ideas are extremely difficult to comprehend, and we must not try to present all of them in one course.

Incidentally, by participating in a broad mathematical science program, we can more easily influence the core which was mentioned in point (a) and we can inject some enthusiasm in its teaching. Since Departments of Mathematics are now discovering that mathematics courses are springing up in many other areas in the universities, I suspect that mathematicians will begin to welcome the cooperation of statisticians since most of us have had a great deal of experience with that type of situation. Nevertheless all of us must recognize that there is much criticism of the teaching of applied mathematics and statistics to students in fields that use the mathematical sciences, and a great deal of development is needed in this educational area. Statisticians, if they want to do so, can make a real contribution here because for the most part, we understand the mathematical needs of the social scientist, the business man, the engineer, the biologist, and so on, better than do most of the pure mathematicians.

So while we admit that the core of mathematics provides an excellent foundation for the study of advanced statistics and probability, it does not necessarily supply the motivation. We must somehow demonstrate to the students, who have some mathematical talents but are interested in application rather than abstractions, the comprehensive nature of statistical and probabilistic methods. I would like to suggest the universal requirement of a course in data analysis, but I am afraid that it would lead to disaster since so few persons are capable of teaching it. That is, it is a sad fact that far too few classroom teachers have had an opportunity to acquire extensive experience in applying statistical methods to real data. Nevertheless, statisticians who are capable of teaching data analysis should not only be permitted but urged to do so. For the health of statistics, we need the stimulation afforded by effective exchanges with other areas. These could very well be the necessary ingredient to attract those students who are now abandoning major studies in the mathematical sciences. Simply the realization of the mass of important data which will be collected in the near future convinces me that we must help direct these students to positions which desparately need to be filled by those who have some flair for applying mathematics. Otherwise the resulting analyses will be, for the most part, very inadequate, for we must remember that the data will be analyzed by someone, possibly by persons without the proper training.

I know that in presenting this report I have taken a rather pessimistic view. However, I do not feel that all is lost. On the contrary, I view a highly successful future for statistics, but one that will not materialize without the appropriate effort. Accordingly, I urge all of us, who are in positions of influence in our educational processes, to take some measures that will generate student interest in statistics. Certainly, the same scheme will not work in all situations, but let us at least consider appropriate responses to the situation now at hand and then try to implement as many of these as are reasonably possible. If we will but do this, I am confident that statistics will capture the interest of many capable students who will enter graduate work. Moreover, if the concept of statistics as a discipline which is concerned with developing and evaluating methods to be used to gain insight into the real world is to permeate, success in this effort is required.

A Note on The Variance of A Predicted Response in Regression

ROBERT C. WALLS and DAVID L. WEEKS, University of Arkansas Medical Center and Oklahoma State University

When planning to use a linear regression equation to predict a response we are faced with the problem of selecting an adequate set of independent variables to include in the equation. A reasonable objective is the selection of a set which minimizes the variance plus the squared bias of the predicted response. This idea is used, for example, in an article by Gorman and Toman (1).

In the more recent texts (e.g., Draper and Smith (2), and Williams (3)) the problem of bias has been given deserved attention. However, it is not mentioned explicitly that the addition of a variable to a regression equation can never reduce (and in fact usually increases) the variance of a predicted response. It seems that this is worth mentioning in statistics courses for the sake of the student's understanding, and to serve as an incentive to avoid "overfitting" along with the obvious advantage of having a simpler equation. A convenient way to present the idea follows.

Suppose the vector of observations, Y, has mean EY and covariance matrix:

$$V(Y) = \sigma^2 I \ ,$$

and consider the following two candidates for the form of the expectation of Y in matrix notation (EY need not be given exactly by either):

$$X_1\beta_1 \ ,$$

and

$$X_1\beta_1 + X_2\beta_2 \ .$$

Parameter estimates in these two cases are given by:

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y \ ,$$

and

$$\begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \begin{pmatrix} X_1'Y \\ X_2'Y \end{pmatrix}$$

