5. CONCLUSIONS

Although our presentation is concerned exclusively with the effects of violations of independence assumptions in the one-way model, it is easily adaptable for other designs and regression models when the family $\mathcal{F}$ in (4) is appropriately defined for such models. The pedagogical importance of demonstrating the effects of violations of ANOVA assumptions in the classroom cannot be overstated. We most certainly want to dispel notions that correlation in data does not affect inferences drawn from the F tests of ANOVA. But more important, as teachers our goal should be to balance our students' eagerness to blindly perform F tests with an appreciation for the underlying tenets of those tests. We have found that such "appreciation" is most effectively communicated by discussing the subtleties of the example in Section 4 in the classroom.

[*Received November 1984. Revised December 1986.*]

REFERENCES

International Mathematical and Statistical Libraries (1977), *IMSL Library Reference Manual*, Houston: Author.

SAS Institute, Inc. (1982), *SAS User's Guide*, Cary, NC: Author.

Searle, S. R. (1971), *Linear Models*, New York: John Wiley.

Tiku, M. L. (1965), "Laguerre Series Forms of the Noncentral χ^2 and **F** Distributions," *Biometrika*, 52, 416–427.

Sometimes $R^2 > r^2_{yx_1} + r^2_{yx_2}$
Correlated Variables Are Not Always Redundant

DAVID HAMILTON*

An extreme example of regression on two variables is presented in which there is almost no correlation between y and x_1 and y and x_2, yet the coefficient of determination is 1. This example illustrates the often counter-intuitive nature of multivariate relationships and is also relevant to discussions on multicollinearity and variable selection techniques.

KEY WORDS: Multiple regression; Multicollinearity; Masking variable; Coefficient of determination; Partial correlation; Variable selection.

1. INTRODUCTION

The following statement about multiple regression appears in an introductory textbook (Ott 1984): "If the independent variables are uncorrelated, then

$$R^2 = r^2_{yx_1} + r^2_{yx_2} + \cdots + r^2_{yx_k}. \quad (1)$$

But when the independent variables are themselves correlated, it is difficult to separate R^2 into the predictive contribution of each independent variable. . . . For these situations, about all we can say about the relationship between R^2 and the $r^2_{yx_j}$ is that

$$R^2 \leq r^2_{yx_1} + r^2_{yx_2} + \cdots + r^2_{xy_k}\text{''} \quad (2)$$

(p. 418).

The latter claim reflects the perhaps natural but erroneous belief that correlated explanatory variables contain only redundant information about y. A search of other introductory and more advanced textbooks revealed that several give examples in which the importance of one explanatory variable is reduced by the inclusion of another in the regression equation, with no mention that the contrary can occur. Others contain statements about multicollinearity and variable selection that indicate a lack of understanding about how explanatory variables can act in combination. On the other hand, the *Minitab Handbook* (Ryan, Joiner, and Ryan 1985, pp. 241–245) gives an example of a "suppressor variable" that increases the importance of another variable when it is added to the regression. This example demonstrates that (2) is not always true for $k = 2$ because $R^2 = .473$ and $r^2_{yx_1} + r^2_{yx_2} = .271 + .076 = .348$. Daniel and Wood (1980, pp. 50–53) cautioned against the use of plots of x_i versus y when $k > 1$. They showed examples in which data are generated from an equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (so $R^2 = 1$) yet the separate plots of x_1 versus y and x_2 versus y show little relationship or give misleading information about the values of the parameters. Their first example has $r^2_{yx_1} + r^2_{yx_2} = .682 + .073 = .775$, and their third has $r^2_{yx_1} + r^2_{yx_2} = .441 + .007 = .448$. Kendall and Stuart (1973, p. 331) said that a variable masks the relationship between two others if their partial correlation given the "masking variable" is larger than their simple correlation. They also stated (ex. 27.22, p. 359) that it is possible for the multiple correlation to be 1 even though one of the simple correlations is 0 and the other is arbitrarily close to 0. The relationship between simple, partial, and multiple correlation is discussed in detail for the case $k = 2$. A data set illustrating Kendall and Stuart's extreme case is given.

