DISCUSSION PAPER NO. 220

THE CONSTRAINED LEAST-SQUARES PARADOX

by

Avinoam Perry

Graduate School of Management
Northwestern University
Evanston, Illinois 60201

May, 1976

THE CONSTRAINED LEAST SQUARES PARADOX

ABSTRACT

In this paper we provide mathematical, numerical, and logical explanations for a phenomenon which may be classified as a paradox. Under this title we discuss the case of restricted least squares estimation. We show that in certain cases the constrained linear regression model exhibits better fit than the unrestricted model. This conclusion is correct despite the fact that the residual (or unexplained) sum of squares becomes larger as a result of introducing constraints.

I. Introduction

The problem of fitting regression coefficients $\hat{b}_i$, $i = 0,1,\ldots,k$, to the model:

$$y_j = b_1 x_{1j} + \ldots + b_k x_{kj} + e_j \qquad \text{where } E(e) = 0 \text{ and } \operatorname{cov}(e_i e_j) = \begin{cases} 0 & i \neq j \\ \sigma^2 & i = j \end{cases} \tag{1}$$

is solved by the unconstrained optimization procedure known as Ordinary Least Squares [1]. In matrix form, the unconstrained least squares problem is presented as

$$\text{minimize } e'e = (y - Xb)'(y - Xb) \tag{2}$$

where e and y are $(N \times 1)$ vectors, X is $(N \times k)$ matrix, and b is $(k \times 1)$ vector The vector $\hat{b}$ which minimizes (2) is given by

$$\hat{b} = (X'X)^{-1}X'y \tag{3}$$

The coefficient of determination R^2 which measures the strength of the relationships between the dependent variable y and the independent variables $x_1 \ldots x_k$ is determined by (4)

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \tag{4}$$

where SSE is the minimum value of the objective function (2) (also referred to as the unexplained or residual sum of squares) and SST is the total sum of squares: $y'y$.

In several cases, due to prior hypotheses concerning the regression coefficients b_i, $i = 1,\ldots,k$. The procedure of fitting the regression equation becomes a problem in constrained optimization

$$\text{minimize} \quad e'e = (y - Xb)'(y - Xb) \tag{5}$$

$$\text{subject to} \quad Ab = c \tag{6}$$

where A is a $(m \times k)$ matrix of full row rank, and c is $(m \times 1)$ vector.

Due to the fact that the function $(y-Xb)'(y-Xb)$ is convex, the optimum solution of (2) is a global minimum and it is always less than or equal to the minimum solution of (5)-(6). Nevertheless, R^2 which is resulted in the constrained least squares fit of (5)-(6) may be larger than the coefficient of determination of the unconstrained least squares model.

In this paper we present mathematical explanations for this paradox and support it by a numerical example and an intuitive justification.

II. The Constrained Least Squares Estimates

Consider the problem

$$\text{minimize } e'e = (y-Xb)'(y-Xb) \tag{5}$$

subject to

$$Ab = c \tag{6}$$

This problem can be solved by defining its equivalent Lagrangean problem and solving the linear system of equations

$$\frac{\partial L}{\partial b} = 0 \tag{7}$$

$$\frac{\partial L}{\partial (2\lambda)} = 0 \tag{8}$$

where

$$L = (y-Xb)'(y-Xb) + 2\lambda(Ab-c) \tag{9}$$

and 2λ is the vector of Lagrange multipliers associated with the constraint set. The solution of (7)-(8) satisfies the Kuhn-Tucker conditions [2] which are necessary and sufficient for the optimal solution of the convex program (5)-(6). The solution of (7) and (8) yield the following results:

$$\lambda = [A'(X'X)^{-1}A]^{-1}(A'\hat{b} - c) \tag{10}$$

and

$$\tilde{b} = \hat{b} - (X'X)^{-1}A[(X'X)^{-1}A]^{-1}(A'\hat{b} - c) \tag{11}$$

where $\hat{b}$ is the unconstrained least squares estimate and $\tilde{b}$ is the constrained least squares estimate of b. The unexplained sum of squares of the constrained least squares model is