Letting the vector (U_1', U_2') represent a point in the space of the independent variables, the two estimates of the response, *both of which may be biased*, are given by:

$$y_1 = U_1'\hat{\beta}_1 \ ,$$

and

$$y_2 = U_1'\tilde{\beta}_1 + U_2'\tilde{\beta}_2 \ .$$

Using the well known identities for the inverse of a partitioned matrix we find:

$$\begin{aligned} Cov(y_1,y_2) &= E[(y_1 - Ey_1)(y_2 - Ey_2)] \\ &= \sigma^2 U_1'[(X_1'X_1)^{-1}X_1']\cdot[X_1C_{11} + X_2C_{21}]U_1 \\ &\quad + \sigma^2 U_1'[(X_1'X_1)^{-1}X_1']\cdot[X_1C_{12} + X_2C_{22}]U_2 \\ &= \sigma^2 U_1'[C_{11} - C_{12}C_{22}^{-1}C_{21}]U_1 \\ &\quad + \sigma^2 U_1'[C_{12} - C_{12}]U_2 \\ &= \sigma^2 U_1'(X_1'X_1)^{-1}U_1 \\ &= Var(y_1) \ . \end{aligned}$$

Therefore,

$$\begin{aligned} &E[(y_1 - Ey_1) - (y_2 - Ey_2)]^2 \\ &\qquad = Var(y_1) + Var(y_2) - 2\ Cov(y_1,y_2) \\ &\qquad = Var(y_2) - Var(y_1) \ , \end{aligned}$$

and since this expression is non-negative, we have:

$$Var(y_2) \geq Var(y_1) \ .$$

Equality holds in the above expression only if

$$[U_1'C_{12}C_{22}^{-1} + U_2']C_{22}[U_1'C_{12}C_{22}^{-1} + U_2']' = 0.$$

Since C_{22} is positive definite, this implies only if

$$U_1'C_{12}C_{22}^{-1} + U_2' = \emptyset \ ,$$

or equivalently, only if

$$U_1'(X_1'X_1)^{-1}X_1'X_2 = U_2'$$

will equality hold.

When $X_1'X_2 = \emptyset$, the variances of the two estimates of the response are equal only if $U_2 = \emptyset$; when $X_1'X_2 \neq \emptyset$, equality holds only if the elements of U_2 are particular linear combinations of the elements of U_1.

These results apply to estimated regression coefficients as well as to predicted responses since the variance of a given coefficient corresponds to a particular choice of the vector (U_1', U_2').

An intuitive argument can be given by remembering that an estimator of the response having zero variance would be provided by selecting an arbitrary constant and agreeing to always predict the response to be this value. Although it could be quite *inaccurate*, no other estimator could provide better *precision*. On the other hand, estimating one or more regression coefficients would introduce variability and provide a less precise, but hopefully, more accurate estimator. Taking the definitions of precision and accuracy in this sense, it follows that adding a variable to the equation can never "improve" the precision but only remove possible biases from the various estimates obtained from the regression analysis. Simultaneous *reduction* of both variance and bias may be achieved only by the substitution of a new variable for one already in the equation.

A simple example will illustrate most of the points made. Let $\hat{y}_L$ and $\hat{y}_Q$ be the prediction equations developed from the data (X_i, Y_i) : (1, 5), (2, 7), (3, 7), (4, 10), (5, 16), (6, 20) where

$$\hat{y}_{Li} = \hat{\beta}_1 X_i \ ,$$

and

$$\hat{y}_{Qi} = \tilde{\beta}_1 X_i + \tilde{\beta}_2 X_i^2 \ .$$

For these data we obtain:

$$\hat{y}_{Li} = (91)^{-1}(280\ X_i) \ ,$$

and

$$\hat{y}_{Qi} = (12{,}544)^{-1}(30{,}184\ X_i + 1{,}736\ X_i^2) \ .$$

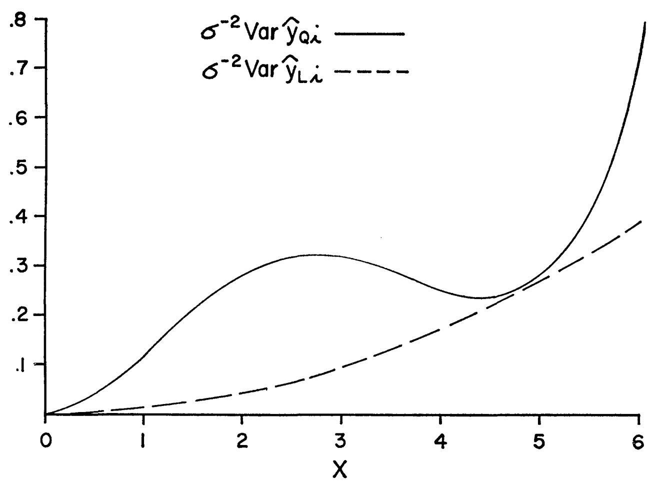
The "standardized" variances of the two predicted responses and their ratio at each observed x-value are given in the following table.

