Int J Biomed Comput, 33 (1993) 1–23
Elsevier Scientific Publishers Ireland Ltd.

A PC PROGRAM TO AID IN THE CHOICE OF THE DESIGN MATRIX IN MULTIPLE LINEAR REGRESSION

AMY M. FUREY[a], CHARLES J. KOWALSKI[b], EMET D. SCHNEIDERMAN[c] and STEPHEN M. WILLIS[c]

[a]*Department of Biostatistics and* [b]*Oral Biology, University of Michigan, Ann Arbor, MI 48109 and* [c]*Department of Oral and Maxillofacial Surgery, Baylor College of Dentistry, 3302 Gaston Ave., Dallas, TX 75246 (USA)*

(Received September 28th, 1992)
(Accepted October 28th, 1993)

A PC program, DESIGN, which can be used to evaluate and compare alternative choices of the design matrix, **X**, in the general linear model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ is described, illustrated and made available to interested readers. Given **X**, the program (1) computes various measures of the 'stability' of **X** and **X'X** and (2) determines the precisions of estimates of the model parameters, β, and of predicted values, $\hat{\mathbf{y}}$, at the given design points. Examples focusing on polynomial regression are given.

Key words: Design matrix; Linear model; Regression; PC program

Introduction

Multiple linear regression is among the statistical tools most often used in biomedical research. Good general accounts are given in literature [1–3], and we make repeated reference to these publications for many of the details that are omitted here. In accordance with Ref. 1 (p. 237), we write the general linear regression model in the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon \qquad (1)$$

for $i = 1, 2, \ldots, n$. This relates the response or 'dependent' variable, y_i, to a number of predictor or 'independent' variables, x_{ij}. The β values are parameters to be estimated from the data and the ϵ_i values represent random errors.

Correspondence to: Emet D. Schneiderman, Department of Oral and Maxillofacial Surgery, Baylor College of Dentistry, 3302 Gaston Ave., Dallas, TX, 75246, USA.

0020-7101/93/$06.00 © 1993 Elsevier Scientific Publishers Ireland Ltd.
Printed and Published in Ireland

The model (1) is more conveniently written in matrix notation (Ref. 1, p. 238) as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{2}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

and $\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

In the above, $\mathbf{y}$ is a $(n \times 1)$ vector of observations, $\mathbf{X}$ is a $(n \times p)$ matrix of constants, β is a $(p \times 1)$ vector of parameters, and ϵ is a $(n \times 1)$ vector of errors or residuals assumed to satisfy

$$\epsilon \sim MVN\,(0, \sigma^2\mathbf{I}) \tag{3}$$

i.e., we assume ϵ has a multivariate normal distribution with mean or expected value $E(\epsilon) = 0$ and covariance matrix $V(\epsilon) = \sigma^2\mathbf{I}$

Given this structure, the mean of $\mathbf{y}$ is $E(\mathbf{y}) = \mathbf{X}\beta$ and its covariance matrix is $V(\mathbf{y}) = \sigma^2\mathbf{I}$ (Ref. 1, p. 238). The least squares estimator of β is (Ref. 1, p. 239)

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{4}$$

and this is *unbiased* with $E(\mathbf{b}) = \beta$ and covariance matrix (Ref. 1, p. 242)

$$V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2 \tag{5}$$

Now, when the above procedure is carried out in the context of a designed experiment, the design matrix, $\mathbf{X}$, is determined or controlled by the investigator and is not a part of the *results* of the experiment (the response $\mathbf{y}$ represents the experimental outcome). The 'best' choice of $\mathbf{X}$ depends on the purpose(s) to which the fitted equation will be put. A good discussion is given in Netter et al. (Ref. 1, p. 175). Even in the simplest case of simple linear regression (where there is but a single x variable and $\mathbf{X}$ is $n \times 2$), among other things, the experimenter will have to consider (Ref. 1, p. 175): (i) How many values of x should be studied? (ii) What shall the two extreme values be? (iii) How should the other values of x, if any, be spaced? and (iv) How many observations should be taken at each x-value? The situation is obviously

more complicated when there are multiple x's: The above points need to be considered for each x, and the combinations of levels of the variables must be selected. We focus on two related but separable aspects of the choice of **X**, namely, the accuracy with which b is computed from Eqn. 4 and the magnitudes of the variances of the elements of **b** as measured in Eqn. 5 (Ref. 4, p. 4).