2. THEORETICAL ASPECTS

An approach taken by several authors is to explain regression on x_1 and x_2 as a sequence of three simple regressions: y on x_1, x_2 on x_1, and the residuals from the first on the

*David Hamilton is Associate Professor, Department of Mathematics, Statistics, and Computing Science, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada.

residuals from the second (with no intercept). Then the overall sum of squares for regression $SSR(x_1, x_2)$ can be partitioned as

$$SSR(x_1, x_2) = SSR(x_1) + SSR(x_2|x_1), \quad (3)$$

where $SSR(x_1)$ is the SSR in the first regression and $SSR(x_2|x_1)$ is the SSR in the third, also called the extra sum of squares explained by x_2 given x_1. Statement (1) is verified using the facts that $R^2 = SSR/TSS$, where TSS is the corrected total sum of squares, and that $SSR(x_2|x_1) = SSR(x_2)$ precisely when x_1 and x_2 are uncorrelated. From (3) it is easy to show that a necessary and sufficient condition for $R^2 > r^2_{yx_1} + r^2_{yx_2}$ is that

$$SSR(x_2|x_1) > SSR(x_2). \quad (4)$$

The partial correlation between y and x_2 given x_1, $r_{yx_2.x_1}$, measures the linear association between the two sets of residuals, and its square is

$$r^2_{yx_2.x_1} = \frac{SSR(x_2|x_1)}{TSS - SSR(x_1)},$$

the proportion of variation explained in the third simple regression. Combining this with (3) gives an expression relating R^2 to one simple and one partial correlation,

$$R^2 = r^2_{yx_1} + r^2_{yx_2.x_1}(1 - r^2_{yx_1}). \quad (5)$$

A correct version of (2) for $k = 2$ is, therefore, $R^2 \leq r^2_{yx_1} + r^2_{yx_2.x_1}$, and the necessary and sufficient condition (4) is

$$r^2_{yx_2.x_1} > r^2_{yx_2}/(1 - r^2_{yx_1}). \quad (6)$$

Since the choice of x_1 as the first explanatory variable is arbitrary, (4) and (6) remain valid conditions when the roles of x_1 and x_2 are interchanged. Kendall and Stuart's definition of a masking variable, that $r^2_{yx_2.x_1} > r^2_{yx_2}$, is a necessary but not sufficient condition for $R^2 > r^2_{yx_1} + r^2_{yx_2}$. If x_1 is a masking variable for x_2, however, x_2 is also a masking variable for x_1.

The partial correlation coefficient can be expressed in terms of the simple correlations between y and x_1 and x_2 and the simple correlation between x_1 and x_2,

$$r_{yx_2.x_1} = (r_{yx_2} - r_{yx_1}r_{x_1x_2})/\sqrt{(1 - r^2_{yx_1})(1 - r^2_{x_1x_2})}. \quad (7)$$

This identity is easily verified by calculating the simple correlation between the residuals $y - \bar{y} - \hat{\alpha}(x_1 - \bar{x}_1)$ and $x_2 - \bar{x}_2 - \hat{\gamma}(x_1 - \bar{x}_1)$, where $\hat{\alpha}$ and $\hat{\gamma}$ are the usual least squares estimates. Substituting (7) in (5) eliminates the partial correlation, giving a relationship between R^2 and the three simple correlations

$$R^2 = (r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1}r_{yx_2}r_{x_1x_2})/(1 - r^2_{x_1x_2}). \quad (8)$$

Similarly, substituting (7) in (6) gives the necessary and sufficient condition for $R^2 > r^2_{yx_1} + r^2_{yx_2}$ in terms of the simple correlation $r_{x_1x_2}$,

$$r_{x_1x_2}\left(r_{x_1x_2} - \frac{2r_{yx_1}r_{yx_2}}{r^2_{xy_1} + r^2_{yx_2}}\right) > 0.$$

The truth of the statement $R^2 > r^2_{yx_1} + r^2_{yx_2}$ is seen to depend directly on the extent of multicollinearity or correlation between the two explanatory variables.

