THE EFFECT OF INADEQUATE MODELS IN SURFACE FITTING

Prepared Under Office of Ordnance Research

Contract No. DA-36-034-ORD-1177 (RD)

by

G. E. P. Box
J. S. Hunter
R. J. Hader

Institute of Statistics
Mimeo Series No. 91
January, 1954

TECHNICAL REPORT NO. 5

THE EFFECT OF INADEQUATE MODELS IN SURFACE FITTING

Prepared Under Contract No. DA-36-034-ORD-1177 (RD)
(Experimental Designs for Industrial Research)

Ordnance Project No. TB2-0001 (832)
Dept. of Army Project No. 599-01-004

Philadelphia Ordnance District
Department of the Army, Department of Defense
with
Institute of Statistics
North Carolina State College of
The University of North Carolina
Raleigh, North Carolina

Technical Supervisor
Ballistics Research Laboratories
Aberdeen Proving Ground
Aberdeen, Maryland

G. E. P. Box
J. S. Hunter
R. J. Hader
Authors of Report

1. INTRODUCTION

In earlier work on statistical problems connected with surface fitting [1], [2], [4] some of the consequences of "lack of fit" were considered briefly. This report is intended to supplement the original discussion and particularly to point out the extremely interesting and fundamental role of the alias matrix. Certain relations derived previously are herein presented in a manner which more clearly reveals their essential nature.

In order to make this report reasonably self-contained we review some important aspects of surface fitting by the method of least squares. Suppose experimental values of a response variable, y, are available for N selected combinations of independent variables $x_1, x_2, \ldots, x_k$. The experimenter postulates that the surface can be adequately described by an equation of the form

$$\eta = \sum_{i=1}^{L} \beta_i x_i \qquad (1)$$

in which the β_i are unknown coefficients and the x_i are the original independent variables and products and powers thereof. Equation (1) may be regarded as the Taylor Series approximation, say to terms of order d, of the true response surface. The object is then to estimate the β_i.

The least squares procedure of estimation is conveniently described in matrix notation. Let Y be the (N x 1) matrix of experimental values of y and η the corresponding (N x 1) matrix of "true" values of the response variable, i.e.,

$$E(Y) = \eta \qquad (2)$$

Also assume

$$E(Y - \eta)(Y - \eta)' = I_N \sigma^2 \qquad (3)$$

where I_N is the (N x N) identity matrix. Equation (3) states that the errors in y are uncorrelated and have common variance, σ^2.

Let X be an N x L matrix in which the elements of the i^{th} column are the numerical values taken in the N experiments by the variable x_i of equation (1). In matrix notation we write the postulated model as

$$\eta = X\beta. \tag{4}$$

The least squares estimate of β is then given by the (L x 1) matrix B

$$B = (X'X)^{-1} X'Y \tag{5}$$

and it can be shown that

$$E(B) = \beta \tag{6}$$

and

$$E(B - \beta)(B - \beta)' = (X'X)^{-1}\sigma^2 = C^{-1}\sigma^2 \tag{7}$$

The estimates thus obtained have minimum variance.

An unbiased estimate s^2 of σ^2 is provided by

$$(N - L)s^2 = (Y - XB)'(Y - XB) \tag{8}$$

$$= Y'Y - Y'XB \tag{9}$$

$$= Y'Y - B'X'XB \tag{10}$$

$$= Y' \left[I - X'(X'X)^{-1} X \right] Y \tag{11}$$

Let $\hat{Y}$ be an N x 1 matrix of predicted values, i.e.,

$$\hat{Y} = XB \tag{12}$$

Writing

$$Y = \dot{Y} + (Y - \dot{Y}) \tag{13}$$

and taking advantage of

$$\dot{Y}'(Y - \dot{Y}) = 0 \tag{14}$$

we have

$$Y'Y = \dot{Y}'\dot{Y} + (Y - \dot{Y})'(Y - \dot{Y}) \tag{15}$$

that is, the sum of squares of the observed y's can be broken up into two parts, $\dot{Y}'\dot{Y}$, the sum of squares of the predicted values and $(Y - \dot{Y})'(Y - \dot{Y})$ the sum of squares of the residuals. It is sometimes convenient to set out these quantities, which are often called "sum of squares due to regression" and "sum of squares about regression", in an analysis of variance table