It is well known that, with respect to the first consideration, it is advantageous to choose **X** so that its columns are *uncorrelated or orthogonal* (Ref. 1, p. 271): A design matrix **X** with mutually orthogonal columns 'represents the best possible experimental data' (Ref. 4, p. 109). It is also true that choosing the columns of **X** to be orthogonal can favorably impact on the variances of the elements of **b**, viz., $V(b_i)$. Theoretical discussions are given Refs. 3 (p. 58) and 5 (p. 193). Graphical demonstrations are given in Refs. 6 and 7 (Chap. 14). It is, of course, not always possible to choose an orthogonal **X** and in such cases both the accuracy of Eqn. 4 and the magnitude of Eqn. 5 will depend on the extent of the correlations between the columns of **X**. The terms *collinearity, multicollinearity* and *ill-conditioned* are used to describe situations in which these columns are highly correlated, and multicollinearity can have profound effects on the accuracy with which the regression coefficients and their variances are estimated (Ref. 1, p. 275). Historically, the existence of multicollinearity has been indicated by a small eigenvalue of $\mathbf{X}'\mathbf{X}$ and/or the associated correlation matrix, **R**, and the extent to which a given **X** is ill-conditioned has be measured in several ways, including *variance inflation factors* (VIFs) and the multiple correlations between a given column of **X** and the remaining columns (Ref. 1, p. 39). These measures are, of course, also useful in situations in which **X** is not under the control of the experimenter but, rather, together with **y**, constitute the outcome of the experiment. Our program may be used in such cases to check on multicollinearity in $\mathbf{X}'\mathbf{X}$ and to assess the extent to which this problem can be ameliorated by remedial measures such as *centering* or *scaling*. Centering refers to subtracting a constant from each value in each of the columns of **X** (e.g. the mean of the values in that column). Scaling refers to dividing each value in a given column of **X** by a constant (e.g., the standard deviation of the values in that column). The two can be used separately or in combination. Centering and/or scaling can effect the stability of $\mathbf{X}'\mathbf{X}$, but it is important to realize that which (if any) to use depends on the problem under consideration. An invaluable reference in this regard is Belsley et al. [4].

As concerns the variances of the estimators, it is seen that $V(\mathbf{b})$ is proportional to $(\mathbf{X}'\mathbf{X})^{-1}$ and it is clear that, when **X** is under the control of the experimenter, it is advantageous to choose it so that $(\mathbf{X}'\mathbf{X})^{-1}$ is 'small'. Two of the more widely used criteria for 'smallness' in this context are the trace (tr) and determinant (det) of $(\mathbf{X}'\mathbf{X})^{-1}$ (Ref. 3, p. 92) The trace of a matrix is the sum of the diagonal elements of the matrix; in this case, the sum of the variances of the elements of **b**. When this is small, the average variance of the b_j will be small. The determinant is sometimes called the *generalized variance* (Ref. 8, p. 139) and, again, it is desirable that this quantity be small. The design **X** is said to be A-optimal if it minimizes $\mathrm{tr}[(\mathbf{X}'\mathbf{X})^{-1}]$. It is said to be D-optimal if it maximizes $\det[(\mathbf{X}'\mathbf{X})]$ (this is the same as minimizing the generalized variance). A-optimal designs minimize the 'total variance'. D-optimal designs minimize the (hyper) volume of fixed level confidence regions for β; they also minimize the maximum variance of any predicted value (Ref. 3, p. 92)

