267



# Working Paper





विद्यातिनियोगाद्विकासः
IIM
AHMEDABAD

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD

# UNWANTED CONSEQUENCES OF LARGE SAMPLE SIZE IN ECONOMETRIC ESTIMATION

by

P N Misra

W P No. 267
Jan. 1979





INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD

# UNWANTED CONSEQUENCES OF LARGE SAMPLE SIZE IN ECONOMETRIC ESTIMATION

by

P.N.Misra

Summary

In this paper we start with the problem of analysing unwanted consequences of large sample size in econometric estimation and find that the problem can be framed as special case to general problem of estimating a model subject to linear restrictions on the parameters. It is proved that use of large sample size leads to biased, ineffcient and inconsistent estimators in the presence of slightest structural change over the observation span. Explanatory power of the model is also shown to fall down. The analysis is extended to provide a general test-statistic that embraces in its ambit almost all the tests known for testing various hypotheses in context to estimation and prediction from linear models. The same test helps in testing hypotheses relating to alternative specifications of variables involved in the model.

The results are utilised to suggest a method of segmentation of a population or observation space in relation to a hypothesised econometric model. The idea so developed is helpful in defining samples and populations when data are required to be collected to estimate a relation. The same idea can be used to group a given number of units into structurally homogenous groups.

# UNWANTED CONSEQUENCES OF LARGE SAMPLE SIZE IN ECONOMETRIC ESTIMATION

by

P.N. Misra

## 1. Introduction

There appears to be universal belief amongst applied researchers that large sample size is basic requirement for reliable inference. It is only the cost of study that usually provides upper bound in most cases. The quest for large sample size is reinforced by prevailing statistical tests based upon F and t statistics. It is well known that for a regression model specified as

$$(1.1) \quad y_i = \sum_{j=0}^{K} \beta_j X_{ji} + u_i$$

an F statistic is computed as below

$$(1.2) \quad F = \frac{n-K-1}{K} \quad \frac{R^2}{1-R^2}$$

where n represents number of sample observations, K represents number of independent variables and $R^2$ represents squared multiple correlation. For testing the null hypothesis relating to $\rho^2$, the population counterpart of $R^2$, as given below

$$(1.3) \quad H_o : \rho^2 = 0$$

one requires the computed F to be larger than the corresponding tabular F. Computed F, as in (1.2), can be made large apparently

by decreasing K and increasing n but decrease in K leads to reduction in $R^2$ also and therefore this alternative is not given any importance. Increase in n is thought to be safer though as we shall see later, it is not so. Often, applied researchers tend to accept a model if n is large enough to provide significant value of F eventhough $R^2$ is quite low. This leaves an uneasy feeling because one tends to believe that low value of $R^2$ should not be tolerated even though F test were significant.

An obvious solution to this problem is to test the null hypothesis,

$$(1.4) \qquad H_o : \rho^2 = \rho_o^2$$

where $\rho_o^2$ is any chosen level inclusive of zero. Here one can test as to whether $\rho^2$ is adequate enough for the model being used for inference and prediction on the basis of sample data. ~~We propose to provide [illegible] in this paper.~~

One may define the following statistic, alternatively as

$$(1.5) \qquad F = \frac{R^2}{I-R^2}$$

where n and K may influence F via influencing $R^2$ only. It is well known that increase in K leads to increase in $R^2$ and therefore F in (1.5) increases as K increases. But precise contribution of change in n to $R^2$ is not known certainly in algebraic sense.

Present paper analyses these issues and provides appropriate solutions.

Some discussion on the nature of sample and population concepts in econometric analyses is due before we proceed further. Generally, we are concerned with estimation of behavioural relations or such other types of relations that change in structural sense over time and space. This fact is not consistent with the usual assumption that a time series data is a simple random sample with replacement from a population represented by the time span $-\infty$ to $\infty$ . A proper definition of a population in econometric sense is collection of those units which satisfy the model to be estimated. Thus the units relating to which the observations satisfy a relation like (1.1) constitute the population because only in such case the unknown coefficients in (1.1) can be supposed to remain invariant over the observation space. A sample from only such a population will be a representative sample. Accordingly, the time span $-\infty$ to $\infty$ can not be defined to constitute population for all kinds of time series data in the presence of structural changes. Only that part of the time span over which the model to be estimated can be supposed to remain invariant is the relevant population and this may naturally vary from model to model. A sample needs to be selected only after the population has been identified. The same holds good for cross-section data. Therefore, a major problem is to segment a given set of observations into homogenous populations where homogeniety is to be ensured with respect to the model to be

estimated. On closer scrutiny, resolution of this problem solves the problem of stratification in multivariate situation, segmentation of markets, clustering, aggregation over relations and variables, and formation of internally homogenous groups, etc. It may be observed that stratification or clustering with respect to magnitudes of the variables involved is neither similar nor preferable to the concept of structural segmentation described above.

