COWLES FOUNDATION DISCUSSION PAPER NO. 29

Note: Cowles Foundation Discussion Papers are preliminary materials circulated privately to stimulate private discussion and critical comment. References in publications to Discussion Papers (other than mere achnowledgment by a writer that he has access to such unpublished material) should be cleared with the author to protect the tentative character of these papers.

A Note on Least Squares Bias in Household Expenditure Analysis

Robert Summers

May 2, 1957

A Note on Least Squares Bias in Household Expenditure Analysis

1. The Method of Least Squares is a classical approach to the problem of estimating the coefficients of a regression equation. If, however, the equation is a member of a set of simultaneous equations, Least Squares applied equation by equation may lead to biased estimates. Attention has been devoted to such biases arising in models describing the aggregate economy and in models describing individual industries. This note examines a bias arising from a recent unjustified use of Least Squares in an analysis of household behavior.

2. A British survey of 22,000 working-class households in 1937-1939 produced a wealth of information about household expenditures of all sorts, classified by household composition and a number of other variables. Unfortunately, the households were not asked for their incomes. In a recent Engel curve analysis of the data of the survey by S.J. Prais and H.S. Houthakker,* total expenditure for each household was used as a proxy for

---

* S. J. Prais and H.S. Houthakker, The Analysis of Family Budgets, (1955) Also see A.M. Henderson, "The Cost of a Family," Review of Economic Studies, Vol. 17, (1949-1950) pp. 127-148.

---

income in an attempt to remedy this deficiency. Regression coefficients were estimated by Least Squares for the relationships between particular kinds of expenditure and total expenditure. Then these coefficients were adjusted so that they would be relevant to the relationship between the

particular kinds of expenditure and income. The adjustments were based upon estimates of coefficients of the relationship between total expenditure and income that were derived from an independent set of data.

3. In estimating the parameters of Engel curves, the assumption is made universally that a household's expenditures on the goods and services in its budget depend upon its income (and possibly other variables), but that its income does not depend upon its expenditures. The basic assumption of the Least Squares theorem is complied with only if there is no "feed-back" between the various expenditures and income. Perhaps there are some households which decide upon the amount of effort they will expend in order to earn income on the basis of how much they wish to spend. In the absence of empirical evidence that such households are common, however, it is not unreasonable to assume for statistical purposes that income does not depend upon expenditures.

It is clear, however, that the assumption that there is no "feed-back" between expenditures on individual items and total expenditure is not acceptable. The biases resulting from estimating regression coefficients by Least Squares when the independent variable is not truly independent of the stochastic term in the regression equation -- which is what the presence of "feed-back" amounts to -- were minimized by Prais and Houthakker. They state that "So long as the item of expenditure is a small proportion of the budget it is not to be expected that serious biases will result ..."*

---

* Prais and Houthakker, op., cit. p. 63

4. In order to examine the bias, the simultaneous equation character of the model underlying the analysis must be specified. A simple linear version of the model consisting of $m+2$ equations is given by (1).

$$(1)\qquad \left.\begin{aligned} e_{ik} &= \alpha_{1i} Y_k + \alpha_{oi} + v_{ik} \qquad i = 1,\ldots,m \\ E_k &= \alpha_1 Y_k + \alpha_o + u_k \\ E_k &= \sum_i e_{ik} \end{aligned}\right\} \quad k = 1,\ldots,n$$

E_k is total expenditure of the k'th household; e_{ik} is expenditure of the k'th household on the i'th commodity; Y_k is any specific function of income of the k'th household; the (v_i)'s are random variables which are independent of income, which have zero means, and which have a distribution, $f(v_1,\ldots,v_m)$, with a covariance matrix $[\sigma_{ij}]$; and u is a random variable, independent of income, with a zero mean and variance σ_u^2.