	Sum of Squares	Expected Value of Sum of Squares
Due to Regression	$B'X'XB$	$L\sigma^2 + \beta'X'X\beta$
About Regression	$Y'Y - B'X'XB$	$(N - L)\sigma^2$
Total	$Y'Y$	$N\sigma^2 + \beta'X'X\beta$

Often we wish to estimate not the individual coefficient β but a set of linear functions of the β's, $\theta_1, \theta_2, \ldots, \theta_p$. Let this set be written as

$$\theta = M\beta \tag{16}$$

where M is a (p x L) matrix of known coefficients. Then the linear estimates $t_1, t_2, \ldots, t_p$ of the θ's which severally have smallest possible variances are

supplied by

$$T = MB \tag{17}$$

where B is the vector of least squares estimates and T is the (p x 1) vector of the estimates $t_1, t_2, \ldots, t_p$. This important theorem is often called the Gauss-Markoff theorem, [3].

The variances and covariances of these estimates are provided by the matrix

$$\mathcal{E}(T - \theta)(T - \theta)' = \mathcal{E} M(B - \beta)(B - \beta)'M' = M(X'X)^{-1}M'\sigma^2 \tag{18}$$

2. EFFECT OF LACK OF FIT OF MODEL

In practice it will often happen that the model (1) used by the experimenter is not completely adequate to describe the surface. Variables, other than those taken account of, may, in fact, be exerting influence. In particular, higher order terms of the Taylor Series, may not really be negligible. In order to investigate the consequences of this so-called "lack of fit" we write the true model as

$$\eta = X_1\beta_{1.2} + X_2\beta_{2.1} \tag{19}$$

where X_1 in an (N x L_1) matrix of the values of the L_1 variables actually fitted for, X_2 is an (N x L_2) matrix of the values of the L_2 variables not taken account of, $\beta_{1.2}$ an (L_1 x 1) matrix of partial regression coefficients for the first set of X's and $\beta_{2.1}$ an (L_2 x 1) matrix of partial regression coefficients for the second set of X's. If now the experimenter uses

$$B_1 = (X_1'X_1)^{-1}X_1'Y \tag{20}$$

then

$$E(B_1) = \beta_{1.2} + A\beta_{2.1} \tag{21}$$

where

$$A = (X_1'X_1)^{-1}X_1'X_2 \tag{22}$$

The (L_1 x L_2) matrix A is a matrix of bias coefficients and may be called the "alias" matrix. As is obvious from (22) these coefficients are in fact regression coefficients of X_2 on X_1. The r^{th} column of A contains the L_1 regression coefficients of the r^{th} column vector in X_2 on the L_1 column vectors of X_1.

The estimate s^2 of the residual error variance σ^2 would also be biassed in this situation. The essential nature of the bias term is best understood if we arrange the model (19) by adding $X_1A\beta_{2.1}$ to the first term and subtracting it from the second, whence we obtain

$$\eta = X_1(\beta_{1.2} + A\beta_{2.1}) + (X_2 - X_1A)\beta_{2.1} \tag{23}$$

We notice that the columns of the matrix ($X_2 - X_1A$) are the vectors of residuals of the regressions of X_2 on X_1, so that this matrix may be written as $X_{2.1}$. Equation (19) may finally then be written in the form

$$\eta = X_1\beta_1 + X_{2.1}\beta_{2.1} \tag{24}$$

where

$$\beta_1 = \beta_{1.2} + A\beta_{2.1} \tag{25}$$

is the set of regression coefficients for X_1 ignoring the X_2 variables.