The choice of **X** has received much attention in the literature and optimal designs exist for many combinations of criteria and experimental structures. Useful reviews are given in the literature [9,10]. Unfortunately, it is not always possible to find *the* optimal design in a given situation and, even when such a design exists, practical considerations and/or physical constraints may preclude its use. It may be of interest in such situations to compare the *possible* with the *ideal* [10]. Moreover, optimal designs tend to concentrate all experimental runs on a small number of design points and, while they may be ideally suited to estimating the coefficients of the assumed model, they provide little or no ability to check for lack of fit [10]. For example, in simple linear regression, if our primary interest is in estimating the slope, β_1, we minimize $V(b_1)$ by using two levels in **X**, at the two extremes for the scope of the model, and placing half the observations at each of the two levels (Ref. 1, p. 175). This is the best design *given the model*, but it provides *no* information concerning the fit of the model. For polynomial regression, D-optimal designs for estimating a polynomial of degree D locate experimental runs at exactly $D + 1$ distinct levels of the predictor variable [10]. This design can provide no indication that a higher degree polynomial may be needed (Ref. 3, p. 186). A number of alternative strategies have been devised to cope with this situation [10] by compromising optimality to allow some ability to test the model and our program can be used to discover some of the properties of these compromise designs and compare their efficiencies with respect to the problem in hand.

In any case, experimenters are often faced with the problem of choosing between competing (possible) designs or assessing the impact of a proposed change in strategy. Our program, DESIGN, was developed to allow the evaluation of a *given* design and the comparison of several designs. It is meant to be used prior to experimentation and does not depend on the values of the response variable, **y**. We emphasize that DESIGN is *not* an optimization program: it does not construct optimal designs. Rather, it is a means for the evaluation and comparison of user-proposed designs. Some of the measures used in this process were alluded to briefly above. In the next section we descibe the program's output in more detail and indicate how each can be used for evaluative and/or comparative purposes. We should note that the potential value of a program of this type was recognized much earlier [11]. Indeed, these individuals developed a program (called EXPLOR) for the General Electric MARK III time-sharing system as early as 1975. Our program, which follows their outline closely, was written both to facilitate our work in polynomial growth curve models [12–18] and to make this useful tool more readily accessible to biomedical researchers. Given our emphasis, we differ from EXPLOR in that we do not consider models without an intercept term (β_0). In addition to the measures given in Meeker et al. [11], we provide measures based on the *singular value decomposition* of **X**, which have a number of desirable properties (Ref. 4, p. 98) Information on obtaining a copy of **DESIGN** is provided in the appendix.

The Program

The menu-driven program is invoked by the single command

gsruni design

The user is then 'prompted' for the location and name of the file containing the matrix, **X**. This may be either an ASCII or GAUSS file. We should note here that while DESIGN is written in GAUSS [19], it is not necessary to have purchased or installed GAUSS to run the program. We allow the use of GAUSS files for the convenience of users who have access to GAUSS, but our program stands alone. The user then is presented with the opportunity to specify the value of σ^2 to be used in the computations (the default value is $\sigma^2 = 1$). If the user can provide a reasonable guess at σ^2, certain of the quantities computed below will be more readily interpretable, especially the confidence intervals for the parameters and predicted values whose lengths will then incorporate this information.

The output includes:

(I) **X**, **R**, det(**R**), $\mathbf{R}^{-1}$, SQRT(**R**), **X**′**X**, and $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. The input matrix, **X**, is echoed so the user may check for accuracy. **R** is the $(p-1) \times (p-1)$ correlation matrix containing the correlations between the columns of **X** (excluding the first column which has every element 1; we denote these columns as $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{p-1}$). Large off-diagonal elements (near unity) in **R** indicate collinearity. det(**R**) has been called, 'an essential part of a good computer regression routine' (Ref. 2, p. 264). Small values of this determinant signal that the estimated regression coefficients may be unstable. $\mathbf{R}^{-1}$ is the inverse of **R** whose use in multivariate analysis — including multiple regression, factor analysis and discriminant function analysis — was considered in Raveh [20]. Our interest in $\mathbf{R}^{-1}$ is in computing the VIFs described below. SQRT(**R**) is the Cholesky or square-root factorization of **R**. It is an upper triangular matrix such that SQRT(**R**)′ × SQRT(**R**) = **R**. Each squared diagonal element of this matrix is of the form $1 - R_{j*}^2$ where R_{j*} is the multiple correlation of $\mathbf{x}_j$ with the preceding x's. These multiple correlations are related to det(**R**), viz., $\det(\mathbf{R}) = \Pi(1 - R_{j*}^2)$, where Π denotes the product. The diagonal elements of $\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ are the variances of the elements of **b**, the estimated regression coefficients. We provide details concerning the computation of $(\mathbf{X}'\mathbf{X})^{-1}$ under IV below.

(II) The VIFs and the multiple correlations between each single **x** and all the other **x** values. The VIFs are the diagonal elements of $\mathbf{R}^{-1}$ and their diagnostic value stems from the relation $\mathrm{VIF}_j = 1/(1 - R_j^2)$ where R_j^2 is the multiple correlation coefficient of $\mathbf{x}_j$ regressed on the remaining explanatory variables. A high value for a given VIF indicates an R^2 near unity, and hence points to multicollinearity. It has been suggested that when the largest VIF exceeds 10, this indicates that multicollinearity may unduly influence the least squares estimates. Alternatively, mean VIF values considerably larger than 1 are indicative of serious multicollinearity problems (Ref. 1, p. 392). The exact nature of these problems may be seen from the relationship [21]

$$V(b_j) = \frac{\sigma^2 \mathrm{VIF}_j}{(n-1)S_j^2} \tag{6}$$

where S_j^2 is the variance of the elements in $\mathbf{x}_j$. This shows that $V(b_j)$ is directly proportional to VIF_j. Confidence intervals for b_j will increase in length by the factor $[\mathrm{VIF}_j]^{1/2}$. It has been suggested [22] that the VIF be used as a measure of how many times larger $V(b_j)$ will be for correlated data than for orthogonal data (where

each VIF is unity). It is the ratio of the variance of b_j to what that variance would be if $\mathbf{x}_j$ were uncorrelated with the remaining $\mathbf{x}$ values. Consider e.g., the case where the correlation matrix has equal off-diagonal elements, viz.,

$$\mathbf{R} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

This can be written in the form $\mathbf{R} = (1 - \rho)[\mathbf{I} - \rho\mathbf{J}]$ where $\mathbf{J}$ is the $(p - 1) \times (p - 1)$ matrix with every element 1. It can be shown that, for $(-1)/(p - 2) < \rho < 1$,

$$\mathbf{R}^{-1} = \frac{1}{1 - \rho}\left[\mathbf{I} - \frac{\rho}{1 + (p - 2)\rho}\mathbf{J}\right]$$

and hence that

$$\mathrm{VIF}_j = \frac{1 + (p - 3)\rho}{(1 - \rho)[1 + (p - 2)\rho]}$$

It is seen that as $\rho \to 1$, $\mathrm{VIF}_j \to \infty$. The values of VIF for several combinations of p and ρ are given in by Mansfield and Helms [22]. For example, for $p = 4$, $\rho = -0.495421$, VIF = 36.844. Changing ρ slightly to $\rho = -0.499$ gives VIF = 167. As ρ approaches -0.5, VIF approaches infinity (for $p = 4$).

The VIF values and the R_j^2 values are also related to the *tolerance* by tolerance = $1/\mathrm{VIF}_j = 1 - R_j^2$. Many computer packages will automatically exclude variables with tolerances less than 0.01 or 0.001 (Ref. 1, p. 393). Tolerance is usually thought of in the context of computational accuracy. As a rule of thumb, the number of leading zeros in the tolerance is the number of significant digits lost if the computation includes the variable in question.

