimator to panel data, which is within estimation (also known as the fixed effects estimator). Rather than studying variations across mean values for cross-sectional units, this estimator looks at variations over time within units and can therefore allow fixed but unobservable household-specific effects to drop out of the analysis. This fixed effects estimator also can be given a particular interpretation in our current context. If there is an intrinsic tendency for under-reporting income, such as for tax evasion purposes, people may self-select into self-employment since it offers potentially greater scope for disguising income than is possible for employees. Since we can control for such intrinsic tendencies with the fixed effects model it might be expected to yield smaller coefficients on the dummy variable for self-employment than does between estimation, which does not control for fixed effects. Therefore a comparison of coefficients from between and within estimation may reveal something about the underlying causes of income under-reporting by the self-employed.

3. Empirical Results

We first estimated equation (5), treating each year of the data as a separate cross-section, and then applied equation (11) to get the upper and lower bounds for $\bar{k}$ in each year. This approach follows the traditional PW method, but applying it in multiple years rather than to a single cross-section. The resulting estimates of the upper bound, lower bound and the interval within which the under-reporting parameter $\bar{k}$ lies are illustrated in Figure 2 for the case of Korea.[5]

Figure 2. Upper and lower bound and interval for under-reporting parameter $\bar{k}$ using the Pissarides and Weber method on KLIPS data, 2000-2005



[5] 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.

Two problems with the traditional PW method are highlighted by the results in Figure 2. The first problem is the large gap between the upper and lower bounds that $\bar{k}$ is estimated to lie within. The interval varies from 0.12 (in 2003) to 0.39 (in 2005), with an average interval over the six years of 0.29.[6] Similarly, when the same approach is applied to the Russian data the interval ranges from 0.03 to 0.38, with an average value of 0.22. Since the upper and lower bounds are themselves stochastic, due to sampling error,[7] there is likely to be a great deal of uncertainty about the actual extent of under-reporting when using the traditional PW method.

The second problem apparent in Figure 2 is that 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 1.94 or as low as low as 1.42. Similarly the lower bound appears to vary between 1.08 and 1.46. 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.

Simply taking the mid-point of the intervals in Figure 2 and then averaging over these medians across the years gives 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.[8] Following this approach, the mean of the medians is 1.45 (1.58 for Russia) while the median of the medians is 1.41 (1.21 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).

6 If instead of estimating year-by-year OLS 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 1.43 and the upper bound to be 1.75. Hence the estimated interval of 0.32 from this pooled approach is very similar to the average interval of 0.29 from year-by-year OLS.

7 The standard errors for the upper and lower bounds that are calculated with the delta method range from 0.11 to 0.29.

8 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.

According to these estimates the food-at-home share in Korea is 2.1 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 2.5 percentage points. The other key parameter readily apparent from Tables 3 and 4 is β, which is -0.05 in Korea and -0.04 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 for self employed households, to β, the coefficient on real income, provides part of the calculation for the extent of under-reporting. It is apparent from comparing these ratios that under-reporting is potentially a larger problem in Russia.

Table 3. Food Engel Curve Estimations of Korea (KLIPS, 2000-05), obs.=6593

Variable	*(1) Between OLS* (KLIPS, 2000-05)	*(2) Fixed Effect* (KLIPS, 2000-05)
Intercept	1.2635 (.0571)***	1.4664 (.1351)***
Log (Real Household Income)	-.0545 (.0035)***	-.0171 (.0028)***
Log (Food CPI/Non-food CPI)	-.6559 (.0522)***	.1881 (.0858)**
Dummy: Self-employed	-.0205 (.0040)***	-.0067 (.0053)
Food Expenditure Share out of home	-.1583 (.0560)***	-.0951 (.0382)**
Age of Householder	-.0001 (.0005)	.0324 (.0305)
Age of Spouse	.0001 (.0006)	-.0590 (.0305)*
Education Years of Householder	-.0032 (.0008) ***	-.0036 (.0023)
Education Years of Spouse	-.0021 (.0009)	.0023 (.0025)
Yearly Hours of Work of Householder	2.13e-9 (1.95e-09)	-6.47e-10 (1.41e-09)
Yearly Hours of Work of Spouse	-8.88e-10 (1.45e-09)	-2.88e-10 (1.28e-09)
Number of children under 15 years old in the household	.0060 (.0022)***	.0016 (.0031)
$\mathbf{R^2}$	.3039	.2183

Note: ***, **, * represent the levels of statistical significance of 1%, 5%, and 10% respectively.