Then

$$\mathcal{E}(N - L_1)s^2 = \mathcal{E}(Y - X_1B_1)'(Y - X_1B_1) = \mathcal{E}\{(Y - \eta) - X_1(B_1 - \beta_1) + X_{2.1}\beta_{2.1}\}' \{(Y - \eta) - X_1(B_1 - \beta_1) + X_{2.1}\beta_{2.1}\} \quad (26)$$

$$= \mathcal{E}\{(Y - \eta) - X_1(B_1 - \beta_1)\}' \{(Y - \eta) - X_1(B_1 - \beta_1)\} - \beta'_{2.1}X'_{2.1}\mathcal{E}\{(Y - \eta) - X_1(B_1 - \beta_1)\} - \mathcal{E}\{(Y - \eta) - X_1(B_1 - \beta_1)\}X_{2.1}\beta_{2.1} + \beta'_{2.1}X'_{2.1}X_{2.1}\beta_{2.1} \quad (27)$$

The second and third terms on the right hand side of (27) are clearly zero, and

$$\mathcal{E}(N - L_1)A^2 = (N - L_1)\sigma^2 + \beta'_{2.1}C_{2.1}\beta_{2.1} \quad (28)$$

where

$$C_{2.1} = X'_{2.1}X_{2.1} = (X_2 - X_1A)'(X_2 - X_1A) \quad (29)$$

$$= X'_2X_2 - A'X'_1X_1A \quad (30)$$

$$= X'_2\{I_N - X_1(X'_1X_1)X_1\}X_2 \quad (31)$$

We see that the bias in s^2 is essentially non-negative since the matrix $C_{2.1}$ is non-negative. Also the elements of $C_{2.1}$ are the sums of squares and products of residuals of the regressions of the vectors of X_2 on those of X_1 and become large therefore when the vectors X_2 have only small inner products with the vectors in X_1. The analysis of variance is now

	Sums of Squares	Expected Value of Sums of Squares
Due to regression	$B'_1X'_1X_1B_1$	$L_1\sigma^2+(\beta_{1.2}+A\beta_{2.1})'X'_1X_1(\beta_{1.2}+A\beta_{2.1})$
About regression	$Y'Y-B'_1X'_1X_1B$	$(N-L_1)\sigma^2+\beta'_{2.1}(X_2-X_1A)'(X_2-X_1A)\beta_{2.1}$
Total	$Y'Y$	$N\sigma^2+\left[\beta'_{1.2}\,\middle|\,\beta'_{2.1}\right]\left[\begin{array}{c}X_1\\\hline X_2\end{array}\right]\left[X'_1\,\middle|\,X'_2\right]\left[\begin{array}{c}\beta_{1.2}\\\hline \beta_{2.1}\end{array}\right]$

From the point of view of the experimenter who is attempting to estimate the coefficients of the X_1 terms only and who believed the model of equation (1) to be adequate, the total sum of squares is influenced by terms involving $\beta_{2,1}$. This additional sum of squares is divided between the components "due to regression" and "about regression", the distribution of the bias between these two quantities being determined by the design used. In no case can the bias in either sum of squares be negative. It is possible however for the whole of the bias to be concentrated in one of the components leaving the other unbiassed.

A test frequently used for adequacy of fit of the mathematical model involves the comparison of the "about regression" sum of squares with an independent estimate of error variance obtained, for example, by replication. Although the incompatibility of the two estimates due to an oversized "about regression" sum of squares will certainly indicate lack of fit, it is now clear that when two such estimates <u>are</u> compatable it would be dangerous to assume that the fit of the equation is necessarily adequate. Quite apart from sampling variation, the extent to which the bias effects will appear in the "about regression" sum of squares will depend on the type of design chosen. In selecting designs and performing the subsequent analysis it is important to bear in mind the alias characteristics of the design. In particular, the matrix A and also the alias pattern to be expected in the analysis of variance should be examined.

For example, suppose we wish to fit the three dimensional planar surface

$$\eta = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \qquad (32)$$

to a set of eight observations. The eight levels of (x_1, x_2, x_3) employed would determine eight points in the three dimensional space of the variables x_1, x_2, x_3 which we shall call the experimental design D. One arrangement D_1 which would allow

the efficient estimation of the β's would be a set of points arranged at the vertices of a cube that is, a 2^3 factorial. An alternative could be based on a selection of four points chosen from the cube which form a tetrahedron. This is a half replicate of the 2^3 design in which the defining contrast is the three factor interaction. Imagine then an alternative design of eight points in which two such tetrahedra where used, one being obtained from the other by a simple translation (so that corresponding to a point x_1, x_2, x_3 in the first tetrahedra there was a second point $x_1 + a, x_2 + b, x_3 + c$.)

