THE EFFECT OF OBSERVATIONAL ERRORS ON LEAST SQUARES REGRESSION ESTIMATES

by

DR. CRISTINA P. PAREL*

A statistical approach to the problem of how a variable y varies with another variable x which utilizes a mathematical function is the use of some specified measure of central tendency of all values of y having the same fixed x vaues. For each fixed value of x, the measure of central tendency is defined so that the collective result is a mathematical function of x. It is essential to this concept, however, that x be not an observed value, unless the observation is not subject to error, since the mathematical function calls for the determination of the measure of central tendency for each fixed and arbitrary, and hence, presumably not observed, value of x. It is this restriction to fixed values of x which has made regression theory and least squares regression theory formally inapplicable to many practical problems in which the values of x, as well as the values of y, are observed. However, some sort of approximation to the regression function can be obtained where the errors of the x's are known to be small by assuming that they are zero, but this is theoretically unsatisfactory. Nevertheless, it is the process commonly used in many of the situations in which both observations on x and y are subject to error.

For the case of specified x, the regression function g(x) may be used as a mathematical function to describe the general relationship existing between the variables. Since the extent of this relationship may vary from trivial to complete, and since the regression function gives no information as to the extent of this relationship, there is need of another

* Associate Professor, U.P. Statistical Center

mathematical function to measure this extent. For this purpose, the concept of variance is extented to the concept of variance about $g(x)$ for every fixed x, which may be defined as the **scedastic function,** $S(x)$. Thus, when $S(x) = 0$, the dependence is complete and the regression function becomes the function which gives the complete mathetical relationship between y and x. As $R(x)$ increases, the extent of the dependence decreases. When $S(x)$ has become for all x as large as variance of y irrespective of the value of x, the dependence has disappeared.

An approximation does not, in general, have the regression properties (i.e. means zero, and minimum variance for each x), for only the regression function has these propertes; hence, it is not possible to determine the coeficient of the arbitrary function by these properties. However it is possible to apply these properties to the overall collection and to require that the means of all deviates from the proposed regression curve be zero and that the overall variance be a minimum. If the residual (or deviate) from the proposed approximate regression function is denoted by ϵ, then the condition above can be written as

$$\sum_{i=1}^{N} \epsilon_i = 0, \quad \text{or} \quad \bar{\epsilon} = 0 \tag{1}$$

$$\sigma_\epsilon^2 = \frac{\sum_{i=1}^{N} (\epsilon_i - \bar{\epsilon})^2}{N} \quad \text{is a minimum.} \tag{2}$$

Since $\sum \epsilon_i^2 = N(\sigma^2 + \bar{\epsilon}) = N\sigma_\epsilon^2$, the two conditions above result from the minimization of $\sum \epsilon^2$. Hence,

a natural criterion to apply in accordance with the concept of regression, is the principle of least squares, and this indicates that the coeficients in the postulated function are to be determined by minmizing the sum of the squares of the residuals.

A somewhat different but more common problem is that of estimating the regression function by sampling. Here, the form of the regression function is assumed, but the coeficients are unknown. Then the regression function has the same form as the known population regression function and the sample coefficient estimates are determined by least squares. Thus, since it is known that the regression function in the multivariate normal is linear, it is proper to take the least squares regression function linear.

Inasmuch as least squares regression theory is usually extended so as to allow the use of observed values of x even if they are in error, it is the purpose of this paper to show the extent to which observational errors of the values of the variables affect the regression estimates. The discussion in this paper will be centered about linear regression; however, the statements that will be made are applicable for the general case of an arbitrary non-linear regression function inasmuch as any arbitrary non-linear function can be transformed to equivalent linear forms.

For compactness of presentation, matrices and matrix derivatives are used in this paper. Hence, familiarity with elementary matrix theory is assumed. However, a short summary of the techniques of matrix differentiation is taken up in the next section.

Matrix Derivatives.

The derivatives of a matrix Y, elements of which are continuous functions of a scalar variable x, with respect to x is the matrix obtained by defferentiating the elements of the matrix Y with respect to the scalar x. If y is a scalar function of a matrix variable X (i.e. a matrix which has as variables for its elements), the derivative of y with respect to X is the matrix obtained by applying a matrix of -dif ferential operators to y.

