Do Survey Income Reports Mimic Tax Return Records?

The role of measurement error in survey Vs register data

Ana Cabral & Norman Gemmell

Formerly, Inland Revenue *Chair in Public Finance, VBS*

Introduction

- Research questions:
  1. By how much do the self-employed under-report their incomes to the tax authority? (Cabral & Gemmell, 2017)
  2. Does reliance of household surveys (containing measurement error) bias results from these exercises?

1.

- Established Pissarides and Weber (PW, 1989) methods regularly used to estimate income under-reporting by the self-employed.
- Relies on estimating Engle curves (relating consumption expenditure to incomes for the each group) and identifying 'shifts' between employees and self-employed

2.

- Measurement error – 'validation studies' [e.g. in labour market literature] test for regression impact of using reported versus true records of employee incomes ⇒ 'attenuation biases' in regression parameters.
- Can tax return data be used to 'validate' survey-based underreporting estimates *for the self-employed?*

The PW Model

- We have two types of households: self-employed and employed households.
- All households, i, are assumed to report their expenditure on items, j, this is, C_{ij} , correctly.
- Income, however, is assumed to be reported correctly by employed households, hence their true income Y_i^T equals their reported income Y_i^R, $Y_i^R = Y_i^T$
- But self-employment income may be misreported. Thus for self-employed households

$$Y_i^T = k Y_i^R \qquad k>1 \qquad (1)$$

where k is a random variable that captures the factor by which self-employed income has to be scaled to arrive to their true income. Note that for the employed it follows that $k = 1$.

- From (1), the under-reporting 'income gap', κ, is:

$$\kappa = 1 - \frac{1}{k}$$

The PW Model

The expenditure function for household i for each item of expenditure j can be written as

$$\ln C_{ij} = \beta_j \ln Y_i^P + A' X_i + \varepsilon_{ij},$$

where β_j is the elasticity of consumption for good j, Y_i^P represents permanent income, X_i is a vector of household characteristics, and ε_{ij} is a white noise error.

Empirically, ...

$$\ln C_{ij} = \beta_j \ln Y_i^R + \gamma_j SE_i + \Theta_i' X_i + \Xi_{ij},$$

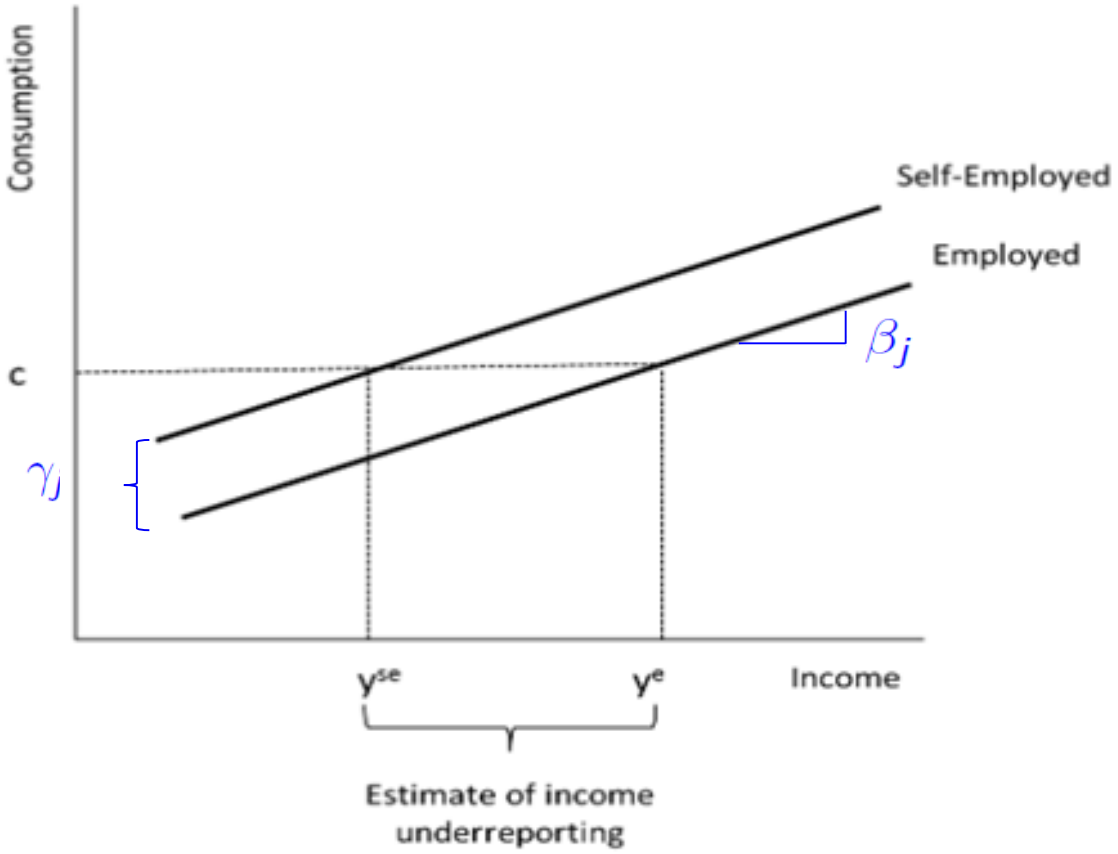
An estimation of the scaling factor k can be obtained from the parameters β and γ as,