D_1 D_2

Suppose now that, in fact, within the region considered, the surface was not planar, that is to say not expressible in terms of equation (32) but instead a second degree equation was required.

$$\eta = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \tag{33}$$

Using equation (22) we would have the following alias patterns for the two designs

D_1	D_2
$b_0 \rightarrow \beta_0 + \beta_{11} + \beta_{22} + \beta_{33}$	$b_0 \rightarrow \beta_0 + \beta_{11} + \beta_{22} + \beta_{33}$
$b_1 \rightarrow \beta_1$	$b_1 \rightarrow \beta_1 + \beta_{23}$
$b_2 \rightarrow \beta_2$	$b_2 \rightarrow \beta_2 + \beta_{13}$
$b_3 \rightarrow \beta_3$	$b_3 \rightarrow \beta_3 + \beta_{12}$

where the arrow notation indicates that the quantity of the left is an unbiassed estimate of the quantity on the right.

The expected values of the sums of squares in the analysis of variance would be as follows:

Expected Values of Sums of Squares

D_1

Due to regression $4\sigma^2 + 8\left[(\beta_0 + \beta_{11} + \beta_{22} + \beta_{33})^2 + \beta_1^2 + \beta_2^2 + \beta_3^2\right]$

About regression $4\sigma^2 + 8\left[\beta_{12}^2 + \beta_{13}^2 + \beta_{23}^2\right]$

D_2

Due to regression $4\sigma^2 + 8\left[(\beta_0 + \beta_{11} + \beta_{22} + \beta_{33})^2 + (\beta_1 + \beta_{23})^2 + (\beta_2 + \beta_{13})^2 + (\beta_3 + \beta_{12})^2\right]$

About regression $4\sigma^2$

With the first design the "due to regression" sum of squares is inflated only by the biassing of β_0 with β_{11}, β_{22}, β_{33}. If, as would be usual, the mean where eliminated the "due to regression" sum of squares would be unbiassed. However, the "about regression" sum of squares would be biassed by all three interaction terms. We see in contrast that using the second design, the sum of squares "due to regression" is biassed by all the second order terms, but now the "about regression" sum of squares is unbiassed. This illustrates two limiting cases which could, of course, be readily treated by more elementary means. They serve to show, however, the care that is required in interpreting lack of fit procedures.

In particular, where, as is usual in multiple regression problems, an equation is fitted to data in which the variables x_1, x_2, x_3 have not been controlled but merely observed, correlation between the x's may lead to pathological situations.

3. ADDITION OF EXTRA CONSTANTS

A problem which sometimes faces the experimenter is that of adding further constants to the model. The first model assumed might be

$$\eta = X_1\beta_1 \tag{34}$$

and the experimenter will have calculated the estimates B_1, the precision matrix $C_1^{-1} = (X_1'X_1)^{-1}$ and the estimate of error s^2 in the manner already described. It may now appear (for example as the result of an analysis of variance such as that described above) that the model is inadequate and that a more elaborate model

$$\eta = X_1\beta_{1.2} + X_2\beta_{2.1} \tag{35}$$

should be assumed. The new model could be fitted of course by calculating ab initio the estimates $B_{1.2}$ and $B_{2.1}$ of $\beta_{1.2}$ and $\beta_{2.1}$. During the calculation the new precision matrix

$$C_{1+2}^{-1} = \left[\begin{array}{c|c} X_1'X_1 & X_1'X_2 \\ \hline X_2'X_1 & X_2'X_2 \end{array}\right] \tag{36}$$

would normally be found and using these quantities the new estimate of error s_{1+2}^2 could also be obtained. However, it would be preferable if possible to use a method in which the labor of obtaining the first estimates could be utilized. As we have seen in section 2, above, we may re-write the model in the form