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 $\bar{k} = 1.614$ (with a standard error of 0.112) and for Russia $\bar{k} = 1.880$ (standard error of 0.596). These estimates are from 11-19 percent (33 percent) higher than the mean (median) of the midpoints in Figure 2. If these estimates are transformed into an under-reporting rate $(= 1 - 1/\bar{k})$ they imply that 38 percent of the income of self-employed households in Korea and 47 percent of the income of Russian self-employed households is not reported.

Table 4. Food Engel Curve Estimations of Russia (RMLS, 1996-2000), obs.=5243

Variable	*(1) Between Estimation* (RMLS, 1994-2000)	*(2) Fixed Effect* (RMLS, 1994-2000)
Intercept	.8147 (.0760)***	1.4856 (.0915)***
Log (Real Household Income)	-.0403 (.0055)***	-.0368 (.0041)***
Log (Food CPI/Non-food CPI)	.1100 (.0436)**	.0589 (.0344)*
Dummy: Self-employed	-.0253 (.0118)**	-.0178 (.0091)*
Food Expenditure Share out of home	-.4824 (.0618)***	-.4919 (.0377)***
Age of Householder	.0023 (.0012)*	-.0047 (.0021)**
Age of Spouse	.0027 (.0011)**	-.0064 (.0019)***
Dummy: Tertiary Education for Head	-.0456 (.0106)***	-.0095 (.0162)
Dummy: Tertiary Education for Spouse	.0045 (.0098)	.0293 (.0161)*
Yearly Hours of Work of Head	-6.15e-6 (4.51e-06)	-1.04e-6 (3.21e-06)
Yearly Hours of Work of Spouse	-6.76e-6 (4.68e-06)	-1.04e-6 (3.09e-06)
ln (household size)	.0407 (.0248)	-.0262 (.0292)
% of household ≤ 2 years old	.1262 (.0802)	.2477 (.0692)***
% of HH 3-14 year old boys	.0921 (.0473)*	.0836 (.0555)
% of HH 3-14 year old girls	.1220 (.0467)***	.2349 (.0557)***
% of HH 15-17 year old boys	.1004 (.0752)	.0574 (.0529)
% of HH 15-17 year old girls	-.1330 (.0819)	.2234 (.0528)***
Dummy: detached dwelling	-.0259 (.0158)	.0342 (.0285)
R^2	.1876	.0994

Note: ***, **, * represent the levels of statistical significance of 1%, 5%, and 10% respectively.

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, $\bar{k}$	1.614 (0.112)	1.880 (0.596)
Under-reporting rate $(= 1 - 1/\bar{k})$	0.380	0.468

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.

The results in Table 5 appear to be robust to changes in the estimation sample. The first sensitivity check was to drop 22 observations that were potential outliers, having food-at-home shares that were either less than 0.05 or more than 0.80. This deletion changed the estimate of $\bar{k}$ only slightly, from 1.614 (±0.112) to 1.605 (±0.110) when using the KLIPS data. The second sensitivity check was to drop 1588 observations where the household received some transfer income, since such income might be spent in a different way than other income and thereby change the food shares. This deletion also made only a small difference, changing the estimate of $\bar{k}$ to 1.546 (±0.126) when using the KLIPS data.

In addition to these sensitivity analyses we also changed the estimation method from between estimation to within estimation. The Engel curve results when household-specific fixed effects are included in the regression are reported in the second columns of Tables 3 and 4. According to these within estimates, the food share in Korea is only 0.7 percentage points lower (and insignificantly different from zero) for self-employed households who otherwise have the same reported income and same demographic characteristics as employee households, while for Russia the food share is 1.8 percentage points lower. In both countries, the within estimates for the coefficient on the self-employed dummy variable are smaller than the between estimates. Hence, the impact of adding the household-specific fixed effects appears consistent with the hypothesis that people with an intrinsic tendency to under-report for tax evasion purposes may self-select into self-employment. If instead, the under-reporting behavior was mainly from the occupational characteristics then the addition of the household-specific fixed effects would not have been expected to have the same attenuating impact on the coefficient estimates.