$$k = exp\left(\frac{\gamma_j}{\beta_j}\right)$$

and the corresponding income-gap κ,

$$\kappa = 1 - \frac{1}{k}.$$

Engle Curves



Survey versus Register Data

- Most PW studies: rely on survey sources for reported income and expenditure data … and difference between employees and self-employed
- Slemrod and Weber (2011): Income-gaps obtained from the survey provide a valid estimation *if reports to the survey = reports to tax administration*.
- But survey data subject to measurement error
- Previous "validation studies" of labour market variables (wages, hours worked) compare survey reports to employer (PAYE) or tax records:
  - Applied to *employees only*
  - Confirm attenuation biases when, e.g. wages used as explanatory variable
- We compare under-reporting results using *survey*-based versus *register*-based income data (all expenditures are survey-only)

⇒ We find:

- Income under-reporting estimates ***much lower*** using survey (HES) income data
- This substantively due to attenuation biases; but especially to lower income reports on average to the register by the self-employed
- i.e. "survey answers are noisy and mean biased" (Kreiner et al. 2015, for Denmark)

Data and Self-Employment Definitions

- Full Expenditure HES questionnaire: 2006/07; 09/10; 12/13.
- Everyone in the household has been successfully matched to their tax records (if applicable).
- HRP is in employment and less than 60 years of age (Aguiar and Hurst, 2005)
- 2500 households.

Definition of Self-Employment

- *Opportunity Definition*✓: A household is defined as self-employed if it draws any income from self-employment sources (net profit, shareholder salary, income from partnership); and employed otherwise. Register data allows us to identify the legal form (not rely on self-reports from survey).
- *25% rule*: Self-Employed if more than 25% of household labour income (employees and business income) comes from self-employment sources.

Measuring Expenditure and Income

Measuring Expenditure

- *Food*
- *Non-Durables Basket*: Clothing, Food, Utilities.

Measuring Income Two measures of income that are comparable across survey and the register:

- *Labour Income*: WAS, income from self-employment: net profit, self-employment income for partnerships, shareholder salary, withholding payments.
- *Total Comparable Income*: Labour Income, Rental Income, Pensions, Other income (ACC, taxable benefits, student allowance and Paid Parental Leave). IDI Tax Data: Do not observe most investment income.

The Method

Empirically, ...

$$\ln C_{ij} = \beta_j \ln Y_i^R + \gamma_j SE_i + \Theta_i' X_i + \Xi_{ij}, \tag{8}$$

- Expanding on the covariates X_i: Demographics of the household (number of children, marital status, region, age and sex of HRP); Wealth (Survey: type of tenure, type of dwelling, number of rooms and of stories, area of the household as parted in the AS; Register: Variability of the income flow (measure of income risk and its growth.)
- Two-stages least squares to correct for the use of reported vs. permanent income in (8).

Income-Gap Estimates

Table: Estimation of the Income-Gap.