A number of problems of estimation and inference, which have been discussed in relevant literature in somewhat isolated fashion, have been synthesised in this paper. All the testing procedures relating to linear models including t, F and those in Chow (1960) and Fisher, F.M. (1970) are shown to be special cases of the test procedure emerging from our approach. The problems of prediction, specification error, dummy variables, qualitative variables, covariance analysis, compatibility of prior information, pooling of time-series and cross-section data, etc., are shown to be derivable from our approach. The idea of discriminant analysis has been extended in context to populations satisfying multivariate regression models. The problem of random coefficients is seen to be extension of the problem discussed in this paper.

## 2 Effect of Structural Changes on Estimated Models

Since users of large samples are likely to treat differing structurers to be similar, the problems of large samples and structural changes are two sides of the same coin. Let us consider an observation

span consisting of n units. Let it be divided into two groups or populations consisting of $n_1$ and $n_2$ units, respectively, so that observations in each group are internally homogenous with respect to a specified model such as (1.1). Structural homogeneity implies that the model remains the same over various observational units and the same holds good for the coefficients. Using usual notations in econometrics, we may express the models in matrix form as

$$(2.1) \quad y_1 = X_1\beta_1+u_1$$

for $n_1$ observations in the first population and

$$(2.2) \quad y_2 = X_2\beta_2+u_2$$

for $n_2$ observations in the second population. The terms population and group are used interchangeably in proper context throughout this paper. The vectors $y_1$ and $u_1$ are each of size $n_1x1$, matrix $X_1$ is of size $n_1x(K+1)$, vector $\beta$ is of size $(K+1)x1$ and the same holds good for model (2.2) with $n_1$ replaced by $n_2$. The total number of units n on both the populations taken together is given as

$$(2.3) \quad n = n_1+n_2$$

Models (2.1) and (2.2) can be written together as

$$(2.4) \quad y^* = X^*\beta^*+u^*$$

where the notations are given by

$$(2.5)\quad y^* = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad X^* = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix},$$

$$\beta^* = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad u^* = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

Let us assume that vectors $\beta_1$ and $\beta_2$ are linearly related as

$$(2.6)\quad R_1\beta_1 + R_2\beta_2 = r$$

This can be written, alternatively, as

$$(2.7)\quad R\beta^* = r$$

where R is of size $C\times(2K+2)$ and is defined as

$$(2.8)\quad R = [R_1 \; R_2]$$

and $R_1$, $R_2$ are each of size $C\times(K+1)$ where C represents the number of constraints. The elements of R and r are supposed to be known constants. In particular, if

$$(2.9)\quad R = [I \; -I]$$

$$r = 0$$

where I is $(K+1) \times (K+1)$ identity matrix and 0 is $(K+1)\times 1$ null vector, the restriction (2.7) simplifies to

$$(2.10)\quad \beta_1 = \beta_2 = \beta$$

and model (2.4) gets simplified as

$$(2.11) \quad y^* = X\beta + u^*$$

$$X = X^* \begin{bmatrix} I \\ I \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}.$$

The model (2.11) is now defined in terms of pooled observations in terms of n observations on the variables involved. Thus estimation of model (2.11) in terms of large sample size implies estimation of model (2.4) under the restriction (2.10). In other words, the problem at hand is special case of estimation under the restriction (2.7). We shall, therefore, consider the problem of estimating model (2.4) under the restriction (2.7).

## 2.1 Estimation

Let us consider a Cx1 vector of Lagrangian constants, $\lambda$ , and define the following function

$$(2.12) \quad \phi = u^{*'}u^* + 2\lambda'(R\beta^* - r)$$

First order condition of optimisation of $\phi$ is given by

$$(2.13) \quad X^{*'}X^*\beta^* + R'\lambda = X^{*'}y^*$$

$$R\beta^* = r$$

Second order condition of optimisation can be obtained by examining the Hessian

$$(2.14) \qquad H = \begin{vmatrix} 2X^{*'}X^{*} & R' \\ R & 0 \end{vmatrix} = -\left|X^{*'}X^{*}\right| \left|R'(X^{*'}X^{*})^{-1}R\right|$$

Principal minors of this Hessian are negative definite, Therefore solution of $\beta^{*}$ and $\lambda$ from (2.13) minimise the function (2.12). Constrained least squares solution for $\beta^{*}$ and $\lambda$ can be written as

$$(2.15) \qquad \begin{aligned} b^{*} &= \hat{\beta}^{*} + (X^{*'}X^{*})^{-1}R'\hat{\lambda} \\ \hat{\lambda} &= N(r - R\hat{\beta}^{*}) \\ \hat{\beta}^{*} &= (X^{*'}X^{*})^{-1}X^{*'}y^{*} \\ N &= \left[R(X^{*'}X^{*})^{-1}R'\right]^{-1} \end{aligned}$$

where $\hat{\beta}^{*}$ is unrestricted least-squares estimator of $\beta^{*}$. It can be easily verified that

$$(2.16) \qquad \begin{aligned} Rb^{*} &= R\hat{\beta}^{*} + r - R\hat{\beta}^{*} \\ &= r \end{aligned}$$

so that the estimator $b^{*}$ satisfies the constraint (2.7) but the same cannot be said for estimator $\hat{\beta}^{*}$. In general one may observe that