$$(y-X\tilde{b})'(y-X\tilde{b}) = (y-X\hat{b})'(y-X\hat{b}) + (\hat{b}-\tilde{b})'(X'X)(\hat{b}-\tilde{b}) \tag{12}$$

and from (11) we obtain

$$SSE_r = (y-X\tilde{b})'(y-X\tilde{b}) = SSE + (A'\hat{b}-c)'[A'(X'X)^{-1}A]^{-1}(A'\hat{b}-c) = SSE + V \tag{13}$$

where

$$V = (A\hat{b}-c)'[A'(X'X)^{-1}A]^{-1}(A\hat{b}-c)$$

and where SSE is the optimum solution of the unconstrained least squares problem (2). Since $(X'X)$ is positive definite it follows that $SSE_r \geq SSE$.

This result is also supported by plain intuition. Given a minimization program, the optimum value of the unconstrained objective function is always smaller than the value of the same objective if restricted to a given region by a set of active constraints. However, although the unexplained sum of squares is smaller in the unconstrained least squares case, the overall regression goodness of fit (which is measured by R^2) may become larger as a result of incorporating constraints into the program. This phenomenon is explained by resorting to an alternative approach for solving the constrained least squares program.

III. The Reduced Model Approach

The matrix A of constraints coefficients is partitioned such that

$$A = [A_1, A_2]$$

where A_1 is a nonsingular matrix of dimensions m x m and A_2 is a matrix of dimensions m x (k-m). The vector b is also partitioned such that

$$b = [b_1, b_2]$$

and

$$Ab = c \tag{6}$$

becomes

$$A_1 b_1 + A_2 b_2 = c \tag{14}$$

from (14) it follows that

$$b_1 = A_1^{-1}(c - A_2 b_2) \tag{15}$$

and the constrained program (5)-(6) is replaced by an equivalent unconstrained reduced program which incorporates the constraints into the objective function.

This reduced program is the minimization problem

$$\text{Minimize } \tilde{e}'\tilde{e} = [y - X_1 A_1^{-1}(c - A_2 b_2) - X_2 b_2]'[y - X_1 A_1^{-1}(c - A_2 b_2) - X_2 b_2] \tag{16}$$

(where Xb is replaced by $X_1 b_1 + X_2 b_2$). (16) can be written as

$$\text{Minimize } \tilde{e}'\tilde{e} = [y - X_1 A_1^{-1} c - (X_2 - X_1 A_1^{-1} A_2) b_2]'[y - X_1 A_1^{-1} c - (X_2 - X_1 A_1^{-1} A_2) b_2] \tag{17}$$

and upon the following substitutions

$$z = y - X_1 A_1^{-1} c \tag{18}$$

$$W = X_2 - X_1 A_1^{-1} A_2 \tag{19}$$

(17) is reduced to

$$\text{Minimize } \tilde{e}'\tilde{e} = (z - Wb_2)'(z - Wb_2) \tag{20}$$

where the minimum is obtained at the point

$$\tilde{b}_2 = (W'W)^{-1} W' z \tag{21}$$

$$\tilde{b}_1 = A_1^{-1}(c - A_2 \tilde{b}_2) \tag{15}$$

<u>Proposition</u>: The optimum point $[b_1, b_2]$ in [(15), (21)] is equal to the point b in (11).

<u>Proof</u>: The minimization problem (5)-(6) is a convex program. Therefore, the solution (11) of (5)-(6) is a global optimum. Since the only constraints restricting b are given in (6), the reduced problem (17) is equivalent to (5)-(6) and the only optimal solution of the convex unconstrained reduced problem [(15), (21)] is the same global optimum obtained by equating the gradient of the Lagrangean (9) to zero.

<u>Corollary</u>: Given that $\tilde{b}$ in [(15), (21)] is equal to the expression given in (11), it follows that SSE_r is equal to $\tilde{e}'\tilde{e}$.