		(1): Register	(2): Survey
Panel A: Self-Employment: Opportunity			
Expenditure	**Income**	**Income-gap**	
Food	Labour	0.200***	0.114*
		(0.057)	(0.063)
Food	Comparable	0.193***	0.120*
		(0.048)	(0.062)
Non-Durables	Labour	0.204***	0.119**
		(0.047)	(0.051)
Non-Durables	Comparable	0.196***	0.124**
		(0.040)	(0.050)
Panel B: Self-Employment: 25% Rule			
Expenditure	**Income**	**Income-gap**	
Food	Labour	0.216***	0.107
		(0.066)	(0.075)
Food	Comparable	0.206***	0.111
		(0.055)	(0.073)
Non-Durables	Labour	0.254***	0.153***
		(0.053)	(0.059)
Non-Durables	Comparable	0.239***	0.158***
		(0.045)	(0.057)

- Survey estimates around 50-66% of register estimates ...

Validation of Survey Income

- Validation studies of income are scarce and focus mainly on employees' income. Self-employed are usually excluded from analysis. (Bound and Krueger, 1999; Bound *et al.*, 1994; Kreiner *et al.*, 2015).
- Previous validation studies seek a 'true' income measure for validation
- But we are interested in how well *underreported income* (in the tax register) is captured by the survey
- Therefore, in our case, register data is validated data (the 'gold standard') and survey reports measure this with error
- Measurement error can then be defined as: $u_i = Y_i^{Survey} - Y_i^{Register}$.

Validation of Survey Income

In our context, consider the 'true' Engle curve relationship in (1):

$$E_i^S = \beta Y_i^R + \varepsilon_i \qquad (1)$$

where E_i = reported expenditure by individual i, Y_i = i's income; 'S' and 'R' superscripts refer to Survey and Register sources respectively, and ε_i is a random error term. Both incomes and expenditures are measured in natural logarithms.

However, where there is measurement error in observed survey incomes, then:

$$Y_i^S = Y_i^R + u_i \qquad (2)$$

Estimating (1) using only survey data gives:

$$\begin{aligned} E_i^S &= \beta\left(Y_i^S - u_i\right) + \varepsilon_i \\ &= \beta Y_i^S + (\varepsilon_i - \beta u_i) \end{aligned} \qquad (3)$$

[E_i^S also measured with error but this 'only' reduces efficiency of estimate]

Note: 'R' = Register, not 'Reported'

Validation of Survey Income

In our context, consider the 'true' Engle curve relationship in (1):

$$E_i^S = \beta Y_i^R + \varepsilon_i \qquad (1)$$

where E_i = reported expenditure by individual i; Y_i = i's income; 'S' and 'R' superscripts refer to survey and register sources respectively, and ε_i is a random error term. Both incomes and expenditures are measured in natural logarithms.

However, where there is measurement error (and mean error $\neq 0$) in observed survey incomes (e.g. for self-employed), let:

$$Y_i^S = Y_i^R + u_i = Y_i^R + \bar{u} + v_i \qquad (2)$$

where $v_i = (u_i - \bar{u})$, $E(v_i) = 0$; $\bar{u} \neq 0$

Estimating (1) using only survey data gives:

$$\begin{aligned} E_i^S &= \beta\left(Y_i^S - \bar{u} - v_i\right) + \\ &= \beta Y_i^S - \beta\bar{u} + (\varepsilon_i - \beta v_i) \end{aligned} \qquad (3)$$

Therefore: (i) attenuation bias due to error term $(\varepsilon_i - \beta v_i)$; and
(ii) systematic downward bias of expenditures, E_i^S, by $\beta\bar{u}$, if $\bar{u} > 0$.

Note: 'R' = Register, not 'Reported'

Validation of Survey Income

For classical measurement error, where Y_i^R and u_i are uncorrelated, the bias can be summarised by:

$$\text{plim } \hat{\beta} = \gamma\beta \tag{4}$$

where: $\gamma = \frac{\sigma_{YR}^2}{\sigma_{YR}^2+\sigma_u^2}$ is the variance ratio or 'attenuation factor'.

