AD-A039 547  FLORIDA STATE UNIV TALLAHASSEE DEPT OF STATISTICS  F/G 12/1
CORRELATION IN POLYNOMIAL REGRESSION.(U)
MAR 77 R A BRADLEY, S S SRIVASTAVA  N00014-76-C-0394
UNCLASSIFIED  FSU-STATISTICS-M409  NL

1 OF 1
AD
A039547

END
DATE
FILMED
6-77

12
NW

# CORRELATION IN POLYNOMIAL REGRESSION

Ralph A. Bradley and Sushil S. Srivastava

FSU Technical Report No. M409
ONR Technical Report No. 111

March, 1977

The Florida State University
Department of Statistics
Tallahassee, Florida 32306

This work was supported by the Office of Naval Research under Contract No. N00014-76-C-0394. Reproduction in whole or in part is permitted for any purpose of the United States Government.

See 1473

# CORRELATION IN POLYNOMIAL REGRESSION*

Ralph A. Bradley and Sushil S. Srivastava**

This paper considers the effects of centering on correlation and ill conditioning in polynomial regression. Standard statistical texts on regression give little attention to the computational problems involved. Recent articles on regression, and they are numerous, sometimes hint at the need for centering and scaling of the independent variables but do not give detail. It is hoped that the demonstration given below may dramatize the need for serious attention to centering in polynomial regression. While it is difficult to believe that this demonstration is new, we have not seen it before and believe that it merits attention.

Marquardt and Snee [3], in commenting on common practices, write:

> "One common practice we note is failure to remove nonessential ill conditioning through the use of standardized predictor variables. ... The ill conditioning that results from failure to standardize is all the more insidious because it is not due to any real defect in the data, but only to the arbitrary origins of the scales on which the predictor variables are expressed. ... In a linear model centering removes the correlation between the constant term and all linear terms. In addition, in a quadratic model centering reduces, and in certain situations completely removes, the correlation between linear and quadratic terms."

Brown [1] considers centering and scaling but places emphasis on the use of ridge regression to avoid centering. Hocking [2], in an excellent major review of analysis and selection of variables in linear regression, does not address the centering problem.

*Research supported at the Florida State University by the Army, Navy and Air Force through ONR Contract N00014-76-C-0394. Reproduction in whole or in part is permitted for any purpose of the United States Government.

**Department of Statistics, Florida State University, Tallahassee, Florida.

**Before the easy availability of computers and regression programs, it was standard practice to "code" the variables to simplify the computational problems of regression. Thus the centering problem did not arise. But regression programs seemed to eliminate computational difficulties and their naive uses, particularly by computationally unsophisticated users, seems likely to have led to misuses with failure to center being a common problem.**

**The purpose of this note is to highlight a basic data analytic problem which can arise in fitting a polynomial regression equation to an observation matrix when the independent variables are not centered. It suffices to consider a quadratic model in one independent variable although the same problem occurs with more complex polynomials.**

**Consider the regression model,**

$$E(Y) = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2, \qquad (1)$$

**when $E(Y)$ is the expectation of a random variable $Y$ dependent on selected values of variables $x_0$, $x_1$ and $x_2$. Least squares estimation of the regression coefficients $\beta_0$, $\beta_1$ and $\beta_2$ is desired from $N$ observation vectors, $(y_\alpha, x_{0\alpha}, x_{1\alpha}, x_{2\alpha})$, $\alpha = 1, \ldots, N$. Let $X$ be the $N\times 3$ matrix with $\alpha$-th row $(x_{0\alpha}, x_{1\alpha}, x_{2\alpha})$. It is well known that the estimation depends on $(X'X)^{-1}$ and that $X'X$ will be singular (nearly singular or ill conditioned) if the sample correlation between $x_1$ and $x_2$ is unity(near unity) in absolute value.**

**The quadratic model is**

$$E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 \qquad (2)$$

and it is a special case of (1) with $x_0 \equiv 1$, $x_1 = x$, $x_2 = x^2$. Here it is the sample correlation between $x$ and $x^2$ that is of interest. Computer programs permit direct use of (1) and, in polynomial regression, the programs are used with the specifications needed to obtain (2), often without concern for the centering of x. We investigate the sample correlation between $x$ and $x^2$ when x is not centered.

Let $x_\alpha = x'_\alpha + \bar{x}$, $\bar{x} = \frac{1}{N}\sum_\alpha x_\alpha$; $x'$ is centered at zero, $x$ is centered at $\bar{x}$ which may be any value depending on the choice of origin for x. Let the three pertinent central sample moments of $x$ be $m_k = \sum_\alpha (x'_\alpha)^k$, $k = 2, 3, 4$. It is easy to show that the sample variances of $x$ and $x^2$ are $m_2$ and $(m_4 - m_2^2 + 4m_3\bar{x} + 4m_2\bar{x}^2)$ respectively and that the sample covariance between $x$ and $x^2$ is $m_3 + 2m_2\bar{x}$. The desired sample correlation coefficient between $x$ and $x^2$ is

