the Advisory Committee to the Surgeon General of the Public Health Service, Washington, D.C., Government Printing Office, 1964.
30. U.S. Public Health Service: *The Health Consequences of Smoking,* Washington, D.C., U.S. Government Printing Office, 1967.
31. *Wall Street Journal:* September 18, 1963, 28.
32. Wilson, Ronald W.: "Comment on 'Review of Claim that Excess Morbidity and Disability can be Ascribed to Smoking'", *Journal of the American Statistical Association,* **68** (March 1973), 85–87.
33. Wright, I. S.; Chairman: "The Final Enovid Report," *Journal on New Drugs,* **3** (1963), 201.
34. Wright, I. S.; Chairman: *Final Report on Enovid,* submitted to the Commissioner of the Food and Drug Administration, September 12, 1963.

Some Notes on the Errors-in-Variables Model

POTLURI RAO*

In studying the properties of least squares estimates we implicitly assume that all the variables are measured without errors. When some, or all, of the variables are subject to errors, even though the estimated equation is a true relation, the regression estimates can be biased. In empirical research we often face variables with errors of some kind or other. Research in this area, generally known as the *errors-in-variables* model, is concentrated mainly on the theoretical properties of the estimates, such as the asymptotic bias, in a two-variable regression model. In a practical situation, however, a researcher is interested in assessing the direction and *approximate* extent of bias in given data in order to decide on whether to keep a variable in the equation. In this paper we provide analytical expressions to evaluate bias when all of the variables of a k-variable regression model are subject to error. These expressions are derived under general conditions, so that a researcher may tailor the expressions to suit the needs of any given situation.

Let the true relationship between the variables be

$$x^*_{k+1,t} = \alpha_1 x_{1t}^* + \alpha_2 x_{2t}^* + \cdots + \alpha_k x_{kt}^* + \epsilon_t, \qquad t = 1, 2, \ldots, T \quad (1)$$

where the variables with asterisks (*) are measured without errors, and the error terms (ϵ_t) follow the assumptions of the classical model.

To make the analysis as general as possible let us assume that all the variables are subject to errors

$$x_{it} = x_{it}^* + f_{it}, \qquad i = 1, \ldots, k+1$$

where x_i is the observed value and f_i is the error.

The true relation (1) may be rewritten in the observed values of the variables as

$$x_{k+1,t} = \alpha_1 x_{1t} + \alpha_2 x_{2t} + \cdots + \alpha_k x_{kt} + z_t + \epsilon_t, \quad (2)$$

where $z_t = f_{k+1,t} - \alpha_1 f_{1t} - \alpha_2 f_{2t} - \cdots - \alpha_k f_{kt}$.

The estimated regression equation from the observed values of the variables is

$$x_{k+1,t} = \alpha_1 x_{1t} + \alpha_2 x_{2t} + \cdots + \alpha_k x_{kt} + \epsilon_t'. \quad (3)$$

For analytical convenience we may interpret the estimated equation (3) as a misspecification of the true relation (2), with z as a left-out variable. We know that omission of a variable in the classical regression model results in biased estimates, and the expression for bias in estimate $\hat{\alpha}_1$, for example, is given in the Yule notation as[1]

$$B(\hat{\alpha}_1) = E(\hat{\alpha}_1) - \alpha_1 = b_{z1.23\ldots k}$$

where $b_{z1.23\ldots k}$ is computationally equivalent to the ordinary least squares estimate in the auxiliary regression equation:

$$z_t = b_{z1.23\ldots k} x_{1t} + b_{z2.13\ldots k} x_{2t} + \cdots + b_{zk.12\ldots k-1} x_{kt} + \epsilon_t.$$

In the matrix notation, $b_{z1.23\ldots k}$ is the first element in the **b** vector defined as

$$\mathbf{b} = \begin{bmatrix} \sum x_1^2 & \sum x_1x_2 & \cdots & \sum x_1x_k \\ \sum x_1x_2 & \sum x_2^2 & \cdots & \sum x_2x_k \\ \vdots & \vdots & \cdots & \vdots \\ \sum x_1x_k & \sum x_2x_k & \cdots & \sum x_k^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum x_1z \\ \sum x_2z \\ \vdots \\ \sum x_kz \end{bmatrix} \quad (4)$$

where $\sum x_ix_j$ stands for $\sum_{t=1}^{T} x_{it}x_{jt}$ and $\sum x_iz$ for $\sum_{t=1}^{T} x_{it}z_t$. Throughout this paper the summation sign ($\sum$) stands for $\sum_{t=1}^{T}$.