The first m equations of (1) are the ones of substantive interest in an investigation of expenditure patterns for particular commodities. The $(m+1)$st equation is included only because it enters into the statistical estimation process used by Prais and Houthakker. The last equation defines total expenditure. From this definition it follows that the $(m+1)$st equation is a linear combination of the first m equations so $u_k = \sum_i v_{ik}$, $\alpha_1 = \sum_i \alpha_{1i}$, and $\alpha_o = \sum_i \alpha_{oi}$. If income data were available, the parameters of the first m equations could be estimated directly by Least Squares applied to each equation individually. For brevity, this

procedure will be called Single Equation Least Squares.

Before computing expected values to investigate the bias resulting from estimating the regression coefficients using an "independent" variable which is correlated with the stochastic term, it is worthwhile recasting (1) into a different form in order to show from a simultaneous equations estimation point of view what gives rise to the bias. We obtain (2) from (1) by replacing Y_k in the first m equations by its value in terms of E_k as determined by the $(m+1)$st equation

$$\left.\begin{aligned} e_{ik} - \beta_{1i}E_k - \beta_{oi} &= w_{ik} \qquad i = 1,\ldots,m \\ E_k - \beta_1 Y_k - \beta_o &= \\ \sum_i e_{ik} - E_k &= 0 \end{aligned}\right\} \quad k = 1,\ldots,n$$

where $\beta_{1i} = \dfrac{\alpha_{1i}}{\alpha_1}$, $\beta_{oi} = \alpha_{oi} - \dfrac{\alpha_{1i}}{\alpha_1}\alpha_o$, $\beta_1 = \alpha_1$, $\beta_o = \alpha_o$, $w_{ik} = v_{ik} - \dfrac{\alpha_{1i}}{\alpha_1}u_k$

and $z_k = u_k$

It is easy to verify that the parameters of (2) are all just-identifiable. Estimates of the parameters of (2) would provide a basis for estimating the parameters of (1), the numbers that are really of interest. (Notice that (1) is a system of equations already in reduced form. In this topsy-turvy way of looking at the problem, reduced form coefficients which are unobtainable because of a data deficiency are wanted and structural coefficients are estimated in order to get them. Prais and Houthakker estimated the (α_{1i})'s from Single Equation Least Squares estimates of the corresponding (β_{1i})'s

and from α_1. Since the (β_{1i})'s should have been estimated by a method that took into account the whole simultaneous equation system, it is clear that the estimates are biased. Consequently, the estimates of the (α_{1i})'s will also be biased. The point can be made in a slightly different way. Since (2) is a recursive system, the parameters of each equation can be estimated without bias by Single Equation Least Squares, provided that in the first m equations a "corrected" version of E obtained from the $(m+1)$st equation is used rather than the observed value. If the correction is not made the estimates will be biased. Prais and Houthakker did not make the correction -- for the very good reason that they did not have the income of each household -- so their estimates are biased.

5. Now the size of the bias will be investigated. The Single Equation Least Squares estimates of a parameter, say β, will be designated $\tilde{\beta}$. The Prais-Houthakker estimate of α_{1i}, designated $\bar{\alpha}_{1i}$, is given by (3).

$$(3) \qquad \bar{\alpha}_{1i} = \tilde{\beta}_1 \cdot \tilde{\beta}_{1i}$$

Since β_1 was estimated from an independent set of data, and it can be assumed that there is no "feed-back" between total expenditure and income,

$$(4) \qquad \mathcal{E}\bar{\alpha}_{1i} = \mathcal{E}\tilde{\beta}_1 \cdot \mathcal{E}\tilde{\beta}_{1i} = \alpha_1 \mathcal{E}\tilde{\beta}_{1i},$$

where $\mathcal{E}$ stands for "expected value." The expected value of $\tilde{\beta}_{1i}$ is

easily evaluated, since

$$(5) \quad \tilde{\beta}_{1i} = \frac{\sum_k (e_{ik} - \bar{e}_i)(E_k - \bar{E})}{\sum_k (E_k - \bar{E})^2}$$

and

$$(6) \quad \tilde{\beta}_{1i} = \frac{\sum_k \left[\alpha_{1i}(Y_k - \bar{Y}) + v_{ik}\right]\left[\alpha_1(Y_k - \bar{Y}) + u_k\right]}{\sum_k \left[\alpha_1(Y_k - \bar{Y}) + u_k\right]^2} .$$