Geometry is a useful tool for explaining multiple regression to more advanced students. This approach has been described by several authors, including Draper and Smith (1981) and Box, Hunter, and Hunter (1978), and in this journal, by Margolis (1979), Herr (1980), and Bryant (1984). Let $Y = (y_1 - \bar{y}, \ldots, y_n - \bar{y})^T$ denote the vector of deviations of the dependent variable from its average, and let X_1 and X_2 contain the corresponding derivations for the explanatory variables. Then the squared length of Y is TSS. Simple regression of y on x_i is achieved by projecting Y on X_i, $SSR(x_i)$ is the squared length of the projected vector $\hat{Y}_i$, and r_{yx_i} is the cosine of the angle between Y and X_i. Regression on both variables is achieved by projecting Y on the plane spanned by both X_1 and X_2, $SSR(x_1, x_2)$ is the squared length of the projected vector $\hat{Y}$, and R^2 is the square of the cosine of the angle between Y and $\hat{Y}$. The projection

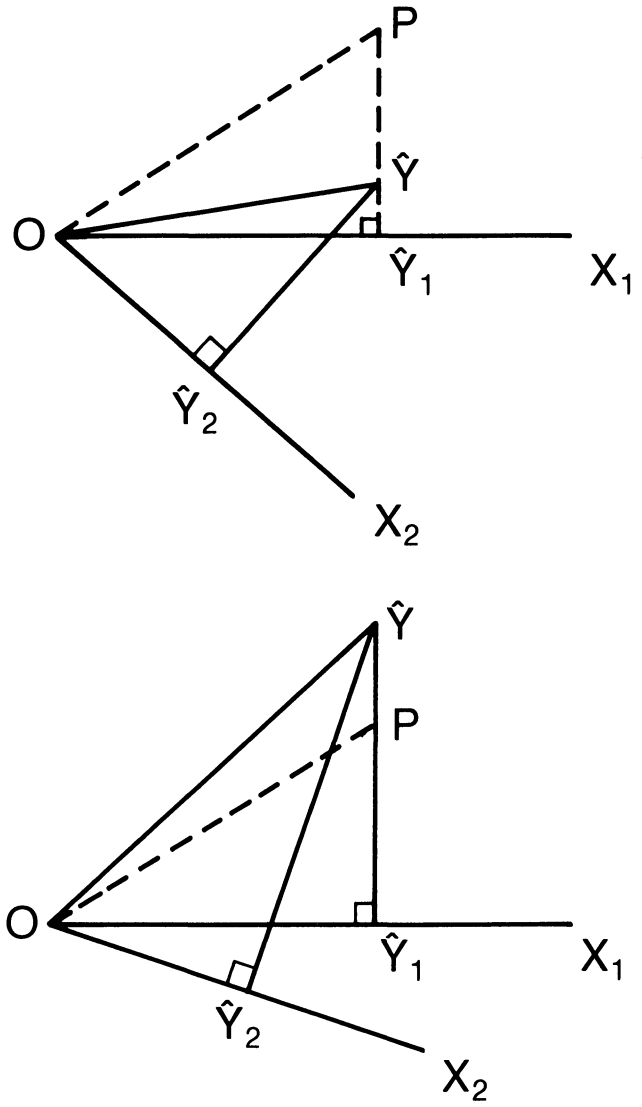

Figure 1. Illustration of Regression on Two Variables. The average-corrected vectors X_1 and X_2 span a plane that contains the average-corrected fitted values $\hat{Y}$. The results for simple regression, $\hat{Y}_1$ and $\hat{Y}_2$, are the projections of $\hat{Y}$ on X_1 and X_2. The line segment OP is constructed to have squared length $SSR(x_1) + SSR(x_2)$ and can easily be compared with $O\hat{Y}$, which has squared length $SSR(x_1, x_2)$. The top panel illustrates a case in which $R^2 < r^2_{yx_1} + r^2_{yx_2}$; the lower panel shows the reverse inequality.