In order to simplify the algebra consider the following equality:

$$\begin{bmatrix} \sum x_1^2 & \sum x_1x_2 & \cdots & \sum x_1x_k \\ \sum x_1x_2 & \sum x_2^2 & \cdots & \sum x_2x_k \\ \vdots & \vdots & \cdots & \vdots \\ \sum x_1x_k & \sum x_2x_k & \cdots & \sum x_k^2 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} \dfrac{1}{\sum x^2_{1.23\ldots k}} & \dfrac{-b_{12.3\ldots k}}{\sum x^2_{1.23\ldots k}} & \cdots & \dfrac{-b_{1k.23\ldots k-1}}{\sum x^2_{1.23\ldots k}} \\ \dfrac{-b_{21.3\ldots k}}{\sum x^2_{2.13\ldots k}} & \dfrac{1}{\sum x^2_{2.13\ldots k}} & \cdots & \dfrac{-b_{2k.13\ldots k-1}}{\sum x^2_{2.13\ldots k}} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{-b_{k1.23\ldots k-1}}{\sum x^2_{k.12\ldots k-1}} & \dfrac{-b_{k2.13\ldots k-1}}{\sum x^2_{k.12\ldots k-1}} & \cdots & \dfrac{1}{\sum x^2_{k.12\ldots k-1}} \end{bmatrix} \quad (5)$$

* Dept. of Economics, Univ. of Washington, 301 Guthrie Hall, Seattle, Wash. 98195.

[1] See Rao [1].

This content downloaded from 128.235.251.160 on Thu, 1 Jan 2015 15:23:41 PM
All use subject to JSTOR Terms and Conditions

where $x_{1.23\ldots k,t}$ is defined in the Yule notation as the tth residual in the auxiliary regression equation with x_1 as a dependent variable and $(x_2, x_3, \ldots, x_k)$ as independent variables, and $\sum x^2_{1.23\ldots k} = \sum_{t=1}^{T} x^2_{1.23\ldots k,t}$. A proof of the equality (5) may be obtained by premultiplying and postmultiplying the original matrix by the presumed inverse and using the following properties of regression equations:[2]

$$\sum x_1 x_{1.23\ldots k} = \sum x^2_{1.23\ldots k}$$
$$\sum x_2 x_{1.23\ldots k} = 0$$
$$b_{12.3\ldots k} = \sum x_{1.3\ldots k} \cdot x_{2.3\ldots k} / \sum x^2_{2.3\ldots k}$$

By using equality (5) and noting that $\sum x^2_{1.23\ldots k} = \sum x_1^2(1 - R^2_{1.23\ldots k})$, where $R_{1.23\ldots k}$ is the multiple correlation coefficient between x_1 and $(x_2, x_3, \ldots, x_k)$, we may express the bias term in two different forms as

$$B(\hat{\alpha}_1) = \frac{\sum x_1 z - b_{12.3\ldots k}\sum x_2 z - \cdots - b_{1k.23\ldots k-1}\sum x_k z}{\sum x_1^2(1 - R^2_{1.23\ldots k})} \tag{6}$$

or

$$B(\hat{\alpha}_1) = \frac{1}{1 - R^2_{1.23\ldots k}} \cdot \frac{\sum x_1 z}{\sum x_1^2} - \frac{b_{21.3\ldots k}}{1 - R^2_{2.13\ldots k}} \cdot \frac{\sum x_2 z}{\sum x_2^2} - \cdots - \frac{b_{k1.23\ldots k-1}}{1 - R^2_{k.12\ldots k-1}} \cdot \frac{\sum x_k z}{\sum x_k^2}. \tag{7}$$

In these expressions for bias, the b's and R's are known, as they can be computed from the observed values of the variables. The only unknown quantities are the terms involving the variable z. In a practical situation, however, the researcher is interested in knowing the nature of bias if the errors were generated in a particular way. Let us consider a few practical situations to illustrate the point.