i	σ^{-2}Var $\hat{y}_{Li}$	σ^{-2}Var $\hat{y}_{Qi}$	Var $\hat{y}_{Qi}$/Var $\hat{y}_{Li}$
1	.0110	.1183	10.75
2	.0440	.2790	6.34
3	.0989	.3214	3.25
4	.1758	.2589	1.47
5	.2747	.2790	1.02
6	.3956	.7433	1.88

In the figure below the graph of $\sigma^{-2}Var\ \hat{y}_{Li}$ versus X appears as the dashed line and that of $\sigma^{-2}Var\ \hat{y}_{Qi}$ appears as the solid line.

In this example we see that, over the set of points at which data were taken, use of the second degree term can increase the variance of the predicted response to more than ten times its value when the simpler equation is used. However, this is quite apart from any consideration of which, if either, of the two equations is "correct," i.e., estimates from both equations may be biased in amounts which depend upon the nature of the true model.

Some additional specific examples can be found in a related discussion by R. L. Anderson (4).



REFERENCES

[1]Gorman, J. W. and Toman, R. J.: "Selection of Variables for Fitting Equations to Data.", *Technometrics* 29, 1966, pp. 27–51.

[2] Draper, N. R. and Smith, H.: *Applied Regression Analysis*, John Wiley and Sons, New York, 1966.

[3] Williams, E. J.: *Regression Analysis*, John Wiley and Sons, New York, 1959.

[4] Anderson, R. L., "Some Statistical Problems in the Analysis of Fertilizer Response Data," E. L. Baum et al., editors, *Economic and Technical Analysis of Fertilizer Innovations and Resource Use*, Iowa State College Press, 1957, pp. 187–206.

RECENT PUBLICATIONS

Aitchison, J. *Solving Problems in Statistics*. Oliver & Boyd, Edinburgh and London, 1968, viii, pp. 168.

Ayres, Robert U. *Technological Forecasting*. McGraw-Hill, New York, 1969, pp. 256, $15.00.

Bancroft, T. A. *Topics in Intermediate Statistical Methods*. Volume I. Iowa State University Press, 1968, xiii, pp. 129, $5.90.

Bartlett, M. S. *Biomathematics*. An Inaugural Lecture Delivered Before the University of Oxford on May 28, 1968. The Clarendon Press, Oxford, 1968, pp. 28.

Bell, W. J., and Mather, J. L. *Business Statistics Simplified*. Heinemann, London, 1968, vii, pp. 141.

Bevington, Philip R. *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, New York, 1969, pp. 320, $7.50.

Bjerrum, Chresten A. *Forecast 1968–2000 of Computer Developments and Applications*. Parsons & Williams, Nyropsgade 43, Copenhagen, Denmark, pp. 64, $12.50.

Brown, William, and Palermo, Carmen. *Random Processes, Communications, and Radar*. McGraw-Hill, New York, 1969, pp. 400, $16.00.

Cohen, Lawrence B. *Work Staggering for Traffic Relief*. An Analysis of Manhattan's Central Business District. Frederick A. Praeger, New York, 1968, pp. 670, $18.50.

Eisen, Martin. *Introduction to Mathematical Probability*. Prentice-Hall, Englewood Cliffs, New Jersey, 1969, pp. 496, $12.95.

Ellis, Brian. *Basic Concepts of Measurement*. The University Press, Cambridge, 1968, pp. 220.

Fels, Rendigs, and Hinshaw, C. Elton. *Forecasting and Recognizing Business Cycle Turning Points*. National Bureau of Economic Research, 1968, xvii, pp. 131, $4.50.