Hence the bias can be given by:

$$-(1-\gamma)\beta = \frac{\sigma_u^2}{\sigma_{YR}^2+\sigma_u^2}\beta \tag{5}$$

However, if Y_i^R and u_i are correlated – as might be expected if survey income reports for higher (register) income taxpayers are subject to more, or less, reporting error – then it can be shown that (4) becomes:

$$\text{plim } \hat{\beta} = (1 - b_{uYs})\beta \tag{6}$$

where b_{uYs} is the estimated coefficient of a regression of u_i on Y_i^S

Errors in Register vs. Survey Incomes

Table: Moments of the distribution of the error by household

	Employees			Self-Employed		
	Register	Survey	Error	Register	Survey	Error
Panel A: Labour Income						
N	1914	1914	1914	663	663	663
Mean	10.984	10.973	-0.011	11.136	11.215	0.079
SD	0.814	0.811	0.453	0.867	0.769	0.507
P25	10.682	10.671	-0.086	10.813	10.881	-0.11
P50	11.104	11.095	-0.005	11.249	11.303	0.023
P75	11.467	11.47	0.06	11.627	11.655	0.247
Panel B: Comparable Income						
N	1914	1914	1914	663	663	663
Mean	11.064	11.025	-0.04	11.168	11.252	0.085
SD	0.627	0.736	0.426	0.799	0.745	0.542
P25	10.714	10.709	-0.096	10.828	10.922	-0.114
P50	11.127	11.123	-0.006	11.261	11.32	0.032
P75	11.472	11.478	0.074	11.639	11.681	0.254

- Measurement error is more severe for the self-employed than for the employed
- Unconditional difference in mean errors $\bar{u} \sim 0.09$

Conditional Errors in Register vs. Survey Incomes

Income Variable: Labour Income	(1) Register	(2) Survey	(3) Survey-Register
Age	0.079***	0.055***	-0.025***
	(0.011)	(0.011)	(0.008)
Age (Sq)	-0.001***	-0.001***	0.000***
	(0)	(0)	(0)
Female	-0.112***	-0.134***	-0.022
	(0.026)	(0.026)	(0.019)
Couple	0.833***	0.764***	-0.070***
	(0.031)	(0.03)	(0.022)
Number Children	-0.202***	-0.156***	0.045***
	(0.024)	(0.024)	(0.017)
.	.	.	.
.	.	.	.
.	.	.	.
Growth (Income)	0.168***	0.084***	-0.084***
	(0.031)	(0.031)	(0.022)
Volatility Income	-0.380***	-0.236***	0.144***
	(0.045)	(0.045)	(0.032)
Self-Employed	-0.090***	0.009	0.099***
	(0.031)	(0.031)	(0.022)
Constant	8.961***	9.471***	0.510***
	(0.253)	(0.253)	(0.181)
Observations	2,577	2,577	2,577
R-squared	0.400	0.366	0.039

- SE effect on income, conditional on: age, sex, single/couple, children, house characteristics (7), Accom. Supp. area (4), region (5), year, (past) average income growth/volatility
- Conditional mean error difference ~ 0.10 (log income higher in survey)

Measurement Error

	Attenuation biases	Reliability ratios
Labour	0.14	0.86
Comparable	0.28	0.72

Measurement error is non-classical

Table: Summary Statistics of Re

		Means (SD)					
Earnings Variables	(1) N	(2) Survey	(3) Register	(4) Error	(5) Variance Ratio (γ)	(6) b_{uY^S}	(7) b_{vY^R}
Panel A							
Labour Income	2577	11.036	11.024	0.013	0.242	0.139***	-0.187***
		(0.806)	(0.830)	(0.469)		(0.011)	(0.01)
Comparable: Total Income	2577	11.084	11.092	-0.008	0.318	0.280***	-0.127***
		(0.744)	(0.676)	(0.462)		(0.011)	(0.013)
Panel B: Omit outliers							
Labour Income	2532	11.041	11.031	0.010	0.152	0.063***	-0.119 ***
		(0.729)	(0.753)	(0.318)		(0.009)	(0.008)
Comparable: Total Income	2526	11.096	11.095	0.001	0.220	0.175***	-0.093***
		(0.647)	(0.616)	(0.327)		(0.009)	(0.008)

- Estimated biases (0.139, 0.280) are lower than the variance ratios (0.242, 0.318)
- Due to the negative correlation of the error with true income value - see column (7).
- Coefficients show the expected magnitude of the attenuation bias on income parameters from a regression where survey income is used as an independent variable as opposed to the register measure.