For example, if $X = [x_{ij}]$, $(i = 1, 2, \ldots, n;$ $j = 1, 2, \ldots, m)$, $\partial y/\partial X = \left[\frac{\partial}{\partial x_{ij}}\right] y = \left[\frac{\partial y}{\partial x_{ij}}\right]$, where $\left[\frac{\partial y}{\partial x_{ij}}\right]$ is a matrix of the same row and column orders as the matrix X.

Suppose Y is a matrix product involving the matrix variable X. The derivative of Y with respect to a particular element x_{ij} may be obtained in which case the matrix $[\partial Y/\partial x_{ij}]$, which has the same row and column orders of Y is formed; or, the derivative of a particular element y_{ik} of Y with respect to the matrix variable X may be obtained in which case the matrix $[\partial y_{ik}/\partial X]$, which has the same row and and column orders of X is formed. For example,

$$Y = MX,$$

where $M = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \\ m_{31} & m_{32} \end{bmatrix}$ and $X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$.

Differentiating Y with respect to each element of the vector X, we have

$$\frac{\partial Y}{\partial x_{11}} = \begin{bmatrix} m_{11} & 0 \\ m_{21} & 0 \\ m_{31} & 0 \end{bmatrix} ; \quad \frac{\partial Y}{\partial x_{12}} = \begin{bmatrix} 0 & m_{11} \\ 0 & m_{21} \\ 0 & m_{31} \end{bmatrix}$$

$$\frac{\partial Y}{\partial x_{21}} = \begin{bmatrix} m_{12} & 0 \\ m_{22} & 0 \\ m_{32} & 0 \end{bmatrix} ; \quad \frac{\partial Y}{\partial x_{22}} = \begin{bmatrix} 0 & m_{12} \\ 0 & m_{22} \\ 0 & m_{32} \end{bmatrix}$$

Dwyer and Macphail[1/] combine these four equations into the single equation

$$\frac{\partial Y}{\partial x_{ij}} = M_{32} \, J_{ij} \, ,$$

where J_{ij} is a matrix of the same order as X and having unity in the ijth position and zero elsewhere. (Matrices of the type of J_{ij} are called "basis" matrices.) The derivative of Y with respect to each element of the variable X is a Type J derivative.

Similarly, the derivatives of the elements of Y with respect to the matrix variable X are

---

1/ "Symbolic Matrix Derivatives" by P.S. Dwyer and M.S. Macphail. *The Annals of Mathematical Statistics*, Vol. XIX, No. 4, Dec. 1948.

$$\frac{\partial y_{11}}{\partial X} = \begin{bmatrix} m_{11} & 0 \\ m_{12} & 0 \end{bmatrix}; \quad \frac{\partial y_{12}}{\partial X} = \begin{bmatrix} 0 & m_{11} \\ 0 & m_{12} \end{bmatrix}$$

$$\frac{\partial y_{21}}{\partial X} = \begin{bmatrix} m_{21} & 0 \\ m_{22} & 0 \end{bmatrix}; \quad \frac{\partial y_{22}}{\partial X} = \begin{bmatrix} 0 & m_{21} \\ 0 & m_{22} \end{bmatrix}$$

$$\frac{\partial y_{31}}{\partial X} = \begin{bmatrix} m_{31} & 0 \\ m_{32} & 0 \end{bmatrix}; \quad \frac{\partial y_{32}}{\partial X} = \begin{bmatrix} 0 & m_{31} \\ 0 & m_{32} \end{bmatrix}$$

These equations are combined in a single equation

$$\frac{\partial y_{qr}}{\partial X} = M_{32}^T \, K_{qr} ,$$

where K_{qr} is a "basis" matrix of the same order as the matrix Y. The derivative of element of Y with respeect to the matrix variable X is called a Type K derivative.