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 traditional Pissarides and Weber method using 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 38.0 percent in Korea and 46.8 percent in Russia, so that the true incomes are 1.61 and 1.88 times the reported incomes for households with self-employment income. Our estimate of $\bar{k}$ is 11-19 percent (33 percent) higher than the mean (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 (18 percent in Russia). This wide range of estimates for 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 multiyear 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

Beatty, T. and Larsen, E. 2005. "Using Engel curves to estimate bias in the Canadian CPI as a cost of living index" *Canadian Journal of Economics* 38(2): 482-499.

Chung, C., Kim, B., and Park, M. 2007. "CPI bias in Korea," *Journal of International Economic Studies* 11(2): 261-284.

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.

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.

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

Appendix
Description of the Datasets

Korea

The Korean data are drawn from the Korean Labor Income Panel Study (KLIPS) an on-going nationally representative longitudinal household survey fielded since 1998 by the Korea Labor Institute. KLIPS collects data on an exhaustive list of individual and household characteristics including detailed income and expenditure data. We use six rounds of KLIPS data from 2000 to 2005,[9] and combine these with the annual CPI for food and non-food that is calculated for each of the 16 regions of Korea. We use a sample of two-adult families which are headed by a man, with or without children, where the adults are between 20-65 years old. We drop the households who had experienced changes in their composition during the sample period to remove the effects of food consumption changes due to newly added members or exits of original members. The resulting sample size is 6593 households.

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 2000 average base), relative food price changes, demographic, educational and employment characteristics. The model also includes the budget share for food out of the home. This form of consumption 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. The self-employment variable is based on whether self-employment is the main job of the household head.

A description of the dependent and explanatory variables is shown in Appendix Table 1. The dependent variable, which is the expenditure share of food consumption at home, averages 23.8 percent for the sample period. The share of food out of home averages 3.6 percent. Reported real total household income including labor income and financial income averages 3,400 million Korean won which is approximately equal to USD 30,000 in 2003. On average the household head is 41.2 years old and has 12.7 years of schooling while the spouse has one year less and is about three years younger. The share of self-employed averages 33.5 percent which did not change much during the sample period.

[9] The collection of data on food expenditure at home starts only in 2001, so earlier waves of KLIPS data cannot be used in this study.

Russia

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.[10] 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 52.6 percent for the sample period. The household head averages 44.3 years old and 25.6 percent of household heads have tertiary education. Spouses are about three years younger in age and 28.4 percent have tertiary education. The share of self-employed households averages 25.6 percent for the sample period.

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

Appendix

Table 1. Descriptive Statistics of the KLIPS data, obs.=6593

Variable	*Mean*	*S.D.*	*Min*	*Max*
w (Food Expenditure Share at Home)	.2317	.1048	.013 2	.9
X_{res} (Food Expenditure Share our of Home)	.0361	.0369	0	.4
$\ln(Y/P)$ (Log Transformed Household Real Income)	17.04	.6200	10.6 4	19.78
Age of Householder	41.40	7.20	21	65
Age of Spouse	38.48	7.07	20	65
Education Years of Householder	12.77	3.03	0	27
Education Years of Spouse	11.91	2.73	0	25
Yearly Hours of Work of Householder	2743	1010	0	8400
Yearly Hours of Work of Spouse	1171	1393	0	8400
Dummy: Self-Employed	.3321	.4710	0	1
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.=5243

Variable	*Mean*	*S.D.*	*Min*	*Max*
w (Food Expenditure Share at Home)	.526	.220	.0152	.989
X_{res} (Food Expenditure Share our of Home)	.040	.084	0	.830
$\ln(Y/P)$ (Log Transformed Household Real Income)	12.70	.934	7.16	16.54
Age of Householder	44.29	10.02	21	65
Age of Spouse	41.88	10.94	21	65
Dummy: Tertiary Education for Head	.256	.436	0	1
Dummy: Tertiary Education for Spouse	.284	.451	0	1
Yearly Hours of Work of Head	1382.86	1155.92	0	5600
Yearly Hours of Work of Spouse	1351.70	1141.78	0	7000
Dummy: Self-Employed	.250	.433	0	1
Ln (household size)	1.107	.295	.693	2.302
% of household ≤ 2 years old	.0178	.0708	0	0.5
% of HH 3-14 year old boys	.0846	.1437	0	0.6
% of HH 3-14 year old girls	.0843	.1442	0	0.6
% of HH 15-17 year old boys	.0248	.0802	0	0.5
% of HH 15-17 year old girls	.0244	.0797	0	0.5
Dummy: detached dwelling	.081	.274	0	1