$$\eta = X_1(\beta_{1.2} + A\beta_{2.1}) + (X_2 - X_1A)\beta_{2.1} \tag{37}$$

that is

$$\eta = X_1\beta_1 + X_{2.1}\beta_{2.1} \tag{38}$$

where

$$\beta_1 = \beta_{1.2} + A\beta_{2.1} \quad . \tag{39}$$

Now

$$X_1'X_{2.1} = X_1'X_2 - X_1'X_1(X_1'X)^{-1}X_1X_2 = 0 \tag{40}$$

That is, the vectors of $X_{2.1}$ are orthogonal to the vectors of X_1.

Whence

$$B_{2.1} = (X_{2.1}'X_{2.1})^{-1}X_{2.1}Y \tag{41}$$

and using (17) with (39)

$$B_{1.2} = B_1 - AB_{2.1} \tag{42}$$

This provides the desired procedure which requires only the inversion of an $(L_2 \times L_2)$ matrix $C_{2.1} = [X_{2.1}'X_{2.1}]$ and utilizes the already inverted C_1^{-1}. The precision matrix C_{1+2}^{-1} is readily found by writing the modified model (37) in the form

$$\eta = \left[\begin{array}{c|c} X_1 & X_2 \end{array}\right]\left[\begin{array}{c} \beta_{1.2} \\ \hline \beta_{2.1} \end{array}\right] = \left[\begin{array}{c|c} X_1' & X_2 \end{array}\right]\left[\begin{array}{c|c} I & -A \\ \hline 0 & I \end{array}\right]\left[\begin{array}{c|c} I & A \\ \hline 0 & I \end{array}\right]\left[\begin{array}{c} \beta_{1.2} \\ \hline \beta_{2.1} \end{array}\right] \tag{43}$$

thus

$$\left\{\left[\begin{array}{c} X_1 \\ \hline X_2 \end{array}\right]\left[\begin{array}{c|c} X_1 & X_2 \end{array}\right]\right\}^{-1} = \left\{\left[\begin{array}{c|c} I & 0 \\ \hline A^{-1} & I \end{array}\right]\left(\left[\begin{array}{c} X_1 \\ \hline X_2 \end{array}\right]\left[\begin{array}{c|c} X_1 & X_2 \end{array}\right]\right)\left[\begin{array}{c|c} I & A \\ \hline 0 & I \end{array}\right]\right\}^{-1} \tag{44}$$

and equals

$$\left[\begin{array}{c|c} I & -A \\ \hline 0 & I \end{array}\right]\left[\begin{array}{c|c} C_1^{-1} & 0 \\ \hline 0 & C_{2.1}^{-1} \end{array}\right]\left[\begin{array}{c|c} I & 0 \\ \hline -A' & I \end{array}\right] \tag{45}$$

where

$$C_{2.1} = X'_{2.1}X_{2.1} = C_2 - AC_1A \quad . \tag{46}$$

Consequently

$$C_{1+2} = \begin{matrix} C_1^{-1} + AC_{2.1}^{-1}A & -AC_{2.1}^{-1} \\ -A'C_{2.1}^{-1} & C_{2.1}^{-1} \end{matrix} \tag{47}$$

The same result may of course be obtained by the direct inversion of $C_{1.2}$. The above method of derivation is preferred since it indicates more clearly the essential nature of the procedure carred out.

Finall because $X_{2.1}$ and X_1 are orthogonal we have

$$(N - L_1 - L_2)s^2_{1+2} = (N - L_1)s^2 - B'_{2.1}C_{2.1}B_{2.1} \tag{48}$$

REFERENCES

(1) Box, G. E. P. and Wilson, K. B., "On the Experimental Attainment of Optimum Conditions", Journal of the Royal Statistical Society, Series B, XIII, 1, p. 1, 1951.

(2) Box, G. E. P., "Multi-Factor Designs of First Order", Biometrika 39:49, 1952.

(3) Plackett, R. L., "Some Theorems on Least Squares", Biometrika 36:458, 1949.

