ADDENDUM TO BU-1475-M: "HEIRARCHICAL REGRESSION"

by

Walter T. Federer
Cornell University

ABSTRACT

A discussion of using centered independent variates X_i versus using non-centered values of X_i is presented. For the centered independent variates, the sums of squares and cross-products used to obtain solutions for the regression coefficients are corrected for their means. For non-centered values of the X_i, the uncorrected sums of squares and cross-products are used. Despite the mathematical and perhaps computational convenience attributed to the latter, it is recommended that it not be used for most practical situations. The reason is difficulty in interpretation of the resulting regression coefficients. Also, for exploratory model selection with trend analysis of spatial variation, it is desired to use as few regressions as possible to explain the variation.

Key words: Polynomial regression, centered and non-centered independent variables, exploratory model selection.

BU-1493-M July, 2000
In the Technical Report Series of the Department of Biometrics, Cornell University, Ithaca, New York 14853

INTRODUCTION

In the Technical Report BU-1475-M, Statement 6 reads:

There is a "heirarchical principle" that needs to be adhered to.

Then an example is presented to refute this statement. The example of (Y,X) pairs is
(1,1), (3,2), (5,3), (3,4), (1,5), (3,6), (5,7), (3,8), (1,9).
It is stated that the regression equation $Y = a + bX$ would fit the first three observations, that the regression equation $Y = c + dX^2$ would fit the first five observations, and that the

regression equation $Y = e + fX^4$ would fit the nine pairs of observations. This statement is not correct as stated and will be clarified below.

The so-called hierarchical principle may be stated as

Denote a p^{th} order regression as $\Sigma\beta_i X_i$ ($i = 0, 1, ...,p$). As a general rule, if the p^{th} order regression is significant, then all $\beta_i X_i$ terms when $i < p$ must remain in the regression equation as well.

There are several forms of a regression equation and the different forms determine which regression coefficients are retained in the regression equation. The question arises as to whether one should use centered X variable values or not. In order to be making inferences in the neighborhood of the values of X in the study, one should be using centered X variable values. Considering only the mathematical aspects, it is inviting to use non-centered X variable values. However, the statistical and practical aspects would appear to dictate that centered Y's and X's should be used. To illustrate, consider the two linear regression equations

$$Y = a + bX \text{ and } Y = \bar{y} + b(X - \bar{x}).$$

When a is zero, the solution for b in the first equation is $\Sigma X_i Y_i / \Sigma X_i^2$. In the second equation the solution for b is $\Sigma(X_i - \bar{x})(Y_i - \bar{y}) / \Sigma(X_i - \bar{x})^2$. For most situations, it is desirable to fit the regression through the point $(\bar{x}, \bar{y})$ rather than through the intercept a. Working in the center of one's data set would seem to be a reasonable thing to do. It is sometimes stated that teachers of introductory statistics courses are using the first form without explaining the consequences. This is an undesirable trend if true.

To show the effects of using different regression equations on the values of the regression coefficient, some examples are presented. First, however, some notation is needed to simplify the presentation. Let Y_i be the value of the i^{th} value of the dependent variable Y, **Y** be the column vector of n observations, X_i be the non-centered value of the i^{th} independent variable X, **X** be the n by b design matrix for the regression coefficients, **B** be the column vector of regression coefficients for the regression equation, **X B = Y** be the regression equation to be solved, x be the centered value of X, b_{YX} be the linear regression coefficient of Y on X, b_{yx} be the regression coefficient of centered Y (i.e., y) on centered X (i.e., x), $b_{y2.1}$ be the regression coefficient of the X^2 or x^2 variable with the residuals of the regression of Y on X or x (that is, this is the effect on the squared variable independent of the linear term), the deviation of i^{th} centered value from i^{th} mean be $x_i = x^i - \Sigma x^i / n$, and let four different models be represented by the equations:

equation 1 be $Y = b_0 + b_1X + b_2X^2 + b_3X^3 ...,$
equation 2 be $Y = c_0 + c_1x + c_2x^2 + c_3x^3 ...,$
equation 3 be $Y = \bar{y} + b_{y1}x + b_{y2.1}x_{2'} + b_{y3.12}x_{3'} + ...,$ ($x_{i'}$ are orthogonal polynomial coefficients)
equation 4 be $Y = \bar{y} + b_{y1}x_1 + c_2x_2 + c_3x_3 + ...$

Other forms are possible

EXAMPLE 1. First five pairs of above example are used.

								Regression coefficients		
Y	X	X^2	x	x^2	$x_{2'}$	**X'Y**	**x'y**	Eq. 1	Eq. 2	Eq.3
1	1	1	-2	4	2	13	13	-3.4	4.314	2.6
3	2	4	-1	1	-1	39	0	5.143	0	0
5	3	9	0	0	-2	131	14	-0.857	-0.857	-0.857
3	4	16	1	1	-1	-	-	-	-	-
1	5	25	2	4	2	-	-	-	-	-
Total	13	15	53	0	10	0				

The only consistent solution is for the quadratic regression corrected for linear regression. It is obvious that the linear regression through the point ($\overline{x}$,$\overline{y}$) is zero but Model or Equation 1 indicates a large value for the slope. Note also that the vector **X'Y** is the cross products of the uncentered X and Y values whereas **x'y** is the sums of cross products for the deviations from the means.