Consider a case where only one of the variables, say x_k, is subject to errors. Suppose we are interested in knowing the nature of the bias in $\hat{\alpha}_1$ if the errors in x_k were generated such that the error term f_k is uncorrelated with the true values of the variables x^*. In this case we want to know the bias if $\sum f_k x_j^*$ were equal to zero. We can easily evaluate the terms involving z under this assumption. Then expression (7) is convenient to interpret, as it can be readily reduced asymptotically to

$$B(\hat{\alpha}_1) = \frac{\alpha_k \cdot b_{k1.23\ldots k-1}}{1 - R^2_{k.123\ldots k-1}} \cdot \frac{\sum f_k^2}{\sum x_k^2}$$

Even though the variable corresponding to a regression coefficient x_1 is error-free, error in the other variables can introduce bias in its regression coefficient. In this example, bias in $\hat{\alpha}_1$ depends on the ratio, $\lambda_k = \sum f_k^2 / \sum x_k^2$, the proportion of variation in x_k due to the error term. The larger the contribution of the error term to the variance of x_k, the larger the bias in all the regression coefficients. Even though λ_k may be large, if $\alpha_k b_{k1.23\ldots k-1}$ is very small the extent of bias in $\hat{\alpha}_1$ can be negligible. The bias depends also on the multiple correlation between the variable x_k and all of the other independent variables in the regression. When this multiple correlation is large, even though the variance of errors may be small relative to the variance of x_k, the bias can be substantial.

When the researcher has prior knowledge of the sign of the parameter α_k, which is usually the case in many empirical works, the direction of bias is obvious from the sign of $b_{k1.23\ldots k-1}$, which can be computed from the available data. In order to know the extent of bias, however, we need to know the term $\alpha_k\lambda_k$. In many cases it may be possible to obtain an approximate value of the term $\alpha_k\lambda_k$ on the basis of extraneous information.

The expressions for bias given in (6) and (7) are general and may be adopted to any practical situation. To demonstrate the flexibility of these expressions let us consider the case where all the variables are subject to the same error. This kind of a problem is frequently found in log-linear models where all the variables are deflated by a "wrong" index in adjusting for trend.[3] In this case

$$f_i = f \text{ for all } i.$$

Suppose we are interested in knowing the bias if the error (f) were uncorrelated with the real values of all the variables; we can evaluate the terms involving the variable z under this assumption. In this case expression (6) is convenient to interpret as it reduces asymptotically to

$$B(\hat{\alpha}_1) = \lambda_1(1 - \alpha_1 - \alpha_2 - \cdots - \alpha_k) \times \frac{(1 - b_{12.3\ldots k} - \cdots - b_{1k.23\ldots k-1})}{1 - R^2_{1.23\ldots k}}.$$

In this case bias depends not only on λ_1 and $R_{1.23\ldots k}$, but also on the degree of the function being estimated $(\alpha_1 + \alpha_2 + \cdots + \alpha_k)$.

In empirical research we often worry about the consequences of errors in the variables. By using the expressions derived in this paper, we can perceive the nature of bias under different error term assumptions. One major advantage of this particular analysis is that the expressions for bias are derived in terms of the observed values of the variables, whereas the theoretical results in the literature are usually derived in terms of the real values of the variables. These results should provide some guidance in practical situations.

The author is grateful to Professors Zvi Griliches and Juan Zapata for their helpful comments.

REFERENCES

[1] Rao, P. (1971): "Some Notes on Misspecification in Multiple Regressions," *The American Statistician*, 25 (December), pp. 37–9.

[2] Rao, P., and R. L. Miller (1971): *Applied Econometrics*, Wadsworth Publishing Company. Belmont, Calif.

[3] Yule, G. U., and M. G. Kendall (1950): *An Introduction to the Theory of Statistics*, 14 ed., Charles Griffin & Co. London.

[2] See Yule and Kendall [3, pp. 285–8].

[3] For an example of this kind of problem in the context of a production function see Rao and Miller [2, p. 182].

This content downloaded from 128.235.251.160 on Thu, 1 Jan 2015 15:23:41 PM
All use subject to JSTOR Terms and Conditions