Table 1. Data for the Extreme Example

x_1	x_2	y
2.23	9.66	12.37
2.57	8.94	12.66
3.87	4.40	12.00
3.10	6.64	11.93
3.39	4.91	11.06
2.83	8.52	13.03
3.02	8.04	13.13
2.14	9.05	11.44
3.04	7.71	12.86
3.26	5.11	10.84
3.39	5.05	11.20
2.35	8.51	11.56
2.76	6.59	10.83
3.90	4.90	12.63
3.16	6.96	12.46

onto the plane spanned by both X_1 and X_2 is the sum of the separate projections onto X_1 and onto the component of X_2 which is orthogonal to X_1. The latter component is parallel to $X_2 - \hat{X}_2$, the vector residuals from the simple regression of x_2 on x_1, and the squared length of the projection of Y onto this component is $\mathrm{SSR}(x_2|x_1)$. Equation (3) is simply the application of the Pythagorean theorem to the triangle $O\hat{Y}_1\hat{Y}$. Whether or not $R^2 > r^2_{yx_1} + r^2_{yx_2}$ clearly depends on the configuration of the three vectors in sample space and not just on the angles between Y and X_1 and Y and X_2.

Any three lines starting at the origin can be used to illustrate X_1, X_2, and $\hat{Y}$. The points $\hat{Y}_i$, however, must be chosen so that $O\hat{Y}_i$ is orthogonal to $\hat{Y}_i\hat{Y}$. The upper panel of Figure 1 shows a case in which the projection of Y onto the plane spanned by X_1 and X_2 is only slightly longer than the separate projections onto X_1 and X_2. Here $|O\hat{Y}_2|^2 = \mathrm{SSR}(x_2)$ exceeds $|\hat{Y}_1\hat{Y}|^2 = \mathrm{SSR}(x_2|x_1)$, so by condition (4), $R^2 < r^2_{yx_1} + r^2_{yx_2}$. This inequality can be illustrated by extending the line segment $|\hat{Y}_1\hat{Y}|$ to the point P so that $|\hat{Y}_1P|^2 = |O\hat{Y}_2|^2 = \mathrm{SSR}(x_2)$. Then by Pythagorus, $|OP|^2 = \mathrm{SSR}(x_1) + \mathrm{SSR}(x_2)$, which in this case is clearly larger than $|O\hat{Y}|^2 = \mathrm{SSR}(x_1, x_2)$. The lower panel of Figure 1 shows a case in which the projection of Y onto the plane spanned by X_1 and X_2 is considerably longer than the separate projections onto X_1 and X_2. Here $|O\hat{Y}_2|^2 = \mathrm{SSR}(x_2)$ is less than $|\hat{Y}_1\hat{Y}|^2 = \mathrm{SSR}(x_2|x_1)$ and $|OP|^2 = \mathrm{SSR}(x_1) + \mathrm{SSR}(x_2)$ is less than $|O\hat{Y}|^2 = \mathrm{SSR}(x_1, x_2)$, so $R^2 > r^2_{yx_1} + r^2_{yx_2}$. Note that the simple regressions of y on x_i are the same in the two panels of the figure. It is the change in angle between X_1 and X_2 that leads to the change in the ordering of R^2 and $r^2_{yx_1} + r^2_{yx_2}$.

3. AN EXTREME EXAMPLE

Kendall and Stuart's example is best understood using the geometric approach. For R^2 to be 1, Y must lie in the plane spanned by X_1 and X_2, and for r_{yx_1} to be 0, Y and X_1 must be orthogonal. The other simple correlation, r_{yx_2}, can be made arbitrarily small by making X_2 nearly orthogonal to Y and, therefore, nearly parallel to X_1. In this extreme example, then, x_1 and x_2 must be very highly correlated, with $r^2_{x_1x_2} = 1 - r^2_{yx_1}$. This correlation is also obtained from (8) by substituting $R^2 = 1$ and $r_{yx_2} = 0$.

Data illustrating this example with any desired value of r_{yx_2} and $r_{x_1x_2}$ can be generated as follows. First, find two orthogonal vectors with mean 0 and unit length u_1 and u_2. These can be obtained from the second and third columns of the Gram–Schmidt orthogonalization of a matrix with three columns, where the first is a vector of ones and the second and third are arbitrary. Then let $X_1 = u_1$, $X_2 = au_1 + bu_2$, and $Y = u_2$, which gives $Y = (X_2 - aX_1)/b$ and $r^2_{yx_2} = b^2/(a^2 + b^2)$. Introducing shift and scale terms, $\overline{X}_1$, $\overline{X}_2$, $\bar{y}$, and c_{x_1}, c_{x_2} and c_y, in each variable does not effect r_{yx_2} but gives $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where $y = \bar{y} + c_y Y$, $x_i = \overline{X}_i + c_{x_i} X_i$, $\beta_0 = \bar{y} + \beta_1 \overline{X}_1 - \beta_2 \overline{X}_2$, $\beta_1 = ac_y/bc_{x_1}$, and $\beta_2 = c_y/bc_{x_2}$.