Since v_i and u are independent of income,

$$(7) \quad \mathcal{E}\tilde{\beta}_{1i} = \frac{\alpha_{1i}\alpha_1\sigma_Y^2 + \sigma_{v_i u}}{\alpha_1^2\sigma_Y^2 + \sigma_u^2},$$

where $\sigma_{v_i u}$ is the covariance of v_i and u. Therefore,

$$(8) \quad \mathcal{E}\bar{\alpha}_{1i} = \frac{\alpha_{1i}\alpha_1^2\sigma_Y^2 + \alpha_1\sigma_{v_i u}}{\alpha_1^2\sigma_Y^2 + \sigma_u^2} .$$

In this model $u_k = \sum_i v_{ik}$, so $\sigma_{v_i u} = \Sigma\sigma_{\ell i}$, where $\sigma_{\ell i}$ is the covariance of v_ℓ and v_i. Therefore,

$$(9) \quad \mathcal{E}\bar{\alpha}_{1i} = \frac{\alpha_{1i}\alpha_1^2\sigma_Y^2 + \alpha_1 \sum_\ell \sigma_{\ell i}}{\alpha_1^2\sigma_Y^2 + \sigma_u^2}$$

In the special case where the stochastic elements of the various expenditure equations are mutually independent, (9) simplifies to

$$(10) \quad \mathcal{E}\bar{\alpha}_{1i} = \frac{\alpha_{1i}\alpha_1^2\sigma_Y^2 + \alpha_1\sigma_i^2}{\alpha_1^2\sigma_Y^2 + \sigma_u^2},$$

The ratio of the first term in the numerator to the first term in the denominator on the right side of (10) is equal to α_{1i}. Then the bias stems from the fact that the expected value of $\bar{\alpha}_{1i}$ is equal to a term of the form $\frac{c\,\alpha_{1i}}{c}$ modified by the addition of an expression in

the numerator and another one in the denominator. The expected value would be equal to α_{1i} only if the ratio of these two expressions were equal to α_{1i} ; that is, if

$$(11) \quad \frac{\alpha_1 \sigma_i^2}{\sigma_u^2} = \alpha_{1i}$$

or if

$$(12) \quad \frac{\sigma_i^2}{\sigma_u^2} = \frac{\alpha_{1i}}{\alpha_1} .$$

Thus the Prais-Houthakker estimate of α_{1i} is biased if the ratio of the variance of residuals around the (e_i,Y) regression curve to the variance of residuals around the (E,Y) regression curve is not the same as the ratio of the slope coefficient of the (e_i,Y) regression curve to the slope coefficient of the (E,Y) regression curve. (It should be remembered that Y is any function of income rather than merely income itself.) More specifically, if $\frac{\sigma_i^2}{\sigma_u^2} > \frac{\alpha_{1i}}{\alpha_1}$, $\bar{\alpha}_{1i}$ will be biased upward; and if $\frac{\sigma_i^2}{\sigma_u^2} < \frac{\alpha_{1i}}{\alpha_1}$, $\bar{\alpha}_{1i}$ will be biased downward.

6. By writing (10) in a slightly different form, it will be possible to see if the Prais-Houthakker claim is justified that the bias is not serious if the item of expenditure is small relative to the whole budget.

$$(13) \quad \mathcal{E}\bar{\alpha}_{1i} = \alpha_{1i} + \frac{\alpha_1 \sigma_i^2 - \alpha_{1i} \sigma_u^2}{\alpha_1^2 \sigma_Y^2 + \sigma_u^2}$$

The second member of the right side of (13) is the size of the bias. If the relative bias in estimating α_{1i} is called B_i, we have

$$(14) \qquad B_i = \left[\frac{\alpha_1 \sigma_i^2 - \alpha_{1i} \sigma_u^2}{\alpha_1^2 \sigma_Y^2 + \sigma_u^2} \right] \Big/ \alpha_{1i} ,$$

from which it follows that

$$(15) \qquad B_i = \frac{\sigma_i^2 - \left(\frac{\alpha_{1i}}{\alpha_1} \right) \sigma_u^2}{\alpha_{1i} \alpha_1 \sigma_Y^2 + \left(\frac{\alpha_{1i}}{\alpha} \right) \sigma_u^2} .$$

The item of expenditure will be small relative to the whole budget if $\frac{\alpha_{1i}}{\alpha_1}$ and $\frac{\alpha_{oi}}{\alpha_o}$ are both small. By expressing B_i as in (15), it becomes clear that the relative bias is not merely a function of the relative importance of the item of expenditure.