The coefficient of determination of the reduced model R_r^2 can be expressed as:

$$R_r^2 = 1 - \frac{\tilde{e}'\tilde{e}}{z'z} = 1 - \frac{\tilde{e}'\tilde{e}}{(y-X_1A_1^{-1}c)'(y-X_1A_1^{-1}c)} \tag{22}$$

where SST_r in (22) is expressed as:

$$\begin{aligned} SST_r &= z'z = (y-X_1A^{-1}c)'(y-X_1A_1^{-1}c) \\ &= SST - 2c'A_1'^{-1}X_1'Y + c'A_1'^{-1}(X_1'X_1)A_1^{-1}c = SST + Q \end{aligned} \tag{23}$$

where

$$Q = c'A_1'^{-1}(X_1'X_1)A_1^{-1}c - 2c'A_1'^{-1}X_1'Y \tag{24}$$

It follows that

$$R_r^2 = 1 - \frac{SSE + V}{SST + Q} \tag{25}$$

is not necessarily smaller than

$$R^2 = 1 - \frac{SSE}{SST} \tag{4}$$

$R_r^2 \leq R^2$ when Q = 0, and Q is equal to zero if the constraint set (6) becomes

$$Ab = 0 \tag{26}$$

Under this condition the difference SSR - SSR_r can be viewed as a reduction in the explained sum of squares due to the additional constraints.

IV. Numerical Example

For the following data

Y	x_1	X_2	X_3
8	2.0	4.0	3.0
10	3.0	3.4	3.1
9	2.4	3.5	3.2
5	1.0	10.1	5.6
12	6.0	4.7	5.8
6	3.0	6.0	4.2
7	3.2	6.1	4.8
13	5.1	4.2	5.0
3	1.1	9.6	4.0

The unconstrained least squares estimates of the regression coefficients of the model

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e = Xb + e \tag{27}$$

are

$$b = (X'X)^{-1}X'y = \begin{bmatrix} 8.02 \\ 0.20 \\ -1.22 \\ 1.51 \end{bmatrix} \tag{28}$$

The coefficient of determination of (27) is given in (29)

$$R^2 = 0.899 \tag{29}$$

and the residual sum of squares SSE is given in (30)

$$SSE = 8.585 \tag{30}$$

When restricting the estimated coefficients by imposing the constraint

$$b_3 - b_1 = 6 \tag{31}$$

The original model is reduced to the following

$$y = b_0 + (b_3 - 6)x_1 + b_2x_2 + b_3x_3 + e \tag{32}$$

which then becomes

$$y + 6x_1 = b_0 + b_2x_2 + b_3(x_1 + x_3) + e \tag{33}$$

and the constrained least squares coefficients become

$$\begin{aligned} \tilde{b}_1 &= -1.702 \\ \tilde{b}_2 &= -2.464 \\ \tilde{b}_3 &= 4.298 \end{aligned} \tag{34}$$

The residual sum of squares SSE_r is larger in the constrained model

$$SSE_r = 18.619 \tag{35}$$

But, nevertheless, the coefficient of determination R_r^2 is larger too.

$$R_r^2 = 0.986 \tag{36}$$

V. Concluding Remarks

Although we provided mathematical and numerical proofs for the constrained least squares paradox, we have not yet given an intuitive justification for the phenomenon reported in this paper. We are about to do just that. The residual sum of squares becomes larger as a result of introducing constraints because the minimum of a given unconstrained convex objective function is always at least as small as the minimum of that objective function subject to constraints. On the other hand, the coefficient of determination R^2 may become larger due to the fact that the constrained least squares program is equivalent to a reduced unconstrained least squares problem. The dependent variable of the reduced model is equal to the original dependent variable y plus a linear combination of the independent variables. It is always possible to come up with such linear combinations of independent variables which correlates highly with a linear combination of the same independent variables. If the total sum of squares of the dependent variable of the reduced model is dominated by the variance of the linear combination of the independent variables added to (or subtracted from) it, then the constrained regression model will exhibit a better fit than the unconstrained regression model.

REFERENCES

[1] Draper, N.R. and H. Smith (1966), "Applied Regression Analysis", John Wiley & Sons.

[2] Kuhn, H.W. and A.W. Tucker (1951), "Nonlinear Programming." Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, ed., J. Neyman. University of California Press, Berkeley and Los Angeles, California, 481-492.