The data shown in Table 1 and Figure 2 were generated with $a = -1$ and $b = .4843$, resulting in $r^2_{yx_2} = .19$. Shifts of 3, 7, and 12, and scales of 2, 6, and 3 were used for X_1, X_2, and y, respectively, so $\beta_0 = -4.52$, $\beta_1 = -3.10$, and $\beta_2 = 1.03$. The plots of y versus x_1 and y versus x_2 show little linear relationship. A sharp observer would notice that the plots are nearly mirror images due to the strong correlation between x_1 and x_2.

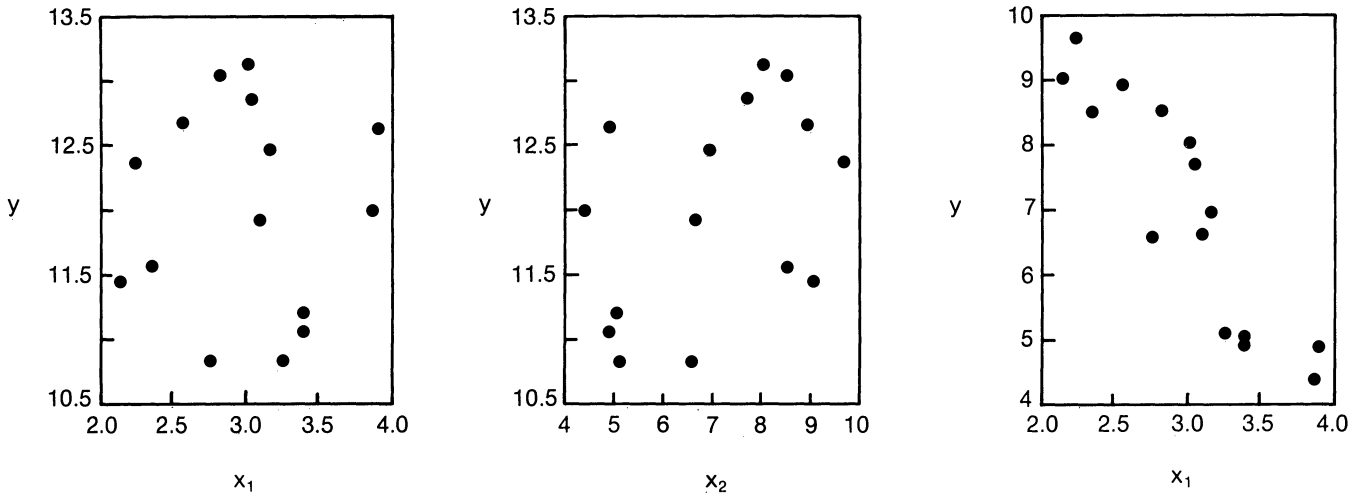

Figure 2. Scatterplots of the Example Data. The left panel shows y and X_1, which have zero simple correlation. The middle panel shows y and x_2, which have squared simple correlation .19. The right panel shows x_1 and x_2, which have squared simple correlation .81. The first two panels can be misleading because the coefficient of multiple correlation is 1.

4. DISCUSSION

The example illustrates the inadequacy of simple measures of association and $x-y$ scatterplots in multiple regression. When $k > 2$ the effect can be more dramatic, because extreme multicollinearity may be present even though the simple correlations between the x's are small. Morrison (1983) stated that the problem of multicollinearity "can be avoided by omitting all but one of the *redundant* variables" (italics mine) and that this should be done "before carrying out the regression analysis" (p. 273). As the example shows, this can amount to throwing out the baby with the bathwater. In discussions of selection techniques, pros and cons of the various approaches are often not given. Draper and Smith (1981, p. 307) stated that backward elimination is "satisfactory" for those not wanting to "miss anything." Chatterjee and Price (1977, p. 203) were more explicit in recommending backward elimination rather than forward selection in cases in which multicollinearity is present. They referred to Mantel (1970) who illustrated, with an example similar to the aforementioned one, that forward selection can fail to detect important variables.