Measurement Error for Employed & Self-employed

Table: Summary Statistics of Reporting Errors

	(1)	Means (SD) (2)	(3)	(4)	(5)	(6)	(7)
	N	Survey	Register	Error	Var. Ratio (γ)	b_{uY^S}	b_{vY^R}
Panel A: Type of household							
Labour Income							
Self-Employment Income > 0	663	11.215	11.136	0.079	0.255	0.081***	-0.278***
		(0.769)	(0.867)	(0.507)		(0.025)	(0.02)
No Self-Employment Income	1914	10.974	10.984	-0.010	0.236	0.152***	-0.158***
		(0.810)	(0.813)	(0.453)		(0.012)	(0.012)
Comparable: Total Income							
Self-Employment Income > 0	663	11.252	11.168	0.085	0.315	0.190***	-0.295***
		(0.745)	(0.799)	(0.542)		(0.027)	(0.024)
No Self-Employment Income	1914	11.026	11.065	-0.040	0.317	0.305***	-0.041***
		(0.734)	(0.626)	(0.426)		(0.011)	(0.016)
Panel B: Omit Outliers							
Labour Income							
Self-Employment Income > 0	645	11.192				0.137***	-0.214***
		(0.707)				(0.023)	(0.021)
No Self-Employment Income	1890	10.990				0.034***	-0.093***
		(0.730)	(0.754)	(0.266)		(0.008)	(0.008)
Comparable: Total Income							
Self-Employment Income > 0	645	11.232			0.276	0.159***	-0.238***
		(0.670)				(0.025)	(0.023)
No Self-Employment Income	1881	11.050	11.075	-0.026	0.184	0.173***	-0.019*
		(0.633)	(0.582)	(0.276)		(0.009)	(0.011)

Reliability ratios ($1 - b_{uY^S}$):
self-employed = 0.86
employees = 0.97

Reliability ratios ≈ 0.84

Measurement Error & Attenuation Biases

Survey-Register (S-R) Parameter Differences

Dependent variable:		Food Expenditure			
Data source:		Register	Survey	Register	Survey
Income type:		Lab.	Lab.	Comp.	Comp.
A: Coefficients					
Income	$(\hat{\beta})$	0.460	0.443	0.545	0.443
S/R ratio		*0.963*		*0.813*	
SE Dummy	$(\hat{\gamma})$	0.103	0.0537	0.117	0.0565
S/R ratio		*0.521*		*0.483*	
B: Estimates of underreporting					
Multiplier	(k)	1.25	1.129	1.239	1.136
S/R ratio		*0.903*		*0.917*	
Income-gap	(κ)	0.200	0.114	0.193	0.120
S/R ratio		*0.570*		*0.622*	

