A Quick and Dirty Estimate of Measurement Error in Household Survey Consumption Expenditures: Application to Indonesian Data.

Pritchett, Lant and Sumarto, Sudarno and Suryahadi, Asep

Harvard Kennedy School of Government, The SMERU Research Institute, INDONESIA, The SMERU Research Instittue, INDONESIA

17 August 2000

A Quick and Dirty Estimate of Measurement Error in Household Survey Consumption Expenditures: Application to Indonesian Data*

Lant Pritchett
The World Bank
Jakarta Stock Exchange Building, Tower 2, 12 th floor
Jl. Jenderal Sudirman Kav. 52-53
Jakarta 12190, Indonesia
Phone: 62-21-5299 3000, Fax: 62-21-5299 3111
Email: Lpritchett@worldbank.org

Sudarno Sumarto
Social Monitoring & Early Response Unit
Jl. Kusumaatmaja No. 69, Menteng
Jakarta 10310, Indonesia
Phone: 62-21-314 1224, Fax: 62-21-392 4659
Email: Ssumarto@smeru.or.id

Asep Suryahadi
Social Monitoring & Early Response Unit
Jl. Kusumaatmaja No. 69, Menteng
Jakarta 10310, Indonesia
Phone: 62-21-314 1224, Fax: 62-21-392 4659
Email: suryahadi@netscape.net

* We thank Wenefrida Dwi Widyanti for her research assistance. We are grateful to BPS, UNDP, and UNICEF for providing access to the data.

A Quick and Dirty Estimate of Measurement Error in Household Survey Consumption Expenditures: Application to Indonesian Data

Abstract

Measurement error is an enormous problem in empirical work. In some types of analysis, it is often ignored for various reasons. In some others, however, it cannot be ignored because it affects the results of analysis significantly. We use a simple procedure to estimate the extent of measurement error in household consumption expenditures data by comparing the coefficients obtained from OLS and instrumental variable regressions. We applied the procedure into school participation, durable goods expenditures, and clothing expenditures models using 10 different household survey data sets from Indonesia. The results indicate that measurement error accounts for less than 20 percent of total variance found in cross section household expenditures. In general, surveys which use detailed consumption questionnaire have smaller measurement error than those which use aggregated consumption questionnaire.

I. Introduction

Measurement error is an enormous problem in empirical economic analysis. Even carefully constructed survey data do not always conform exactly to the variables the analyst have in mind and the biases introduced by measurement error can be severe (Greene, 1990). In this study, we focus specifically on measurement error in household consumption expenditure data which are collected from household surveys. Household consumption expenditures are an important and widely used variable, particularly in household welfare analysis. For example, in most poverty analysis based on current consumption expenditure deficit (CCED), households are classified into poor and non-poor by the *measured* expenditures. However, measured expenditures are only a rough measure of actual expenditures and an even rougher measure of long run or "permanent" expenditures.

Observed expenditures (e) of household h at time t can be decomposed into the "permanent" (P) component of expenditures (that is the part which is time-varying but is expected at time t to persist), a "transitory" (T) component of expenditures (time varying and expected not to persist), and a measurement error term (v):

$$(1) \qquad e_h^t = e_h^{P,t} + e_h^{T,t} + v_h^t$$

If the three variances (σ^2) of the three components in equation (1) are uncorrelated with each other, then ratio of measurement error (or "noise") to total variance (n) is:

$$(2) \qquad n = \frac{\sigma_v^2}{\left(\sigma_P^2 + \sigma_T^2 + \sigma_v^2\right)}$$

In household welfare analysis, measurement error has several consequences. First, it tends to "flatten" poverty profiles by lowering the measured expenditure gap between groups, so the difference in poverty rates between households by educational status, landholding, etc., is understated. Second, as pointed out by Ravallion (1994), if the poverty line is below the mode, measurement error will increase measured poverty. Third, in inequality analysis, measurement error can cause higher measured than the actual state of inequality (Luttmer, 2000). Fourth, when using panel data for poverty transition analysis of the same households over time, measurement error can give a wrong impression that there are a lot of movements into and out of poverty (Baulch and Hoddinot, 2000; Deaton, 1997).