$$(2.17) \qquad R\hat{\beta}^{*} \neq r$$

If (2.17) were not true, that is $R\hat{\beta}^{*}=r$, one would find that

$$(2.18) \qquad b^{*} = \hat{\beta}^{*}$$

implying that the restriction (2.7) is in fact redundant. Thus it follows that restricted least-squares is needed only when (2.17) holds, or, when one wishes to impose restrictions that are not supported by unrestricted least-squares. In particular, if restrictions are given by (2.10) and $R\hat{\beta}^* = 0$, that is

$$(2.19) \quad \hat{\beta}_1 = \hat{\beta}_2$$

then, (2.18) can be expressed as

$$(2.20) \quad b^* = \begin{bmatrix} I \\ I \end{bmatrix} \hat{\beta}$$

$$\hat{\beta} = (X'X)^{-1}X'y^*$$

where $\hat{\beta}$ can be estimated directly from the pooled model (2.11). It follows, therefore, that pooling is desirable when (2.19) holds good. Any divergence between $\hat{\beta}_1$ and $\hat{\beta}_2$ is suggestive of treating the two populations separately. Pooling can be tolerated in statistical sense provided $\hat{\beta}_1$ and $\hat{\beta}_2$ are different from each other only insignificantly. Properties of the estimator $b^*$ in (2.15) were analysed by Misra (1973) by treating the parameters to be mis-specified over the sample space. In what follows, these are examined in slightly different way to afford greater generalisation.

## 2.2 Bias

Combining (2.4) with (2.15) we obtain

$$(2.21) \quad b^* - \beta^* = A(r - R\beta^*) + A_1 u^*$$

$$A = (X^{*'}X^*)^{-1}R'N$$

$$A_1 = (I - AR)(X^{*'}X^*)^{-1}X^{*'}$$

Taking expectation on both sides of (2.21) under usual least square assumptions we have

$$(2.22) \quad E(b^*-\beta^*) = A(r-R\beta^*)$$

This shows that estimator $b^*$ is biased in case the restriction (2.7) is not true. The bias can be estimated by replacing $\beta^*$ by $\hat{\beta}^*$. Again the estimated bias will not disappear if data are such that (2.17) is true. It follows therefore, that quest for large samples will generally lead to biased estimates in case pooling of observations is done as in (2.11).

The estimator $\hat{\beta}$ of $\beta$ in model (2.11) can be expressed, alternatively, as

$$(2.23) \quad \hat{\beta} = W_1\hat{\beta}_1 + W_2\hat{\beta}_2$$

$$W_1 = (X'X)^{-1}X_1'X_1$$

$$W_2 = (X'X)^{-1}X_2'X_2$$

$$W_1 + W_2 = I$$

which shows that $\hat{\beta}$ is weighted average of $\hat{\beta}_1$ and $\hat{\beta}_2$ and therefore differs from both of these unless (2.19) holds good. This again shows that $\hat{\beta}$ is unable to estimate $\beta_1$ or $\beta_2$ unbiasedly. It may be noted that $\beta$ is simply artificially constructed parameter vector to substitute the parameter vectors $\beta_1$ and $\beta_2$.

## 2.3 Efficiency

Using (2.21) and definition

$$(2.24) \quad E\, u^*u^{*'} = \Sigma^*$$

we obtain variance-covariance matrix of estimator b* as follows

$$(2.25) \quad E(b^*-\beta^*)\,(b^*-\beta^*)' = A(r-R\beta^*)\,(r-R\beta^*)'A' + A_1\Sigma^*A_1'$$

The expressions in (2.25) can be further simplified as

$$(2.26) \quad E(b^*-\beta^*)\,(b^*-\beta^*)' = V_1+A(V_2-V_3)A'$$

where

$$(2.27) \quad V_1 = (X^{*'}X^*)^{-1}X^{*'}\Sigma^*X^*(X^{*'}X^*)^{-1}$$

$$V_2 = (r-R\beta^*)\,(r-R\beta^*)'$$

$$V_3 = E\left[(r-R\hat{\beta}^*)-(r-R\beta^*)\right]\left[(r-R\hat{\beta}^*)-(r-R\beta^*)\right]'$$

$$= R\,V_1\,R'$$

Further, since $\hat{\beta}^*$ is unbiased estimator of $\beta^*$, we can derive

$$(2.28) \quad V_3 = V_4-V_2$$

$$V_4 = E(r-R\hat{\beta}^*)\,(r-R\hat{\beta}^*)'$$

Combining (2.28) with (2.26) we get

$$(2.29) \quad E(b^*-\beta^*)\,(b^*-\beta^*)' = V_1+AV_2A'+A(V_2-V_4)A'$$

The result (2.29) shows that the estimator b* is likely to be inefficient as compared to $\hat{\beta}^*$ owing to several reasons. Firstly, $\beta^*$ is best estimated by $\hat{\beta}^*$ and in that case estimated value of $V_2-V_4$

will be zero so that estimated variance of b* is larger than that of $\hat{\beta}^*$ by an amount $A\hat{V}_2A'$ where $\hat{V}_2$ is estimate of $V_2$. Secondly, it is likely in most situations that structural changes are so frequent that $n_1$ and $n_2$ represent population sizes. In that case both $V_1$ and $V_3$ will disappear and

$$(2.30) \quad E(b^*-\beta^*)(b^*-\beta^*)' = AV_2A'$$

which does not vanish unless both $R\beta^*=r$ and $R\hat{\beta}^*=r$ hold good.

