184

WP: 184

Working Paper

विद्याविनियोगाद्विकास:
IIM
AHMEDABAD

INDIAN INSTITUTE OF MANAGEMENT AHMEDABAD

CHOICE OF ESTIMATED ECONOMETRIC MODELS

by

P.N. Misra

W P No. 184
Nov.1977

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD

CHOICE OF ESTIMATED ECONOMETRIC MODELS

B y

P.N. Misra

Indian Institute of Management, Ahmedabad

1 Introduction

One of the problems faced by researchers in the field of quantitative economics relates to choice of appropriate functional forms from amongst many that can be estimated on the basis of available data on a given set of causal and effect variables. Most economic phenomena could alternatively, be stated alternatively as an effect variable y depending upon K causal variables, namely x_1, x_2, ... and x_k. A good understanding of economic reasoning both in theory and practice will help a lot to specify, define and quantify the above mentioned variables but it seldom comes to one's rescue while one battles to understand the mode of dependence between y and x variables. The only way that appears to be available at this stage is to collect adequate data on these variables and try to figure out the mode of dependence on the basis of the data at hand. Algebraically, the statment that y depends upon $x_1, \ldots, x_k$ can be expressed as

$$y = f(x_1, \ldots x_k, u) \qquad (1.1)$$

where f stands for 'function of' and u denotes error in the

function that persists in spite of best efforts to identify, define and quantify the variables associated with the phenomenon that one wants to explain. This function may be linear:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \tag{1.2}$$

exponential:

$$\log y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \tag{1.3}$$

double-log:

$$\log y = {}_0 + \beta_1 \log x_1 + \ldots + \beta_k \log x_k + u \tag{1.4}$$

or, of any other form. The β coefficients and error term u in the above forms are understood to be different in case of different relations. In general, Taylor expansion can transform any relation like (1) into a polynomial of appropriate degree. Such a polynomial of second degree can be written as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \alpha_1 x_1^2 + \ldots \alpha_k x_k^2 + \ldots + u \tag{1.5}$$

where terms of higher order of x's could be included if necessary.

Estimation of an economic relation, however, is only an intermediate step because often the estimated relation is to be employed to obtain forecast on the effect variable. Forecasts will be close to true value if mistakes are avoided in respect

of every step of the whole exercise including choice of right functional form. It is, therefore, better to examine this problem from the angle of getting good forecasts.

The problem has been discussed in several texts on econometrics and applied econometrics. One of these [2] provides a good account for the benefit of applied economists but the discussion is not conclusive. The results reported in this note will hopefully narrow down the ambiguity that exists on the subject.

2 An useful Statistic

In order to understand the nature of the statistic in relatively less algebraic term, we will start with a brief discussion of forecasts. Forecasts are of two types. The first is known as ex-ante forecast where the specified model is used to generate data on effect variable for the sample periods or units depending upon whether the data is time-series or cross-section. The other type of forecasts are known ex-post where the estimated model is used to obtain forecasts on effect variable for the periods or units outside the sample. Obviously, the accuracy of ex-post forecasts will depend upon the accuracy of ex-ante forecasts besides other factors.[1]

[1] Discussion on various factors that effect accuracy of forecasts is available in [1] where the issues have been examined empirically as well as theoretically.

It should be borne in mind that accuracy of ex-ante forecasts helps in getting better ex-post forecasts but does not guarantee the same because other factors, as pointed out above, have their own role to play. Therefore, we may consider the accuracy of ex-ante forecasts as a pre-requisite for the accuracy of ex-post forecasts.

Let us examine the problem of developing a measure for testing the accuracy of ex-ante forecasts from alternative specifications implied by the model (1.1). Supposing that n observations are available on the variables involved, then, for the i-th sample unit, we can write relation (1.1) as

$$y_i = f(x_{1i}, \ldots, x_{ki}, u_i) \qquad (2.1)$$

$$i = 1, \ldots, n.$$

Further, whatever may be the specified form, let us denote the estimated form f(.) so that ex-ante forecast from the specification under consideration are given by

$$\hat{y}_i = \hat{f}(x_{1i}, \ldots, x_{ki}) \qquad (2.2)$$

$$i = 1, \ldots, n$$

The ex-ante forecast, namely, $\hat{y}_i$ will be perfect if it is equal to the observed value y_i for each i. The measure of closeness will vary from one form to other. For all estimating methods

used the estimates $\hat{y}_i$ depend upon y_i, and if both are measured around their respective means then closeness between the two can be measured by finding correlation coefficient for the model.

$$\hat{y}_i = \beta y_i + u_i \tag{2.3}$$

where the error term u_i tends towards zero as β approaches unity. Sinex y_i may be computed in relation to different forms of $f(x_1, \ldots, x_k)$ the model (2.3) will yield as many correlations as the number of forms of which the highest one may be denoted by η^2, usually known as correlation ratio. This is a measure of explanatory power of the model. Thus we may write

$$\eta^2 = \text{maximum correlation from amongst those between } y \text{ and various forms of } f(x_1,\ldots,x_k) = \max(\eta_1^2, \ldots, \eta_F^2) \tag{2.4}$$

where F denotes possible number of functional forms and η_i^2 is correlation ratio corresponding to i-th forces. In particular if $f(x_1,\ldots,x_k)$ is linear as in (1.2) and coefficients are estimated by least squares procedure, then $\eta^2 = \rho^2$ where

$$\rho^2 = \text{maximum correlation between } y \text{ and linear form (1.2) of } f(x_1,\ldots,x_k) \tag{2.5}$$

The maximum is achieved when β's are estimated according to least squares. From this it follows that

(2.6) $\eta^2 \geq \rho^2$

Keeping the concept of y to be same and altering functional forms of $f(x_1,\ldots,x_k)$, the choice should obviously go in favour of the model that leads to higher value of η^2.