Fifth, in vulnerability analysis, if the standard deviation of "true" expenditures is less than the observed variability due to measurement error, this will imply that the level of vulnerability faced by households will be overstated.[1] The analysis of vulnerability using the estimates uncorrected for measurement error can be thought of as measuring the likelihood a household will have an episode of *appearing* to be in poverty, which could be the result of either that households are actually in poverty or simply that there is measurement error in their expenditures.

Using the first differences of expenditures exacerbates measurement error by reducing the role of permanent expenditures in the total expenditure variability. Assuming that the variances of permanent, temporary, and measurement error are constant across time, then the ratio of noise to total variance (noise plus signal) in the changes in expenditures is:

[1] Our definition of vulnerability is a probability, the risk a household will experience at least one episode of poverty in the near future (see Pritchett *et al*, 2000).

$$(3) \qquad n_{t,t-1} = \frac{2\sigma_v^2 - 2\sigma_{v_{t,t-1}}^2}{\left(2\sigma_P^2 - 2\sigma_{P_{t,t-1}}^2\right) + \left(2\sigma_T^2 - 2\sigma_{T_{t,t-1}}^2\right) + \left(2\sigma_v^2 - 2\sigma_{v_{t,t-1}}^2\right)}$$

In the special case in which permanent income is time invariant and the innovations in temporary and measurement are uncorrelated with the previous period's innovation, then this implies that the measurement error problem in estimating changes is worse by eliminating the permanent component by first differencing:

$$(4) \qquad \frac{\sigma_v^2}{\sigma_T^2 + \sigma_v^2} > \frac{\sigma_v^2}{\sigma_P^2 + \sigma_T^2 + \sigma_v^2}$$

So, while we can estimate the latter expression from a cross section, to move to the former requires some estimate of the permanent versus transitory innovations and the persistence of "permanent" innovations to expenditures.

II. Estimating Measurement Error: A Quick and Dirty Approach

What to do about measurement error? The first imperative is to have some reasonable estimate of its magnitude. Here we propose one easy heuristic way in estimating the extent of measurement error in household consumption expenditure data. Simply estimate any equation in which expenditures are a right hand side variable using both ordinary least squares (OLS) and instrumental variable (IV) techniques. Since the expression for the attenuation bias in OLS estimates in a bivariate regression is:

$$(5) \qquad \text{plim}\, \beta_{OLS} = \beta_o \left(1 - \frac{\sigma_v^2}{\sigma_*^2}\right)$$

where "*" represents the total (noise plus signal) variances, we just need to reverse this equation. If there exists an instrumental variables estimate that is consistent (plim $\beta_{IV} = \beta_o$), then one minus the ratio of the OLS to the IV estimate is a consistent estimate of the noise to total variance ratio:[2]

$$(6) \qquad \text{plim}\left(1 - \frac{\beta_{OLS}}{\beta_{IV}}\right) = \frac{\sigma_v^2}{\sigma_*^2}$$

A) Instruments

The key to estimating equation (6) empirically is to have good instruments for household expenditures. Fortunately, most household surveys which have household expenditures as a variable typically also contain a number of variables that are plausible instruments for expenditures. Examples of such variables are education level of household head, asset ownership, and housing conditions. The availability of such instruments makes it is possible to estimate equation (6) empirically, as implemented in the next subsection.

B) Empirical applications

Probably the simplest model with expenditures as a right hand side variable is an Engel curve, where the food share in household consumption expenditures is a function of the log of expenditures. However, since the right hand side is a non-linear transformation (i.e. natural log) of a variable (total expenditures) that is in the denominator of the left hand side, the

[2] This procedure is similar to that used in Luttmer (2000).

model does not satisfy the classical measurement error assumption. Hence, in this case one would have to apply the more advanced technique for non-linear measurement error (a la Hausman, Newey, and Powell (1995)).