(4) Box, G. E. P., Hader, R. J., Hunter, J. S., "Experimental Designs for Multi-Factor Experiments: Preliminary Report", Prepared under Contract DA-36-034-ORD-1177 (RD) for the Office of Ordnance Research by the Institute of Statistics, N. C. State College, June 1953.

DISTRIBUTION LIST

Agency	No. of Technical Copies
Office of Ordnance Research Box CM Duke Station Durham, North Carolina	10
Office, Chief of Ordnance Washington 25, D. C. ATTN: ORDTB	1
Director National Bureau of Standards Washington 25, D. C. ATTN: Statistical Engineering Laboratory	1
District Chief, Phila. Ordnance District 1500 Chestnut Street Philadelphia 2, Pa. Attn: Chief, Artillery-Small	2
Commanding General Aberdeen Proving Ground, Maryland ATTN: BRL	2

Agency	No. of Technical Copies
Commanding Officer Frankford Arsenal Bridesburg Station Philadelphia 37, Pa.	1
Commanding Officer Picantinny Arsenal Dover, New Jersey	1
Commanding Officer Redstone Arsenal Huntsville, Alabama	1
Commanding Officer Rock Island Arsenal Rock Island, Illinois	1
Commanding Officer Watertown Arsenal Watertown 72, Mass.	1
Chief, Bureau of Ordnance (AD3) Department of the Navy Washington 25, D. C.	1
Commander U. S. Naval Proving Ground Dahlgren, Virginia	1
Director Applied Physics Laboratory Johns Hopkins University 8621 Georgia Avenue Silver Spring 19, Maryland	1
Corona Laboratories National Bureau of Standards Corona, California	1
U. S. Naval Ordnance White Oak, Silver Spring 19, Md. ATTN: Library Division	1
The Director Naval Research Laboratory Washington 25, D. C. ATTN: Code 2021	1

Agency	No. of Technical Copies
Commander, U. S. Naval Ordnance Test Station, Inyokern China Lake, California ATTN: Technical Library	1
Commanding General Air University Maxwell Air Force Base, Ala. ATTN: Air University Library	1
Commanding General Air Research and Development Command Baltimore, Maryland ATTN: RDR	1
Commanding General Air Research and Development Command Baltimore, Maryland ATTN: RDD	1
Commanding General Air Material Wright-Patterson Air Force Base Dayton 2, Ohio ATTN: Flight Research Lab. F. N. Bubb, Chief Scientist	1
NAC for Aeronautics 1724 F Street, Northwest, Washington 25, D. C. Attn: Mr. E. B. Jackson, Chief Office of Aeronautical Intelligence	1
U. S. Atomic Energy Commission Document Library 19th and Constitution Avenue Washington 25, D. C. Attn: Director, Division of Research	1
Technical Information Service P. O. Box 62 Oak Ridge, Tennessee Attn: Reference Branch	1
Commanding General Research and Engineering Command Army Chemical Center, Maryland	1

Agency	No. of Technical Copies
Commanding Officer Signal Corps Engineering Laboratory Fort Monmouth, New Jersey Attn: Director of Research	1
Commanding Officer Engineer Research and Development Laboratories Fort Belvoir, Virginia	1
Scientific Information Section Research Branch Research and Development Division Office, Assistant Chief of Staff, G-4 Department of the Army Washington 25, D. C	1
Director National Bureau of Standards Washington 25, D. C.	1
Jet Propulsion Laboratory California Institute of Technology 4800 Oak Grove Drive Pasandena 3, California	1
Chief, Ordnance Development Division National Bureau of Standards Washington 25, D. C.	1
Commanding General Air Research and Development Command PO Box 1395 Baltimore 3, Maryland ATTN: Office of Scientific Research (RDR)	1
Commanding General Air Research and Development Command PO Box 1395 Baltimore 3, Maryland ATTN: RDD	1
Document Service Center U. B. Building Dayton 2, Ohio Attn: DSC-SD	1
Chief of Naval Research c/o Reference Department Technical Information Division Library of Congress Washington 25, D. C.	3