The result (2.30) shows that estimator b* will be inconsistent if the restriction (2.7) is not true. In other words, use of large sample size in the presence of structural change is most likely to lead to inconsistent estimate of unknown coefficients even though all the least-squares assumptions were true.

4 Explanatory Power

Using estimator b* of $\beta^*$ we can express

$$(2.31) \quad y^* = \hat{y}^* + \hat{u}^*$$

where

$$(2.32) \quad \hat{y}^* = X^*b^* = M^*y^*+X^*A(r-R\hat{\beta}^*)$$

$$\hat{u}^* = My^*-X^*A(r-R\hat{\beta}^*)$$

$$M^* = X^*(X^{*'}X^*)^{-1}X^{*'}$$

$$M = I-M^*$$

and A is same as defined in (2.21). Remembering that

$$(2.33)\quad M^*M^* = M^* \qquad MM = M \qquad MM^* = MX^* - X^{*\prime}M = 0$$

we can derive the following results:

$$(2.34)\quad \hat{y}^{*\prime}\hat{y}^* = y^{*\prime}M^*y^* + r'Nr - \hat{r}'N\hat{r}$$
$$\hat{u}^{*\prime}\hat{u}^* = y^{*\prime}My^* + (r-\hat{r})'N(r-\hat{r})$$
$$\hat{y}^{*\prime}\hat{u}^* = -\hat{r}'N(r-\hat{r}) - (r-\hat{r})'N(r-\hat{r}) = -r'N(r-\hat{r})$$
$$r = R\hat{\beta}^*$$

It can be easily verified that

$$(2.35)\quad \hat{y}^{*\prime}\hat{y}^* + \hat{u}^{*\prime}\hat{u}^* + 2\,\hat{y}^{*\prime}\hat{u}^* = y^{*\prime}M^*y^* + y^{*\prime}My^*$$

The result (2.34) shows that $\hat{y}^{*\prime}\hat{u}^*$ is not zero unless $r = \hat{r}$ or $r = 0$. This gives rise to what is known as covariance analysis. In other words, for all kinds of restrictions where $r = 0$, the components $\hat{y}^*$ and $\hat{u}^*$ are orthogonal. In that case explained and residual sum of squares can be written as

$$(2.36)\quad \hat{y}^{*\prime}\hat{y}^* = y^{*\prime}M^*y^* - \hat{r}'N\hat{r}$$
$$\hat{u}^{*\prime}\hat{u}^* = y^{*\prime}M\,y^* + \hat{r}'N\hat{r}$$

This result shows that restrictions similar to (2.10) are capable of reducing explained sum of squares and increasing the residual sum of squares and this undesirable property gets further enhanced

as $R\hat{\beta}^*$ is found to deviate from the null vector. In other words, attempt to use large samples in the presence of structural change or to mix samples from two different populations leads to reduction in explanatory power. The reduction could be substantial if the structures are widely apart from each other.

## 3 Testing of Hypotheses

A careful examination of the test statistics developed for testing various hypotheses including those by Chow (1960) and Fisher (1970) reveals that the tests have to be constructed separately for every hypothesis. All such statistics are found to be special cases of the statistic, developed in this section to test the hypothesis that restriction (2.7) holds. For predictive tests, however, the statistics have to be developed separately when $n_2 < K+1$. Most of the results developed so far, can be adapted to restricted estimation of a single equation model. The results can also be extended to situations containing more than two regressions. We propose to develop an appropriate statistic for testing (2.7) and then show that most of the known as well as unknown tests are special cases of the same. Besides theoretical neatness, the proposed approach can prove helpful in affecting economy in computerisation of various tests. Predictive tests are developed separately.

### 3.1 Test for $R\beta^* = r$

We may express explained variation in (2.34) as

$$(3.1)\quad \hat{y}^{*\prime}\hat{y}^{*} = S_f - S_r$$

$$S_f = y^{*\prime}M^{*}y^{*}$$

$$S_r = \hat{r}'N\hat{r} - r'Nr$$

where $S_r$ represents variation ascribable to restriction (2.7) and $S_f$ represents variation ascribable to model (2.4) when it is free from any restriction. Remembering that

$$(3.2)\quad \hat{r} = r + R(X^{*\prime}X^{*})^{-1}X^{*\prime}u^{*}$$

we may express $S_r$, alternatively, as

$$(3.3)\quad S_r = 2\, r'NR(X^{*\prime}X^{*})^{-1}X^{*\prime}u^{*} - u^{*\prime}X^{*}(X^{*\prime}X^{*})^{-1}R'NR(X^{*\prime}X^{*})^{-1}X^{*\prime}u^{*}$$

Using (2.33) we observe that both of the components of $S_r$ are statistically independent with the quadratic form

$$(3.4)\quad S_{ef} = y^{*\prime}My^{*} = u^{*\prime}Mu^{*}$$

where $S_{ef}$ represents residual sum of squares of the unrestricted model (2.4).