As examples of empirical application of equation (6), we define a generic model:

$$(7) \qquad y_k = \alpha + \beta \cdot e + \gamma \cdot X + v$$

where y_k is the dependent variable, e is the log of expenditures, X is a vector of variables presumably affecting the dependent variable, v is a white noise, and k = 1, 2, 3. For the set of dependent variables, y_1 is the proportion of household members who are enrolled in schools, y_2 is the log of household expenditures on durable goods, and y_3 is the log of household expenditures on clothing. The X variables are education level of household head, age of household head, household size, sector of occupation of household head, urban-rural dummy variable, and district dummy variables.[3]

In these applications, the models are estimated using 10 different data sets from Indonesia.[4] All the surveys are carried out by the same organization, i.e. Statistics Indonesia (BPS). These surveys use either (i) a detailed Susenas consumption module questionnaire that contains 339 goods with recall period of one week for food and one month or one year for non-food, or (ii) the aggregated Susenas core questionnaire that contains only 23 goods with the same recall period as the detailed questionnaire.[5]

---

[3] The last two variables are included to take into account regional variability of school, durable goods, and clothing supllies.
[4] The descriptions of these data sets are available in Suryahadi *et al* (2000).
[5] Susenas is the National Socio-Economic Survey.

The results of estimations are presented in Table 1 for school participation, Table 2 for durable goods expenditures, and Table 3 for clothing expenditures. Since we are only interested in estimating measurement error in household expenditure data, only the coefficient of expenditures variable is presented for each estimation.

C) Comparing the results across models

The results for school participation model in Table 1 indicate that the estimates of ratio of measurement error to total variance range between 22 and 68 percent. Meanwhile, the magnitude of measurement error estimates from durable goods expenditures and clothing expenditures models in Tables 2 and 3 give a range of 1 to 33 percent and 3 to 19 percent respectively.

The school participation model, in which most of measurement error estimates are more than 40 percent, gives the highest estimates of the magnitude of measurement error in household expenditure data. The clothing expenditures model, in which most of measurement error estimates are less than 15 percent and certainly there is no estimate of more than 20 percent, gives the lowest estimates of the magnitude of measurement error. Meanwhile, the durable good expenditures model, in which most of measurement error estimates range between 9 and 33 percent, is in between.

This pattern of magnitudes of measurement error estimates can be traced back to the nature of each model. School participation is a relatively stable decision of a household. It does not vary very much with income. Once a child is enrolled in school, he or she will be most likely to stayed enrolled, at least within a given year, even though the household may experience declining fortunes during the course of the year. Hence, school participation is more likely to be affected by permanent income or expenditures only, but not by the transitory

ones. This means that the estimates in Table 1 are not a measure of measurement error only, but also include transitory expenditures.

Clothing expenditures, on the other hand, are affected by both permanent and transitory expenditures. When a household fortunes improve, they will likely buy more and better clothing. On the other hand, when the household fortunes go down, clothing is one of the first expenditure item that will be cut from the household budget. This means that the estimates in Table 3 are most likely to represent the net estimates of the magnitude of measurement error in household expenditure data.

Meanwhile, durable goods expenditures are also affected by transitory expenditures, but probably not as much as clothing expenditures. The reason is that durable goods in general are more expensive than clothing. Therefore, a slight increase in household fortunes may not induce the household to buy more or better durable goods. However, such a slight increase in household fortunes may be enough to induce the household to buy more or better clothing. This implies that the estimates in Table 2 contain both measurement error and some transitory expenditures, even though not as much as the estimates from school participation model.

D) Comparing the results across surveys

The results in Tables 2 and 3 indicate that the estimated ratio of measurement error to total variance in data generated from a detailed consumption questionnaire in general tends to be smaller than in data generated from an aggregated questionnaire. This implies that using a more detailed questionnaire helps reducing measurement error in household total expenditure data. This is probably because such a detailed questionnaire helps the respondents in recalling the actual household expenditures better than a more aggregated questionnaire.

Such an indication, however, do not appear in the results in Table 1. This means that if the focus of the survey is on permanent expenditures only, then using a detailed questionnaire does not give an advantage over using an aggregated questionnaire.

III. Concluding Remarks

In some types of analysis, netting out measurement error from total variance is needed because it affects the results significantly. We use a simple procedure to do so by utilizing regression coefficients obtained from the commonly used ordinary least squares and instrumental variable techniques. One minus the ratio of these coefficients is a measure of the proportion of measurement error from the total variance.

We apply the procedure to school participation, durable goods expenditures, and clothing expenditures models using 10 different household survey data sets from Indonesia. We find that the estimates of measurement error, net of transitory expenditures, are less than 20 percent of total variance in expenditures across households. In general, surveys which use a detailed consumption questionnaire have smaller measurement error than those which use an aggregated consumption questionnaire. However, the aggregated questionnaire is sufficient to capture permanent expenditures.