$\hat{\beta}_S / \hat{\beta}_R$ →

$\hat{\gamma}_S / \hat{\gamma}_R$ →

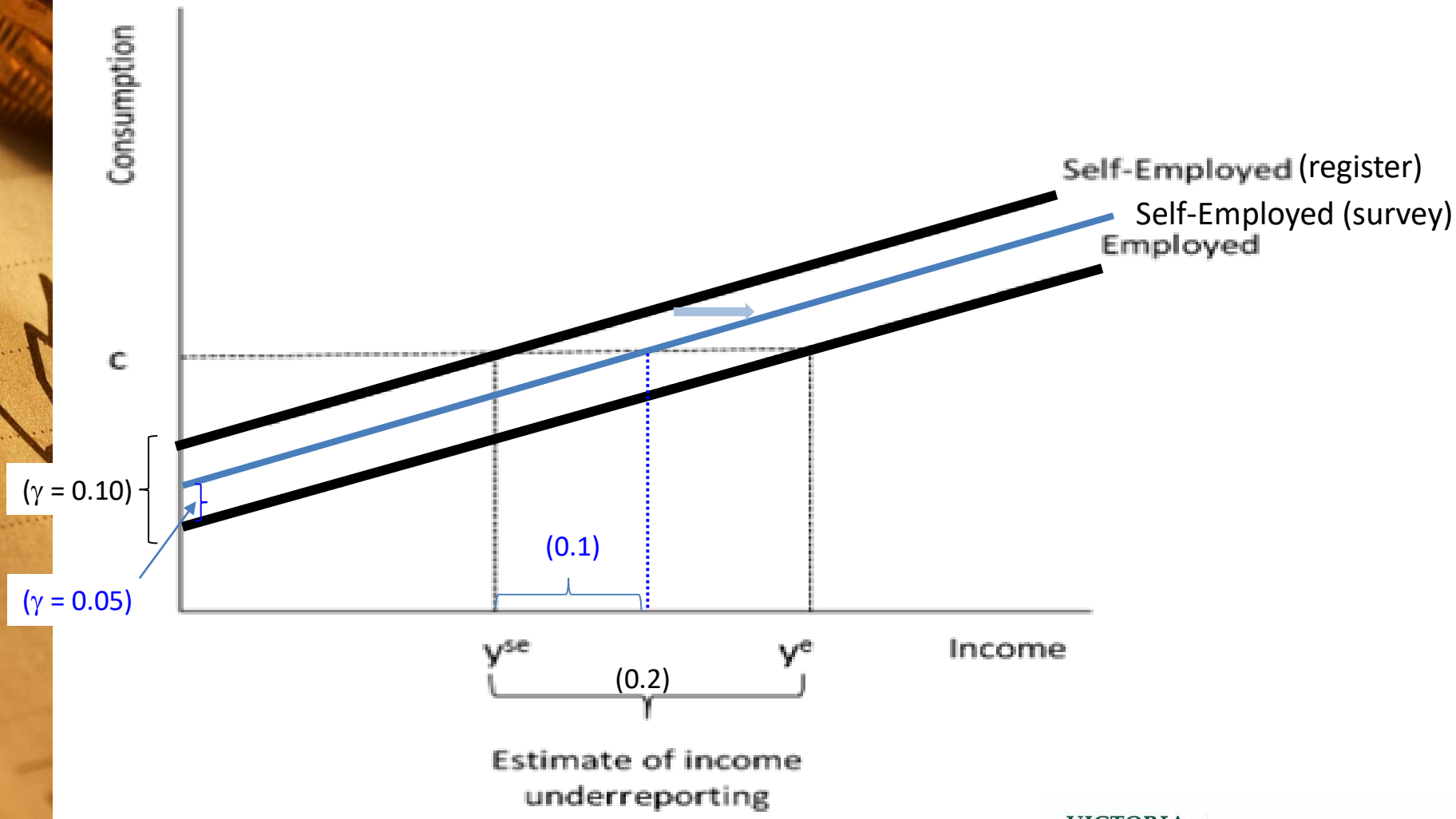
$k = exp(\hat{\gamma}/\hat{\beta})$ →

$\kappa = 1 - (1/k)$ →

Similar to:
Reliability ratios ($1 - b_{uY}s$)
labour = 0.86, 0.97
Comp. = 0.84

- How the two (income and SE dummy) attenuation biases interact to affect biases in income-gap estimates is not straightforward since the income-gap = $\kappa = 1 - (1/k)$, where the 'income scaling factor' $k = exp(\hat{\gamma}/\hat{\beta})$.

Engle Curves



Survey − based estimates: $E_i^S = \beta Y_i^S - \beta \bar{u} + (\varepsilon_i - \beta v_i)$

Conclusions

- Estimates of self-employment income gaps vary substantially depending on whether tax register, or survey-reported, income data are used in an Engle curve approach (around 19-21% versus 10-12%)
- Survey reports of income can be expected to be inaccurate as measures of reported taxable income (e.g. due to recall errors and deliberate underreporting by the self-employed)
- Data confirm survey-reported incomes are higher (on average) than register incomes for the self-employed, but very similar for employees.
- These generate substantial attenuation biases in parameter estimates for income in Engle curve regressions ~ up to 20%
- *Large* effects for SE dummy variable. Due to large average positive error (8-10%) for SE (log) incomes in survey data. Equivalent to ~ 4-5% error in (log) expenditures (with $\hat{\beta} \approx 0.5$)
- $\Rightarrow$ -0.05 (i.e. ≈ 50%) bias in SE parameter estimates, $\hat{\gamma}$, using survey data

Do Expenditure Survey Income Reports Mimic Tax Return Records?

The role of measurement error in survey versus register data

Ana Cabral & Norman Gemmell

Formerly Inland Revenue *Chair in Public Finance, VBS*

VICTORIA UNIVERSITY OF WELLINGTON
TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

Chair in Public Finance
Victoria Business School