Assuming that the errors in (2.4) are independent and homoschedastic and have zero means, we can write (2.24) as

$$(3.5)\quad \Sigma^{*} = \sigma^{2}I$$

Using these assumptions we obtain the following results:

$$(3.6) \quad ES_r = \sigma^2 R(X^{*\prime}X^*)^{-1}X^{*\prime}X^*(X^{*\prime}X^*)^{-1}RN = \sigma^2 N^{-1}N$$
$$= \sigma^2 C$$

and

$$(3.7) \quad ES_{ef} = \sigma^2(n-2K-2)$$

where C represents number of constraints on the parameters and also the size of the square matrix N. Therefore we can define an F statistic as

$$(3.8) \quad F\,(C,\ n-2K-2) = \frac{n-2K-2}{C} \quad \frac{\hat{r}'N\hat{r}-r'Nr}{y^{*\prime}My^*}$$

with C and n-2K-2 degrees of freedom to test the restriction (2.7). The statistic F can be computed in any given situation and if it exceeds the tabular value of F, the statistical validity of the restriction (2.7 may be seriously doubted because $S_r$ tends to be larger owing to failure of restriction (2.7) to be compatible with the sample observations.

In case one were dealing with a single regression model like (2.1) and the corresponding restriction as in (2.7), one would obtain an F statistic as

$$(3.9) \quad F\,(C,\ n_1-K-1) = \frac{n_1-K-1}{C} \quad \frac{\hat{r}'N_1\hat{r} - r'N_1r}{y_1'M_1y_1}$$

where the subscripted expressions refer to corresponding concepts when model (2.1) is used instead of model (2.4).

The test holds good for all types of exact restrictions. What one

requires is to use relevant values of R and r to compute the needed statistic. For instance, if one were interested to test constant returns to scale in context to Cobb-Douglas production functions, one can set r = 1 and define R to be an appropriate row vector so that $R\beta$ provides the desired sum of the coefficients.

## 3.2 Special Cases

As pointed out earlier, specific hypotheses can always be transformed in the form (2.7) with specific values of R and r. At the same time one may also determine the corresponding degrees of freedom. In what follows, we shall illustrate in some cases as to how the statistic (3.8) leads to certain well known tests. We will also illustrate in certain other cases as to how certain hypotheses can be transformed in the form (2.7) so that the test (3.8) could be applied.

### Contribution of a Single Variable in a Single Model

Let us consider the hypothesis

$$\beta_j = 0 \tag{3.10}$$

in context to model (1.1). Here we can identify R and r as

$$R = (00..1\ 0...0) \tag{3.11}$$

$$r = 0$$

$$C = 1$$

where j-th place of R contains unity. With these values and formula (3.9) we obtain the relevant statistic as

$$(3.12) \qquad F(1, n-K-1) = \frac{\hat{\beta}_j^2}{\text{var}(\hat{\beta}_j)} = t^2$$

where $\text{var}(\hat{\beta}_j)$ represents unbiasedly estimated variance of $\hat{\beta}_j$. The statistic (3.12) provides the well known t-test.

Simultaneous Contribution of All the Causal Variables

In this situation the hypothesis to be tested is given as

$$(3.13) \qquad \beta_1 = \beta_2 = \ldots = \beta_K = 0$$

If we redefine the model (1.1) with variables in terms of deviations around their sample means, then we have

$$(3.14) \qquad R = I \qquad r = 0 \qquad C = K$$

where I represents KxK identity matrix. Combining (3.14) with (3.9) we obtain the relevant statistic as

$$(3.15) \qquad F(K, n-K-1) = \frac{n-K-1}{K} \quad \frac{y'M^*y}{y'My} = \frac{n-K-1}{K} \quad \frac{R^2}{I-R^2}$$

where observations are measured around their sample means. The test (3.15) is same as (1.2) which is the well known F test for testing the statistical validity of hypothesised model (1.1) in terms of n sample observations.

Alternatively, one may consider a hypothesis in terms of subsets of the parameters and identify the corresponding R. One may also consider equality, proportionality or linearity of two or more coefficients in model (1.1) and in each case identify R and r to define the corresponding test statistic.

Equality of Coefficients in Two Regressions

Let us consider the problem of testing the equality of parametric vectors as in (2.10) with R and r as defined in (2.9). In this case the test statistic in (3.8) gets simplified as

$$(3.16)\quad F(K+1,\ n-2K-2) = \frac{n-2K-2}{K+1}\ \frac{\hat{r}'N\hat{r}}{y^{*\prime}My^{*}}$$

$$\hat{r}'N\hat{r} = (\hat{\beta}_1-\hat{\beta}_2)'\left[(X_1'X_1)^{-1}+(X_2'X_2)^{-1}\right]^{-1}(\hat{\beta}_1-\hat{\beta}_2)$$

$$y^{*\prime}My^{*} = y_1'M_1y_1 + y_2'M_2y_2$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ represent unrestricted estimates of $\beta_1$ and $\beta_2$, respectively, and $y^{*\prime}My^{*}$ represents sum of residual squares in the two regressions.