7. The foregoing conclusion was reached on the assumption of a simple linear model where the only generality present was that the independent variable of (1) could be any specific function of income. As a consequence, in (2), the other way of writing the model, e_i is a linear function of total expenditure. Prais and Houthakker found, however, that in the British survey the best-fitting relationship between e_i and E seemed to be either semi-logarithmic or double-logarithmic, depending upon the item of expenditure. This implies that the basic model is composed of equations like those of (1').

$$(1') \qquad \left.\begin{array}{lll} e_{ik} = \alpha_{1i} Y_k + \alpha_{oi} + v_{ik} & & i = 1,\ldots,m_1 \\ \text{In } e_{ik} = \alpha_{1i} Y_k + \alpha_{oi} + v_{ik} & & i = m_1+1,\ldots,m \\ \text{In } E_k = \alpha_1 Y_k + \alpha_o + u_k & & \\ \sum_i e_{ik} = E_k & & \end{array}\right\} \quad k = 1,\ldots,n$$

The first m_1 equations of (1') concern the items of expenditure related to E in a semi-logarithmic way; the next $(m-m_1)$ equations concern the items related to E in a double-logarithmic way. In this model, however, u_k is not equal to $\sum_i v_{ik}$ and hence $\sigma_{v_i u}$ does not reduce simply to σ_i^2. It can easily be shown that now the expression for the relative bias for any of the m α_{1i}'s is given by (14') and (15').

$$(14') \qquad B_i = \left[\frac{\alpha_1 \sigma_{v_i u} - \alpha_{1i} \sigma_u^2}{\alpha_1^2 \sigma_Y^2 + \sigma_u^2} \right] \Big/ \alpha_{1i}$$

$$(15') \qquad B_i = \frac{\sigma_{v_i u} - (\frac{\alpha_{1i}}{\alpha_1}) \sigma_u^2}{\alpha_{1i} \alpha_1 \sigma_Y^2 + (\frac{\alpha_{1i}}{\alpha_1}) \sigma_u^2}$$

Though a simple expression for $\sigma_{v_i u}$ is not available, under the assumption that the stochastic terms in the various expenditure equations are mutually independent, it is clear that $\sigma_{v_i u} > 0$. The condition that the estimate of α_{1i} will be unbiased, given in (12'), is directly analogous with the condition for the linear model, but it has no simple interpretation.

$$(12') \qquad \frac{\sigma_{v_i u}}{\sigma_u^2} = \frac{\alpha_{1i}}{\alpha_1}$$

As in the case of the linear model, one must expect that the estimate of α_{1i} will be biased even if the item of expenditure is small relative

to the whole budget.

8. A remaining comment should be made. The variables Prais and Houthakker dealt with were deflated to allow for differences in household size and composition. In the simplest case, where the expenditure equations are as given in (1) with e_i, E, and Y all expressed in per-household-member terms, (14) and (15) still give the relative bias. Even when dealing with the linear equations, however, if one household-composition scale is used for deflating Y and E, but different ones are used for deflating each of the e_i's, no more can be said about the bias than that it is given by (14') and (15').

9. Conclusion. The estimating procedure used in a recent analysis of household behavior has been examined. In this analysis, Engel curve coefficients were estimated, but because of a data deficiency, the estimates were derived from two different sets of data. Of necessity, the Least Squares estimating method was used. It has been shown that this approach is equivalent to applying Least Squares equation by equation to a system of simultaneous equations. The resulting bias has been shown to be dependent upon more than just the relative size of the item of expenditure in the total budget.