As instructors of statistics, we should be careful not to leave students with the impression that correlated explanatory variables are always redundant. We should point out the dangers of relying on $x-y$ scatterplots and simple correlations, of discarding variables as a cure for multicollinearity, and of using the forward selection technique with correlated explanatory variables.

[*Received July 1985. Revised December 1986.*]

REFERENCES

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*, New York: John Wiley.

Bryant, P. (1984), "Geometry, Statistics, Probability: Variations on a Common Theme," *The American Statistician*, 38, 38–48.

Chatterjee, S., and Price, B. (1977), *Regression Analysis by Example*, New York: John Wiley.

Daniel, C., and Wood, F. S. (1980), *Fitting Equations to Data* (2nd ed.), New York: John Wiley.

Draper, N., and Smith, H. (1981), *Applied Regression Analysis* (2nd ed.), New York: John Wiley.

Herr, D. G. (1980), "On the History of the Use of Geometry in the General Linear Model," *The American Statistician*, 34, 43–47.

Kendall, M. G., and Stuart, A. (1973), *The Advanced Theory of Statistics* (Vol. 2, 3rd ed.), New York: Hafner Publishing.

Mantel, N. (1970), "Why Stepdown Procedures in Variable Selection," *Technometrics*, 12, 621–625.

Margolis, M. S. (1979), "Perpendicular Projections and Elementary Statistics," *The American Statistician*, 33, 131–135.

Morrison, D. F. (1983), *Applied Linear Statistical Methods*, Englewood Cliffs, NJ: Prentice-Hall.

Ott, L. (1984), *An Introduction to Statistical Methods and Data Analysis* (2nd ed.), Boston: Duxbury Press.

Ryan, B. F., Joiner, B. L., and Ryan, T. A. (1985), *Minitab Handbook* (2nd ed.), Boston: Duxbury Press.

A Note on Stagewise Regression

WILLIAM M. ALLEY*

Suppose one estimates the coefficient β_2 in $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ by stagewise regression. That is, first the model $E[Y] \cong \beta_0 + \beta_1 X_1$ is fit using simple linear regression followed by a simple linear regression of the residuals from this model on X_2 to yield the estimator $\tilde{b}_2$. The ratio of the squared t statistic for the estimate b_2 from multiple regression to the squared t statistic for $\tilde{b}_2$ is greater than or equal to 1.0 and is shown to be a convenient function of correlation coefficients among Y, X_1, and X_2. Examination of stagewise regression can provide useful insights when introducing concepts of multiple regression.

KEY WORDS: Multiple regression; Partial correlation coefficient; Regression; t statistic.

1. INTRODUCTION

Consider the linear regression model $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ estimated as

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2. \tag{1}$$

Rather than fitting this full model, suppose one first fits the simple linear regression model

$$Y = b_0' + b_1' X_1 + e. \tag{2}$$

In a second step, the residuals e are regressed against the second variable to yield the estimator

$$\hat{e} = b_0'' + b_2'' X_2. \tag{3}$$

Combining the results from (2) and (3) yields an estimator for $E[Y]$,

$$\tilde{Y} = \tilde{b}_0 + \tilde{b}_1 X_1 + \tilde{b}_2 X_2, \tag{4}$$

where $\tilde{b}_0 = b_0' + b_0''$, $\tilde{b}_1 = b_1'$, and $\tilde{b}_2 = b_2''$.

The preceding is the simplest case of a procedure that has come to be known as stagewise regression. The procedure was originally referred to as *stepwise least squares* (Goldberger and Jochems 1961) or *residual analysis* (Freund, Vail, and Clunies-Ross 1961a,b). These two names for the procedure have subsequently been dropped for obvious reasons. Prior to the advent of the high-speed computer, stagewise regression was used at times as a simple method of estimating β's in multiple regression.

The relation between b_2 obtained using the full model and $\tilde{b}_2$ obtained using stagewise regression was derived by Freund et al. (1961a,b) and Goldberger and Jochems (1961)

*William M. Alley is with the U.S. Geological Survey, 412 National Center, Reston, VA 22092. The author thanks Brent Troutman, Ed Gilroy, Aldo Vecchia, and two anonymous reviewers for helpful comments.