One may visualise several cases of restriction (2.10). If it is restricted to equality of constant terms only, one may test for

significance of dummy variables, or differences in time unit specifics or cross-section unit specifics in context to estimation of such pooling models as in Misra (1972C, 1976). It may be noted that tests in context to covariance analysis also fall in this category. One may consider equality or any linear combination of a subset of coefficients and use the test statistic (3.8) with appropriate R and r.

There are several interesting applications of the test in (3.16). It can help us in identifying those micro relations which if defined in terms of macro variables will have no aggregation bias in estimates as well as predictions. The truth of this statement can be easily verified by considering relevant results in Misra (1967, 1969a, 1969b). The test can be used to decide as to whether two sample sizes $n_1$ and $n_2$ can be pooled together to obtain larger sample size. Test for aggregability of micro units may help in determining homogenous groups of units (firms, consumers, investors, etc.) or regions where homogeneity is sought to be in terms of hypothesised structures.

Specification of X

A major problem of decision making in actual practice is to decide specification of variables in matrix X. They could be totals, ratios, percentages, first order differences, anticipated values, lagged values, qualitative or any other form and one does not known which one of these is the right choice. Representing two alternative specifications of X by $X_1$ and $X_2$, we can specify the following models to explain the same y:

(3.17) $y = X_1\beta_1+u_1$

$y = X_2\beta_2+u_2$

Now the hypotheses $X_1 = X_2$ and $\beta_1=\beta_2$ are equivalent in context to above specification. Therefore, one may use (3.16) to test (2.10) and if $\beta_1$ is found to be different from $\beta_2$, the choice of the specification in (3.17) leading to higher $R^2$ is more desirable.

Specification of Y

Alternative specifications of y variable corresponding to given X can be tested by testing the restriction (2.10) in context to models

(3.18) $y_1 = X\beta_1+u_1$

$y_2 = X\beta_2+u_2$

If $\beta_1$ is found to be significantly different from $\beta_2$, one may opt for the specification in (3.18) providing higher $R^2$.

The hypotheses $y_1=y_2$ and $X_1=X_2$ considered together are equivalent to testing the restriction (2.10) to lead to test in (3.16). In this case also if $\beta_1\neq\beta_2$ one may opt for the specification that yields higher $R^2$.

Effect of Inclusion of a Variable

Several times one is interested in testing as to whether inclusion of an additional variable, $x_a$, changes the coefficients in a model significantly. Thus given a model as

(3.19) $y = X_1\beta_1+u_1$

one may augment the matrix $X_1$ as

$$(3.20) \qquad X_2 = (X_1 \; x_a)$$

and then estimate the model

$$(3.21) \qquad y = X_2 \beta_2 + u_2$$

The problem is, then, to test as to whether the common coefficients in $\beta_1$ and $\beta_2$ are different from each other when considered, singly or jointly together. In case one is interested in testing the significance of change in $i^{th}$ coefficient, one may consider the following restriction

$$(3.22) \qquad \beta_{1a} = 0$$

$$\beta_{1i} = \beta_{2i}$$

for each i. The coefficient $\beta_{1a}$ represents coefficient of $x_a$ in model (3.19) when $x_a$ is assumed to be hypothetically included in $X_1$. Given the restrictions in (3.22), one may identify R and r and carry on the test as in (3.8).

### 3.3 Predictive Tests

The tests in section 3.1 hold good provided both $n_1$ and $n_2$ are greater than K+1. In actual practice one faces several situations when $n_1 > K+1$ but $n_2 < K+1$. In almost all the short-term prediction exercises $n_2$ includes the periods to be forecasted and these normally

do not exceed the number of explanatory variables. The problem is therefore to test the restriction (2.10) when $\beta_1$ can be estimated from model (2.1) but $\beta_2$ cannot be estimated from model (2.2) owing to shortage of number of observations. We shall discuss two interesting situations.

Equality of $\beta_1$ and $\beta_2$

Assuming (2.10) holds good, one may estimate $\beta$ from (2.11) to obtain the estimator in (2.20). This estimator may be used to obtain

$$(3.23) \quad \hat{y}_1 = X_1\hat{\beta}$$

Alternatively $y_1$ may be estimated directly from model (2.1) as $X_1\hat{\beta}_1$ without imposing any restriction. Using these estimates, we may obtain present version of $S_r$ as

$$(3.24) \quad S_r^* = \hat{\beta}_1'X_1'X_1\hat{\beta}_1 - \hat{\beta}'X_1'X_1\hat{\beta}$$

The hypothesis to be tested in the present case is as to whether $\beta_1 = \beta_2$ and as to whether $y_2$ is predicted accurately by $X_2\beta_1$. This is equivalent to testing the restriction

$$(3.25) \quad \beta_1 = \beta_2 = \beta$$
$$u_2 = 0$$

Using (3.25) we have

$$(3.26) \quad y = X\beta + u_o$$
$$y_1 = X_1\beta + u_1$$
$$u_o = \begin{bmatrix} u_1 \\ \bullet \end{bmatrix}$$

Combining (3.26) with (3.24) and using relevant expressions for $\hat{\beta}_1$ and $\hat{\beta}$ we get variation owing to restriction (3.25) as

$$(3.27) \quad S^*_r = u_1'X_1(X_1'X_1)^{-1}X_1'u_1 - u_1'X_1(X'X)^{-1}X_1'X_1(X'X)^{-1}X_1'u_1 + \text{expressions in terms of } X_1'u_1$$