Computation of η^2 can be made straight away for models like (1.2) and (1.5) where same y is explained by alternative forms. It cannot be done similarly in case of models like (1.3) and (1.4) where estimated models yield estimates of log y rather than y. The measure η^2 for model (1.3) and (1.4) represent correlation between original log y_i and estimated log y_i which is not the same as correlation between y_i and y_i. Estimates of effect variable from all forms of $f(x_1,\ldots,x_k)$ must be transformed back into original dimension and correlations may be computed in terms of those to get comparable estimates of η^2.

An estimate of ρ^2 from a sample of size n is represented by R^2, the multiple correlation, which can be defined as

(2.7) R^2 = Square of correlation between y_i and $\hat{y}_i$ when $\hat{y}_i$ is estimated from a linear form (1.2) according to least-squares.

To find similar measures for models like (1.3) and (1.4) one can proceed as follows. Let log y_i be represented by

w_i and $\hat{w}_i$ be estimate of w_i, then, we can find

(2.8) $\hat{z}_i = \text{anti-log } \hat{w}_i$.

so that $\hat{z}_i$ has same dimension as y_i. We may now define a measure of explanatory power of models like (1.3) and (1.4) as follows:

(2.9) R_1^2 = square of correlation between y_i and $\hat{z}_i$.

The measure R_1^2 is an estimate of η^2 when $f(x_1,\ldots,x_k)$ is exponential. Therefore to see as to whether relation (2.6) is valid in case of models (1.3) and (1.4) one can compare R_1^2 with R^2. Sometimes one is tempted to compare R^2 with R_*^2 defined as

(2.10) R_*^2 = square of correlation between w_i and $\hat{w}_i$

But such a comparison could be misleading because R_*^2 cannot be said to estimate η^2 in any sense. The conclusions could also be misleading in actual empirical situation as described in the next section.

3 Empirical Results

The anomaly involved in irrelevant comparison of explanatory powers is illustrated empirically in this section by using linear and double log models for explaining demand for schools in Gujarat, Bank credit for scheduled Commercial Banks in India

and Sugar for entire India. Data for demand for schools function are cross-section, sampled from those given in 1961 Census of Gujarat State, while for the other two functions time-series data are used. Sample period for bank credit consists of annual data over the period 1948-49 to 1967-68 while that for sugar are again annual figures over the period of 1955-56 to 1972-73.[2]

Using the data various models were estimated. Results for linear and double-log models are reported in the following table.

Table 1
Estimated Demand Functions

Demand Variable	Form	Estimated Model
Bank Credit	Linear	C = -41.49+0.43TD+0.97DD+8.99I
	Double-log	Log C = 0.17 + 0.32 log TD+0.55 log DD + 0.44 log I
Schools	Linear	S = 0.05 + 0.13P - 0.12B
	Double-log	Log S = -1.89 + 1.01 log P - 0.11 log B
Sugar	Linear	SU = -6251.62 - 5.00 PS + 1.29 PG + 183.08UP - 26.83TC - 0.61CC
	Double-log	Log SU = -4.14 - 0.46 log PS + 8.11 log PG + 5.09 log UP - 1.84 log TC +0.14 log CC

[2] Data are reported in Misra [17] and to be made available with details on demand.

The symbols stand for:

C	:	Demand for bank credit
TD	:	Time deposit with scheduled banks
DD	:	Demand deposit with scheduled banks
I	:	Loan rate
S	:	Demand for schools
P	:	Population
B	:	Number of business houses
SU	:	Demand for Sugar
PS	:	Price of Sugar
PG	:	Price of Gur
UP	:	Urban Population
TC	:	Tea Consumption
CC	:	Coffee Consumption

We skip over discussion on relevance, and significance of estimated results and concentrate upon comparison of explanatory power alone for the purposes of this note. Most computer programmes provide results for R^2, as defined in (2.7), for linear models and $\bar{R}^2_*$, as defined in (2.10), for double-log models. We have also computed R^2_1, as defined in (2.9), for double-log models and present the results in the following table:

Table 2

Alternative Estimates of Explanatory Power

Demand Variable	R^2	R^2_*	R^2_1
Bank Credit	0.990	0.987	0.990
Schools	0.940	0.980	0.940
Sugar	0.860	0.970	0.880

The results, as given above, do not suggest that a high or low magnitude of R^2_* does always imply high or low magnitude of R^2_1. For bank credit equation $R^2_* < R^2$ but $R^2_1 = R^2$, for schools equation $R^2_* > R^2$ but $R^2_1 < R^2$, and for sugar $R^2_* > R^2$ and $R^2_1 > R^2$. Thus all kinds of relations are possible and R^2_* does not bear any definite relation with R^2_1. Remembering that R^2_1 is the concept that resembles η^2 and the fact that it does not bear any relation with R^2_* we conclude that measures such as R^2_* should not be used to compare explanatory power of econometric models. At the same time, a meaningful comparison is possible if one computes measures comparable to η^2 and then finds out the model that leads to highest magnitude of η^2. It is desirable that enough number of alternative models are estimated and high magnitude of η^2 is considered as only one of the several other relevant factors to choose the model for inference and