If the tranpose $Y^T = X^T M_{32}^T$ is differentiated with respect to a particular element x_{ji} of X^T,

$$\frac{\partial Y^T}{\partial x_{ji}} = J_{ji}^T \, M_{32}^T \qquad (i, j = 1, 2)$$

where J_{ji}^T is a basis matrix of the same row and column orders as X^T. If a particular element y_{rq} of Y^T is differentiated with respect to the matrix variable X,

$$\frac{\partial y_{rq}}{\partial X} = M_{32}^T \, K_{rq}^T ,$$

where K_{rq}^T is a basis matrix of the same row and volume orders as Y^T.

In obtaining $\partial y_{ij}/\partial X$ from $\partial Y/\partial x_{ij}$, Dwyer and Macphail give the following rules1/:

(a) Each J becomes K and each J^T becomes K^T.
(b) The pre (or post) multiplier of J is changed to its tranpose.
(c) The pre (or post) multiplier of J^T is changed to a post (or pre) multiplier of K^T.

The rules for the differentiation of sum, product, and quotient of matrices are analogous to the rules for the differentiation of the sum, product and quotient of scalar functions of scalar variables in differential calculus.

The Effect of Errors on Least Squares Regression.

Let the linear regression function be given by

$$E(Y) = a_0X_0 + a_1X_1 + \ldots + a_sX_s, \quad \text{where } X_0 \equiv 1.$$

(dummy variable)

1/ "Symbolic Matrix Derivatives" by P.S. Dwyer and M.S. Macphail. *The Annals of Mathematical Statistics*, Vol. XIX, No. 4, Dec. 1948.

Then the residuals are

$$\epsilon_i = Y_i^O - (a_0X_0 + a_1X_{i1} + \ldots + a_sX_{is}),$$

where the upperscript "O" denotes observed value.

Let the "observed" residuals be

$$\epsilon_i^O = y_i^O - (a_0X_0^O + a_1X_{i1}^O + \ldots + a_sX_{is}^O),$$
$$i = 1, 2, \ldots, n.$$

In matrix notation, these n equations are written as

$$\epsilon^O = Y^O - X^OA,$$

where ϵ^O is a column vector (n x 1) of residuals; Y^O is a column vector (n x 1) of observed values of the dependent variable Y; X^O is n x (c + 1) matrix of observed values of the independent variables and is assumed to be of rank (s + 1); A is a column vector, (s + 1) x 1, of regression coeficients. It should be noted that all these observations (except the dummy variable X_O,) may be subject to error. (random errors, not systematic errors.)

If the residuals are uncorrelated and assumed to be of equal weight, the principle of least squares requires that

$$S = \epsilon^{OT}\epsilon^O = (Y^O - X^OA)^T\ (Y^O - X^OA)$$

be minimized with respect to each element of A. Thus

$$\partial S/\partial a_k = -J_k^TX^{OT}\ (Y^O - X^OA) - (Y^O - X^OA)^TX^OJ_k,$$

where J_k and J^T are "basis" matrices of the same order as A. By application of Rules (a), (b), (c),

$$\partial S / \partial A = -X^{OT}(Y^O - X^O A)K^T - X^{OT}(Y^O - X^O A)K,$$

where K and K^T are "basis" matrices of the same order as S. Since S is a scalar, $K = K^T = 1$. Hence, setting $\partial S / \partial A = 0$, we have

$$-X^{OT}(Y^O - X^O A) - X^{OT}(Y^O - X^O A) = O,$$

and

$$(X^{OT}X^O)\hat{A} = X^{OT}Y^O$$

$$\hat{A} = (X^{OT}X^O)^{-1} X^{OT}Y^O \qquad \text{(I)}$$

This estimate of A is in agreement with the usual least squares estimates when the observations on $x_1, x_2, \ldots, x_s$ are not subject to error. It can be used somewhat less satisfactorily when the x's are subject to error, and no knowledge of the size of the error is available. If it is known that the errors in the x's are relatively small with respect to the x's it may be quite satisfactory.