It can be easily verified that individual terms on the right hand side of (3.27) are distributed independently with

$$(3.28) \quad S^*_{ef} = u_1' \left[I - X_1(X_1'X_1)^{-1}X_1'\right] u_1$$

Therefore we may define an F statistic as

$$(3.29) \quad F(d_1, n_1-K-1) = \frac{n_1-K-1}{d_1} \quad \frac{S^*_r}{S^*_{ef}}$$

where $d_1$ is obtainable from

$$(3.30) \quad \sigma_1^2 d_1 = E(S^*_r) = \sigma_1^2 \left[K+1 - \text{tr}(X'X)^{-1}X_1'X_1(X'X)^{-1}X_1'X_1\right]$$

An interesting adaptation of the test in (3.29) is to test as to whether a restriction like

$$(3.31) \quad R\beta = r$$

is statistically compatible with the $\beta$ defined in model

$$(3.32) \quad y = X\beta + u$$

Compatibility of Additional Data to Estimated Structure

Suppose that estimated version of model (2.1) is used to generate $y_2$ as

$$(3.33)\quad \hat{y}_2 = X_2\hat{\beta}_1$$

whereas the true $y_2$ is given by

$$(3.34)\quad y_2 = X_2\beta_1 + u_2$$

This gives

$$(3.35)\quad \hat{y}_2 - y_2 = X_2(\hat{\beta}_1 - \beta_1) - u_2$$

and

$$(3.36)\quad (\hat{y}_2 - y_2)'(\hat{y}_2 - y_2) = (\hat{\beta}_1 - \beta_1)'X_2'X_2(\hat{\beta}_1 - \beta_1) + u_2'u_2 + \text{terms in } u_2$$

Assuming that $u_2$ and $u_1$ are statistically independent, the terms in (3.36) are found to be statistically independent with $S^*_{ef}$ defined in (3.28). Therefore if $u_1$ and $u_2$ are homoschedastic we can define an F statistic as

$$(3.37)\quad F(d_2, n_1-K-1) = \frac{n_1-K-1}{d_2}\quad\frac{(\hat{y}_2 - y_2)'(\hat{y}_2 - y_2)}{S^*_{ef}}$$

where $d_2$ is obtainable from

$$(3.38)\quad \sigma_1^2 d_2 = E\left[(\hat{y}_2 - y_2)'(\hat{y}_2 - y_2)\right]$$
$$= \sigma_1^2 \operatorname{tr} X_2(X_1'X_1)^{-1}X_2' + \sigma_1^2$$

In particular, if $n_2 = 1$, so that $X_2 = x_*'$, we can express $d_2$ as

$$(3.39) \quad d_2^* = x_*'(X_1'X_1)^{-1}x_{*+1}$$

In this case we can couple $d_2^*$ together with $S_{ef}$ and define an F as

$$(3.40) \quad F(1, n_1-K-1) = \frac{(\hat{y}_2-y_2)^2(n_1-K-1)}{d_2^* \; S^*_{ef}}$$

This shows that we can define a t-statistic

$$(3.41) \quad t = \frac{(\hat{y}_2-y_2)\sqrt{n_1-K-1}}{\sqrt{d_2^* \; S^*_{ef}}}$$

which can be used for obtaining interval estimates of $y_2$ at a specified level of confidence.

The test in (3.37) enables one to test as to whether an external data set $(y_2 X_2)$ belongs to an estimated structure $y_1 = X_1\hat{\beta}_1$. In this sense the test enables one to ascertain as to whether an external set of observations belongs to a given linear model. In other words, the test serves the same purpose in context to regression model as discriminant analysis serves in context to multivariate normal distributions. It has been shown by Misra (1972a, 1972b) that least-squares estimated regressions tend to follow t-distribution even though regression errors are found to be heterotypic. Thus the discriminating ability of the tests in this paper may not get retarded even though the regression errors do not follow any known distribution.

## 4 Formation of Structurally Homogenous Groups

Results developed in the preceding sections can be used to subdivide any n observations into internally homogenous group while homogeneity is ensured with respect to a model. This problem has earlier been posed by Misra (1978, Ch.5), in context to market segmentation in relation to general demand functions and the same holds good in general. Examination of result (2.23) provides an useful clue to resolve this problem. The result holds good for sample as well as population sizes and therefore segmentation can be done in either case. The result (2.23) states that the plane in (2.11) passes through the zone falling in between the planes in (2.1) and (2.2).