We would propose a similar analysis in all uses of consumption expenditures as a quick check on the reliability of the data. As feasible instruments are typically available, this procedure is quick, easy, and gives some sense of realism about the validity of results based on expenditures data.

References

Baulch, Bob and John Hoddinot (2000), 'Economic Mobility and Poverty Dynamics in Developing Countries,' *Journal of Development Studies*, forthcoming.

Deaton, Angus (1997), *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Johns Hopkins University Press, Baltimore, MD.

Greene, William H. (1990), *Econometric Analysis*, Macmillan, New York.

Hausman, J.A., W.K. Newey, and J.L. Powell (1995), 'Nonlinear Errors in Variables: Estimation of Some Engel Curves,' *Journal of Econometrics*, Vol. 65, pp. 205-34.

Luttmer, Erzo F.P. (2000), *Inequality and Poverty Dynamics in Transition Economies: Disentangling Real Events from Noisy Data*, World Bank, Washington, D.C., mimeo.

Pritchett, Lant, Asep Suryahadi, and Sudarno Sumarto (2000), *Quantifying Vulnerability to Poverty: A Proposed Measure, with Application to Indonesia*, SMERU Working Paper, May, Social Monitoring & Early Response Unit, Jakarta.

Suryahadi, Asep, Sudarno Sumarto, Yusuf Suharso, and Lant Pritchett (2000), *The Evolution of Poverty during the Crisis in Indonesia, 1996 to 1999 (Using Full Susenas Sample)*, SMERU Working Paper, March, Social Monitoring & Early Response Unit, Jakarta.

Table 1: Estimates of Measurement Error in Expenditures from a School Participation Model Using Various Indonesian Data Sets

Data Set	Estimation Technique	Number of Observation	R^2	Expenditures Coefficient		Estimate of Noise to Total Variance Ratio
				Coefficient	t value	
A. Detailed Susenas Consumption Module Questionnaire:						
Susenas Consumption Module, February 1996	OLS	35,161	0.1031	0.02238	11.95	0.5933
	IV	35,161	0.0952	0.05502	13.80	
Susenas Consumption Module, February 1999	OLS	28,387	0.1232	0.03140	13.04	0.2155
	IV	28,387	0.1228	0.04002	8.15	
Mini Susenas, December 1998	OLS	5,101	0.1453	0.02512	4.93	0.4216
	IV	5,101	0.1430	0.04343	5.13	
Mini Susenas, August 1999	OLS	4,439	0.1533	0.02008	3.92	0.5320
	IV	4,439	0.1493	0.04291	5.07	
B. Aggregated Susenas Core Questionnaire:						
Susenas Core, February 1996	OLS	82,650	0.1133	0.02674	20.40	0.5634
	IV	82,650	0.1059	0.06124	18.50	
Susenas Core, February 1999	OLS	68,431	0.1045	0.02547	15.84	0.6754
	IV	68,431	0.0902	0.07845	19.18	
100 Village Survey, May 1997	OLS	6,122	0.0783	0.03463	7.50	0.4101
	IV	6,122	0.0742	0.05870	6.58	
100 Village Survey, August 1998	OLS	6,578	0.0771	0.02596	6.13	0.4912
	IV	6,578	0.0722	0.05102	5.86	
100 Village Survey, December 1998	OLS	6,618	0.0792	0.02283	5.04	0.4400
	IV	6,618	0.0770	0.04076	4.12	
100 Village Survey, May 1999	OLS	6,538	0.0780	0.02513	5.43	0.3591
	IV	6,538	0.0767	0.03921	4.12	

Notes:
- The Susenas Core samples are excluding those in Susenas Consumption Module samples.
- The instruments used for expenditures are education, gender, housing conditions, and asset ownership variables, except for Susenas 1996 where asset ownership information is not available.
- Other independent variables are education level of household head, age of household head, household size, sector of occupation of household head, urban-rural dummy variable, and district dummy variables.