The first order error approximation to the error in $\hat{A}$ resulting from the consideration of dX^O in the normal equations is given by

$$\begin{aligned} d\hat{A} = & -(X^{OT}X^O)^{-1}[(dX^O)^T X^O + X^{OT}(dX^O)] (X^{OT}X^O)^{-1}X^O Y^O \\ & + (X^{OT}X^O)^{-1} [(dX^O)^T Y_O + X^{OT} dY^O]. \end{aligned} \qquad \text{(II)}$$

If the observations on the X's are not subject to error,

$$d\hat{A} = (X^T X)^{-1} X^T dY^O. \qquad \text{(III)}$$

But this error $d\hat{A}$ in the estimate $\hat{A}$ may be ignored, as by one of the properties of the least squares regression estimate (i.e. with the observations on the X's not subject to error) is that $E(\hat{A})$ — A, the "true" value.

In cases for which it can not be assumed that the errors dX^O are not trivial, the estimates $\hat{A}$ are not adequate, and even the use of (II) above is not completely satisfactory since the values of dX^O are not commonly known, though bounds for them may be available.

To make clear the developments above, let us take the case when we have only 2 variables X and Y. Let the linear regression function be given by

$$Y \doteq a_O X_O + a_1 X_1, \quad \text{where } X_O = 1 \text{ (dummy variable).}$$

Then the "observed" residuals are

$$\epsilon^O = Y^O - X^O A, \quad \text{where } \epsilon^O \text{ is an } (n \times 1)$$

Or, in matrix notation, we have

$$e_i^O = y_i^O - (a_0 X_0 + a_1 X_{i1}^O), \qquad i = 1, 2, \ldots, n.$$

matrix of residuals; Y^O is (n x 1) matrix of observed values of Y; X^O is (n x 2) matrix of observed values of the independent variable and is of rank 2; A is (2 x 1) matrix of regression coefficients. That is,

$$\epsilon^0 = Y^0 - X^0A; \quad \text{or,}$$

$$\begin{bmatrix} \epsilon_1^0 \\ \epsilon_2^0 \\ \vdots \\ \epsilon_n^0 \end{bmatrix} = \begin{bmatrix} Y_1^0 \\ Y_2^0 \\ \vdots \\ Y_n^0 \end{bmatrix} - \begin{bmatrix} 1 & X_{11}^0 \\ 1 & X_{12}^0 \\ \vdots & \vdots \\ 1 & X_{1n}^0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix}$$

$$\hat{A} = \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} = (X^{0T}X^0)^{-1} X^{0T}Y^0$$

$$= \begin{bmatrix} \dfrac{\sum X_{1i}^{02}}{A_{xx}} & \dfrac{-\sum X_{1i}^0}{A_{xx}} \\ \dfrac{\sum X_{1i}^0}{A_{xx}} & \dfrac{n}{A_{xx}} \end{bmatrix} \begin{bmatrix} \sum Y_i^0 \\ \sum X_{1i}^0 Y_i^0 \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{\sum X^{02}Y^0 - \sum X^0 \sum X^0Y^0}{A_{xx}} \\ \dfrac{n \sum X^0Y^0 - \sum X^0 \sum Y^0}{A_{xx}} \end{bmatrix}$$

where

$$A_{xx} = n \sum X^{02} - (\sum X^0)^2.$$

If the observed values of X are with errors $e_{1x}, e_{2x}, e_{3x}, \ldots, e_{nx}$ and those on Y are $e_{1y}, e_{2y}, e_{3y}, \ldots, e_{ny}$, the errors in the estimate of A are

$$dA = \begin{bmatrix} \dfrac{-2a_1\sum X^0\sum e_{ix}X_i^0 - \sum X^0(\sum Y_i^0 e_{ix} + \sum X_i e_{iy})}{A_{xx}} \\ \\ \dfrac{-2na_1\sum e_{ix}X_i^0 + n\,(\sum Y_i^0\, e_{ix} + \sum X^0\, e_{iy})}{A_{xx}} \end{bmatrix} . \quad \text{(IV)}$$

It should be noted here that the errors on the observed values of X and Y are assumed non-systematic (i.e. $\sum e_{ix} = \sum e_{iy} = 0$).

It can be observed in (IV) above that if the errors are not trivial, the regression obtained is not adequate; its slope a_1 is on the average smaller than the "true" slope.

There has been much concern about these observational errors in the determination of least equares regression estimates when all the variables involved are subject to error, and a number of statisticians have proposed methods of minimizing the effects of these errors in the estimates.