Suppose we consider the signs of the residuals corresponding to model (2.11) and group the observations corresponding to positive (inclusive of zero) and negative residuals into $n_1$ and $n_2$ and use these to estimate the models (2.1) and (2.2), respectively. In this situation we have

$$(4.1) \qquad \begin{aligned} X_1\hat{\beta}_1 &\geq X_1\hat{\beta} \\ X_2\hat{\beta} &> X_2\hat{\beta}_2 \end{aligned}$$

Further, remembering the pattern of grouping of the residuals we have

$$(4.2) \qquad \begin{aligned} y_1 - X_1\hat{\beta}_1 &\leq y_1 - X_1\hat{\beta} \\ X_2\hat{\beta} - y_2 &> X_2\hat{\beta}_2 - y_2 \end{aligned}$$

This shows that errors corresponding to models (2.1) and (2.2) are reduced in terms of absolute magnitude when compared to corresponding errors in relation to model (2.11). This implies that the grouping is such that the observations are internally more homogenous with respect to the specified model. In other words, the residual variation corresponding to model (2.1) and (2.2) will be reduced substantially by following the above mentioned subgrouping. The explained variation in both the models will be increased by an amount equal to $S_r$, defined in (3.1), because separate estimation does not require imposition of any restriction like (2.10). The sub-grouping criterion, as outlined above, is therefore capable of reducing error sum of squares and increasing explained sum of squares and these together will lead to improvement in $R^2$. Thus segmentation of observations into homogenous groups should lead to improvement in $R^2$.

This approach holds good for all the models that are estimated in accordance with the style of linear models. To start with, one may consider possible causal variables and divide the total size into two homogenous groups as above. Similar procedure may then be adopted to segment further the two groups separately. The process may be continued till there is no significant improvement in $R^2$ as a result of further segmentation.

If the groupings $n_1$ and $n_2$ are not intra homogenous with respect to a model, the $R^2$ corresponding to pooled model (2.11) is shown by

Misra (1973) to lie in between the $R^2$ values corresponding to models (2.1) and (2.2). The same is found to hold good in respect of empirical estimates of $r^2$ as computed by Fisher (1958, pp.203-204). This is expected in view of the foregoing discussion because any arbitrary division will make one of the groups to be more homogenous as compared to the pooled size and this property will lead to improvement in $R^2$. Correspondingly, the other group is rendered more heterogenous and the result is exhibited in terms of reduced $R^2$.

The concept of stratification in sampling is simply a special case of the present approach when the models are specified as

$$(4.3) \quad y_{1i} = \beta_{1o} + u_{1i} \quad : \quad i = 1, \ldots, n_1$$

$$(4.4) \quad y_{2i} = \beta_{2o} + u_{2i} \quad : \quad i = n_1+1, \ldots, n_2$$

$$(4.5) \quad y_i = \beta_o + u_i \quad : \quad i = 1, \ldots, n$$

where n is same as defined in (2.3)[1]. This shows that stratification may lead to subgroups that may not be proper segments when segmentation is required to be done in relation to models containing explanatory variables.

---

[1] See Misra (1978, Ch.4) where all the sampling results in case of simple random sampling are derived as special cases of a general regression model. The sampling models are, in fact, regression models without having any causal variable in it. The tests in (3.9) and (3.8) can be used for the above simple models to obtain the well-known t tests for testing the significance of sample mean and difference of sample means, respectively.

The idea of segmentation can be used to ascertain the point of structural break in case of time-series as well as cross-section data. The subgroups need not be temporally or spatially contiguous. In time-series data one simply defines a block of time-period to be a sample and the results of the segmentation may provide further insight to redefine the sample span. More often it may turn out to be population span rather than sample span if the structural breaks happen to be more frequent over time. In any case one can group the time-periods into internally homogenous groups and use that structure for predictive purposes that seems to be compatible with the prediction period in question. The same holds good in case of cross-section data.

REFERENCES


Chow, Gregory C., 1960, "Tests of Equality Between Subsets of Coefficients in Two Linear Regressions," Econometrica, 28, pp.591-605.

Fisher, Franklin, M., 1970, "Tests of Equality Between Sets of Coefficients in Two Linear Regressions - An Expository Note," Econometrica, 38, pp. 361-366.

Fisher, R.A., 1958, Statistical Methods for Research Workers, Hafner, New York, pp. 203-204

Misra, P.N., 1967, "Matrix Formulation of the Theory of Linear Aggregation," Sankhya, Series B, 29, pp. 41-46.

Misra, P.N., 1969a, "A Note on Linear Aggregation of Economic Relations," International Economic Review, 10, pp.247-49.

Misra, P.N., 1969b, Some Problems of Estimation and Aggregation in Econometric Models with Applications to Industry, Ph.D. Thesis, Delhi School of Economics, Delhi University.

Misra, P.N., 1972a,"Relation Between Pearsonian Coefficients of Distributions of Least Squares Estimators and the Disturbance Term," JASA, 67, pp.662-68.

Misra, P.N., 1972b, "Distribution of OLS Estimators of Coefficients and Disturbance Term in a General Regression Model," Sankhya, Series B, 34, pp.395-404.

Misra, P.N., 1972C, "Pooling of Time Series and Cross-section Data," Sankhya, Series B, 34, pp.385-394.

Misra, P.N., 1973, "Some Implications of Structural Changes within the sample," Technical Report 5, IIM Ahmedabad, Read at 1973 Indian Econometric Society Conference.

Misra, P.N., 1976, U.K.Srivastava and Deepak Chawla, "An Econometric Approach to Evolve Viable Cattle Development Financing Schemes," Indian Economic Review, XI, pp.91-109.

Misra, P.N., 1978, Forecasting and Control with Applications to Demand and Sales, forthcoming.