Table 2: Estimates of Measurement Error in Expenditures from a Durable Goods Expenditures Model Using Various Indonesian Data Sets

Data Set	Estimation Technique	Number of Observation	R^2	Expenditures Coefficient		Estimate of Noise to Total Variance Ratio
				Coefficient	t value	
A. Detailed Susenas Consumption Module Questionnaire:						
Susenas Consumption Module, February 1996	OLS	47,066	0.4398	2.10096	138.77	-
	IV	47,066	0.4329	1.73791	53.52	
Susenas Consumption Module, February 1999	OLS	34,448	0.3612	1.90798	100.44	0.1326
	IV	34,488	0.3568	2.19977	55.38	
Mini Susenas, December 1998	OLS	6,510	0.4290	1.86627	45.40	0.0149
	IV	6,510	0.4290	1.89449	27.86	
Mini Susenas, August 1999	OLS	5,452	0.4391	1.87406	43.11	0.0518
	IV	5,452	0.4385	1.97654	26.64	
B. Aggregated Susenas Core Questionnaire:						
Susenas Core, February 1996	OLS	111,355	0.3248	1.60881	172.29	0.0920
	IV	111,355	0.3229	1.77179	71.20	
Susenas Core, February 1999	OLS	84,482	0.2918	1.48291	129.98	0.2445
	IV	84,482	0.2769	1.96276	66.50	
100 Village Survey, May 1997	OLS	8,490	0.3496	1.80471	52.66	0.1478
	IV	8,490	0.3432	2.11767	29.97	
100 Village Survey, August 1998	OLS	7,133	0.3045	1.59609	42.85	0.3083
	IV	7,133	0.2689	2.30753	27.60	
100 Village Survey, December 1998	OLS	7,019	0.3267	1.77973	46.44	0.3271
	IV	7,019	0.2777	2.64482	28.26	
100 Village Survey, May 1999	OLS	7,075	0.3261	1.79064	45.65	0.3047
	IV	7,075	0.2879	2.57533	28.93	

Notes:
- The Susenas Core samples are excluding those in Susenas Consumption Module samples.
- The instruments used for expenditures are education, gender, housing conditions, and asset ownership variables, except for Susenas 1996 where asset ownership information is not available.
- Other independent variables are education level of household head, age of household head, household size, sector of occupation of household head, urban-rural dummy variable, and district dummy variables.

Table 3: Estimates of Measurement Error in Expenditures from a Clothing Expenditures Model Using Various Indonesian Data Sets

Data Set	Estimation Technique	Number of Observation	R^2	Expenditures Coefficient		Estimate of Noise to Total Variance Ratio
				Coefficient	t value	
A. Detailed Susenas Consumption Module Questionnaire:						
Susenas Consumption Module, February 1996	OLS	60,374	0.5160	0.94059	169.01	-
	IV	60,374	0.5155	0.89491	74.75	
Susenas Consumption Module, February 1999	OLS	51,031	0.4974	1.03944	149.56	0.0888
	IV	51,031	0.4953	1.14069	77.96	
Mini Susenas, December 1998	OLS	9,445	0.5961	0.93984	62.68	0.0349
	IV	9,445	0.5959	0.97380	37.58	
Mini Susenas, August 1999	OLS	7,776	0.6072	0.90619	60.70	0.1042
	IV	7,776	0.6046	1.01163	39.62	
B. Aggregated Susenas Core Questionnaire:						
Susenas Core, February 1996	OLS	144,671	0.4512	0.95562	206.40	0.0366
	IV	144,671	0.4510	0.99191	81.64	
Susenas Core, February 1999	OLS	122,775	0.4128	1.03979	187.83	0.1369
	IV	122,775	0.4086	1.20477	85.97	
100 Village Survey, May 1997	OLS	11,912	0.3670	0.98206	58.73	0.1865
	IV	11,912	0.3573	1.20724	34.42	
100 Village Survey, August 1998	OLS	11,903	0.3598	0.96050	55.78	0.0609
	IV	11,903	0.3591	1.02278	27.27	
100 Village Survey, December 1998	OLS	11,916	0.3741	0.93597	54.02	0.1716
	IV	11,916	0.3676	1.12981	26.71	
100 Village Survey, May 1999	OLS	11,939	0.4103	0.92238	54.36	0.1490
	IV	11,939	0.4058	1.08384	28.50	

Notes:
- The Susenas Core samples are excluding those in Susenas Consumption Module samples.
- The instruments used for expenditures are education, gender, housing conditions, and asset ownership variables, except for Susenas 1996 where asset ownership information is not available.
- Other independent variables are education level of household head, age of household head, household size, sector of occupation of household head, urban-rural dummy variable, and district dummy variables.