In order to observe the procedure for obtaining the above values, an algebraic formulation is useful. For Model 1 and quadratic regression, the following matrix equation is solved to obtain solutions for a, b, and c:

$$\begin{pmatrix} n & \Sigma X_i & \Sigma X_i^2 \\ \Sigma X_i & \Sigma X_i^2 & \Sigma X_i^3 \\ \Sigma X_i^2 & \Sigma X_i^3 & \Sigma X_i^4 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \\ \Sigma X_i^2 Y_i \end{pmatrix}$$

which for the above example is

$$\begin{pmatrix} 5 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 13 \\ 39 \\ 131 \end{pmatrix}$$

resulting in a = -3.4, b = 5.143, and c = -0.857.

For the quadratic form of Model 2, a solution of the following matrix results in solutions for c_0, c_1, and c_2:

$$\begin{pmatrix} n & 0 & \Sigma x_i^2 \\ 0 & \Sigma x_i^2 & 0 \\ \Sigma x_i^2 & 0 & \Sigma x_i^4 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \Sigma Y_i \\ \Sigma x_i Y_i \\ \Sigma x_i^2 Y_i \end{pmatrix}$$

which for the above example is

$$\begin{pmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 13 \\ 0 \\ 14 \end{pmatrix}$$

The solutions are $c_0 = [13 + 12(10)/14]/5 = 4.31428$, $c_1 = 0$, and $c_2 = -12/14 = -0.85714$.

The quadratic form of Model 3 is:

$$\begin{pmatrix} n & 0 & 0 \\ 0 & \Sigma x_i^2 & 0 \\ 0 & 0 & \Sigma x_{2i'}^2 \end{pmatrix} \begin{pmatrix} \overline{y} \\ b_{Y1} \\ b_{Y2.1} \end{pmatrix} = \begin{pmatrix} \Sigma Y_i \\ \Sigma x_i Y_i \\ \Sigma x_{2i'} Y_i \end{pmatrix}$$

For the above example this is

$$\begin{pmatrix} 5 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 14 \end{pmatrix} \begin{pmatrix} \overline{y} \\ b_{Y1} \\ b_{Y2.1} \end{pmatrix} = \begin{pmatrix} 13 \\ 0 \\ -12 \end{pmatrix}$$

For the above $\overline{y} = 13/5 = 2.6$, $b_{Y1} = 0$, and $b_{Y2.1} = -0.8571$.

All three models have the same predicted values. They are 0.886, 3.457. 4.314, 3.457, and 0.886 for the X valuues 1, 2, 3, 4, and 5. The plotted regression curve is the same for all three models.

EXAMPLE 2

Using n = 5 and $Y_i = x_i^2$ and X_i^2, the regression coefficients for Equations 1, 2, and 3 are

$Y = X^2$			$Y = x^2$		
Eq. 1	Eq .2	Eq. 3	Eq. 1	Eq. 2	Eq.3
0	9	11	9	0	2
0	6	6	-6	0	0
1	1	1	1	1	1

Equation 1 gives the correct answer for $Y = X^2$ and Equation 2 gives the correct answer for $Y = x^2$. All three equations give the correct value for the quadratic regression coefficient.

EXAMPLE 3

Consider the following values

X	X^2	X^3	x	x^2	x^3	x_2	x_3	Y	$Y=X^3$	$Y=x^3$
1	1	1	-3	9	-27	-19	-111	1	1	-27
2	4	8	-2	4	-8	-16	-104	3	8	-8
3	9	27	-1	1	-1	-11	-85	5	27	-1
4	16	64	0	0	0	-4	-48	3	64	0
5	25	125	1	1	1	5	13	1	125	1
6	36	216	2	4	8	16	104	3	216	8

7	49	343	3	9	27	29	231	5	343	27

Values for the regression coefficients from various equations are:

Y				$Y=X^3$			$Y=x^3$		
Eq. 1	Eq. 2	Eq. 3	Eq. 4	Eq. 1	Eq. 2	Eq. 3	Eq. 1	Eq. 2	Eq.3
-6.14	3	3.11	3	0	64	117.1	-64	0	0
9.40	-1.27	0 11	9.40	0	48	46.09	48	0	7
-2.67	0	0.21	-2.67	0	12	10.39	-12	0	0
0.22	1.33	0.78	0.22	1	1	-21.71	1	1	6

COMMENT

In exploratory model selection using trend analyses, it is desirable to use as few regressions as possible to explain the spatial variation present in the experiment. This is not possible using non-centered values of the variates. Also, when the dependent variable is equal to some power of the independent variate, this simple relationship will not be indicated using the non-centered values of the independent variate. For most situations, the use of centered independent variates is recommended.

The above examples illustrate that the so-called "hierarichal principle" relies heavily on the words "in general". Perhaps "in general" should be replaced by "sometimes"or perhaps it is not a principle.