Gregory, S. *Statistical Methods and the Geographer*. 2d ed. Longmans, London, 1968, ix, pp. 277.

Hart, P. E., ed. *Studies in Profit, Business Saving and Investment in the United Kingdom, 1920–1962*. Volume II. Allen & Unwin, London, 1968, pp. 283.

Jenkins, Gwilym M., and Watts, Donald G. *Spectral Analysis and Its Applications*. Holden-Day, San Francisco, 1968, xv, pp. 525, $17.00.

Keyfitz, Nathan, and Flieger, Wilhelm. *World Population: An Analysis of Vital Data*. University of Chicago Press, Chicago, 1968, xi, pp. 672.

King, Leslie J. *Statistical Analysis in Geography*. Prentice-Hall, Englewood Cliffs, New Jersey, 1969, pp. 228, $6.95.

Ku, Harry H., ed. *Precision Measurement and Calibration—Statistical Concepts and Procedures*. National Bureau of Standards Special Publication 300, U.S. Government Printing Office, Washington, D.C., 1969, pp. 436, $5.50.

Kyburg, Henry E., Jr. *Probability Theory*. Prentice-Hall, Englewood Cliffs, New Jersey, 1969, $8.95.

Lambe, C. G. *Elements of Statistics*. 2d ed. Longmans, London, 1968, viii, pp. 138.

Lipschutz, Seymour. *Probability*. McGraw-Hill (Schaum's Outline Series), New York, 1969, pp. 161, $3.50.

Lowe, C. W. *Industrial Statistics*. Volume I. Business Books Ltd., London, 1968, xii, pp. 316.

McGilvray, James. *Irish Economic Statistics*. Institute of Public Administration, Dublin, 1968, viii, pp. 180.

McCleod, John, ed. *Simulation*. McGraw-Hill, New York, 1968, xii, pp. 356, $15.00.

Meditch, J. S. *Stochastic Optimal Linear Estimation and Control*. McGraw-Hill, New York, 1969, pp. 384, $15.00.

Mendenhall, William. *Introduction to Linear Models and the Design and Analysis of Experiments*. Wadsworth Publishing Co., Belmont, California, 1968, xiv, pp. 465.

Murdoch, J., and Barnes, J. A. *Statistical Tables for Science and Engineering*. Macmillan, New York, 1968, pp. 32.

Oldham, P. D. *Measurement in Medicine: the Interpretation of Numerical Data*. English Universities Press, London, 1968, viii, pp. 216.

Panico, Joseph A. *Queuing Theory*. Prentice-Hall, Englewood Cliffs, New Jersey, 1969, pp. 224, $7.95.

Schlaifer, Robert. *Analysis of Decisions Under Uncertainty*. McGraw-Hill, New York, 1969, pp. 832, $16.50.

Scoville, James. *The Job Content of the U.S. Economy*. McGraw-Hill, New York, 1969, pp. 144, $6.95.

Spear, Mary Eleanor. *Practical Charting Techniques*. McGraw-Hill, New York, 1969, pp. 352, $10.00.

Szabady, Egon, ed. *World Views of Population Problems*. Akademiai Kiado, Budapest, 1968, pp. 447.

Thomasian, Aram J. *The Structure of Probability Theory with Applications*. McGraw-Hill, New York, 1969, pp. 832, $18.50.

Thornley, Gail, ed. *Critical Path Analysis in Practice: Collected Papers on Project Control*. Tavistock Publications, London, 1968, xii, pp. 152.

Turney, Billy L., and Robb, George P. *Simplified Statistics for Education and Psychology*. International Textbook Co., Scranton, Pennsylvania, 1968, xi, pp. 140.

Wein, Harold H., and Sreedharan, V. P. *The Optimal Staging and Phasing of Multi-Product Capacity*. MSU Studies in Comparative and Technological Planning, Michigan State University, East Lansing, Michigan, 1968, pp. 131, $14.50.

Yeomans, K. A. *Introducing Statistics: Statistics for the Social Scientist*. Volume I, Penguin Books, Harmondsworth, 1968, pp. 258.

Zangwill, Willard I. *Nonlinear Programming: A Unified Approach*. Prentice-Hall, Englewood Cliffs, New Jersey, 1969, pp. 384, $12.50.