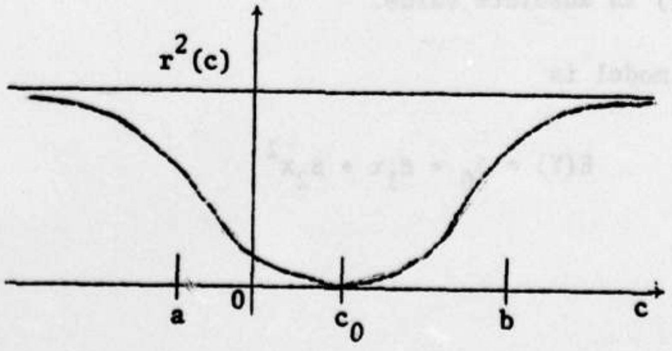
$$r(\bar{x}) = (m_3 + 2m_2\bar{x})/\{m_2(m_4 - m_2^2 + 4m_3\bar{x} + 4m_2\bar{x}^2)\}^{1/2}. \qquad (3)$$

It is clear that $r(\bar{x}) \to \pm 1$ as $\bar{x} \to \pm\infty$. Choice of $\bar{x} = -m_3/2m_2$ gives $r(\bar{x}) = 0$ and is a choice of origin equivalent to the use of orthogonal polynomials, to degree two in this case, in the description of the model (2).

Let $\bar{x} = cs$, $s^2 = m_2$. Then, from (3),

$$r(c) = (m_3 + 2s^3c)/\{s^2(m_4 - s^4 + 4sm_3c + 4s^4c^2\}^{1/2}. \qquad (4)$$

A plot of $r^2(c)$ is as follows:



On the plot, $c_0 = -m_3/2s^3$ and a and b are inflection points where the second derivative of $r^2(c)$ with respect to c changes sign, the values respectively being

$$\frac{-m_3 - \sqrt{(m_4m_2-m_2^3 - m_3^2)/3}}{2s^3} \quad \text{and} \quad \frac{-m_3 + \sqrt{(m_4m_2 - m_2^3 - m_3^2)/3}}{2s^3} .$$

The choice of the N values of x should be open to the experimenter. Thus, if they are chosen in a symmetric fashion defined to mean that $m_3 = 0$ and if x is centered, $c = 0$, $r(0) = 0$, and the assertion of Marquardt and Snee that correlation may be completely removed in certain circumstances is justified.

Suppose that values of x are chosen so that they form a moment pattern, $m_3 = 0$, $m_4 = 3s^4$, like those of a normal variate. Then $r^2(c) = 1/(1 + 1/2c^2)$ and $r^2(c) = .67, .89, .95$ as $c = 1, 2, 3$. Since c is measured in terms of the standard deviation of x, $|r(c)|$ is seen to be close to unity for even relatively small values of c and a choice of origin for x far from $\bar{x}$ in units of standard deviations of x is most unfortunate. More extreme results follow if the values of x are chosen in a uniform pattern over an interval with a moment pattern, $m_3 = 0$, $m_4 = 1.8s^4$, like those of a uniform variate. Then $r^2(c) = 1/(1+1/5c^2)$ and $r^2(c) = .83, .95, .98$ as $c = 1, 2, 3$.

The authors encountered the problem addressed in this paper in some preliminary analysis of a weather modification experiment that involved the fitting of polynomials in two variables with the variables badly centered. The simple analysis given above demonstrates clearly that failure to center the independent variables in polynomial regression can lead to correlations near unity and to ill conditioning. Choices of origins far from centers can be disastrous and it is clearly safest to center the independent variables at their means. The development given is easy and could provide an effective problem assignment in a first course on regression. The development is of sufficient importance to require the introduction of cautions in discussions of regression in texts.

## REFERENCES

[1] Brown, P. J. (1977): Centering and scaling in ridge regression, Technometrics, 19, 35-36.

[2] Hocking, R. R. (1976): The analysis and selection of variables in linear regression, Biometrics, 32, 1-49.

[3] Marquardt, D. W. and Snee, R. D. (1975): Ridge regression in practice, The American Statistician, 29, 3-20.

UNCLASSIFIED (14) FSU-Statistics-M409, TR-111-ONR

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

1. REPORT NUMBER
ONR Report No. 111

2. GOVT. ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and subtitle)
(6) Correlation in Polynomial Regression.

5. TYPE OF REPORT & PERIOD COVERED
(9) Technical Rep[redacted]t.,

6. PERFORMING ORG. REPORT NUMBER
FSU Statistics Report M409

7. AUTHOR(s)
(10) Ralph A./Bradley
Sushil S./Srivastava

8. CONTRACT OR GRANT NUMBER(s)
ONR N00014-76-C-0394 (15)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
The Florida State University
Department of Statistics
Tallahassee, Florida 32306

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
Arlington, Virginia 22217

12. REPORT DATE
(11) Mar[redacted]77

13. NUMBER OF PAGES
5 (12) 8p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS (~~of this~~ report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this report)
Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Polynomial regression, singularities in regression, correlation, ill-conditioned matrices, matrix conditioning, centering in regression.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number).

It has been stated that centering of the independent variable in quadratic regression reduces the correlation between linear and quadratic terms. This in turn reduces or removes ill conditioning of the matrix to be fitted in the use of least squares. It is shown in this paper how the correlation between linear and quadratic terms depends on c, next page

400277 HB

**REPORT DOCUMENTATION PAGE**

**20. ABSTRACT (continued)**

**the departure from centering of the independent variable in terms of a measure of variation in that variable. It is shown that the correlation approaches unity in absolute value as $|c| \to \infty$ and that it is near unity in absolute value for even modest values of $|c|$. It is shown further that with appropriate centering, the correlation may be reduced to zero. The results developed for simple quadratic regression extend easily to polynomial regression with similar effects.**

approaches infinity

the absolute value of c