(III) The matrices $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The matrix $\mathbf{C}$ is sometimes referred to as the *catcher* matrix [23,24]. Since $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{C}\mathbf{y}$, the rows of $\mathbf{C}$ consist of the weights $(c_{i1}, c_{i2}, \ldots, c_{in})$ that enter into the expressions

$$\hat{b}_i = c_{i1}y_1 + c_{i2}y_2 + \cdots + c_{in}y_n \tag{7}$$

which reflect the sensitivity of the estimated coefficients to the (to be observed) responses [11]. The elements of $\mathbf{C}$ are related to the VIFs by [24]

$$\mathrm{VIF}_j = \sum_i c_{ij}^2 \sum_i (x_{ij} - \bar{x}_j)^2 \tag{8}$$

and to the $V(b_j)$ by (Ref. 4, p. 13)

$$V(b_j) = \sigma^2 \sum_i c_{ij}^2 \tag{9}$$

The matrix **H** is known as the *hat* matrix (Ref. 1, p. 220). It is related to the fitted values of **y** by $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$. A number of the properties and uses of **H** were given in Hoaglin and Welsch [25]. For our purposes, the most important is the role that **H** plays in the covariance matrices of $\hat{\mathbf{y}}$ and the residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. viz.,

$$V(\hat{\mathbf{y}}) = \sigma^2\mathbf{H} \tag{10}$$

and

$$V(\mathbf{e}) = \sigma^2 (\mathbf{I} - \mathbf{H}) \tag{11}$$

The rows of **H** consist of the weights $(h_{i1}, h_{i2}, \ldots, h_{in})$ that enter into the expressions

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n \tag{12}$$

reflecting the relative importance of the y_i in predicting the response variable: h_{ij} has the direct interpretation as the amount of leverage or influence exerted on $\hat{y}_i$ by y_j (regardless of the value of y_j since **H** depends only on **X**). With fixed **X**, we can examine, and perhaps modify, the experimental conditions in advance.

It can also be seen from Eqn. 11 that the variance of a given residual, say e_i, is given by $V(e_i) = (1 - h_{ii})$ where h_{ii} is the ith diagonal element of **H**. The element h_{ii} is called the leverage of the ith observation. The larger the value of h_{ii}, the smaller e_i, i.e. the closer the fitted value $\hat{y}_i$ will be to the observed value y_i. Similarly, from Eqn. 12, the larger h_{ii}, the more important y_i is in determining $\hat{y}_i$. It can be shown that (Ref. 1, p. 402)

$$0 \leq h_{ii} < 1$$

$$\text{tr}\{\mathbf{H}\} = p$$

and

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2$$

this last equation showing in fact that h_{ii} summarizes the leverage of y_i on all the fitted values. A given h_{ii} is usually considered 'large' if it exceeds twice the average leverage, viz., $2p/n$ (Ref. 1, p. 403). This, however, is based on an approximation that is accurate only for large n and p. For smaller n and p, $3p/n$ may be more appropriate [24].

The value of h_{ii} may also be interpreted as a measure of the distance between the x values for the ith case and and the means of the x values for all n cases. A large leverage h_{ii} signals that the ith case is distant from the center of all the observations. The Mahalanobis distance of $\mathbf{x}_i$ from the other observations can be written [24]

$$\frac{n(n-2)}{n-1}\left(\frac{h_{ii}-1/n}{1-h_{ii}}\right)$$

When designing an experiment, it is desirable to choose the x values to be roughly equally influential, i.e. to have each $h_{ii} \approx p/n$ (Ref. 4, p. 17) This recommendation is based on results given in the literature [26,27]. Huber [26] studied the effects of outliers on experimental designs for linear models and suggested using designs for which the h_{ii} are all *well below* unity. Indeed, he suggested that $h_{ii} > 0.2$ implies that too much weight is being given to y_i (note that this cut-off point is independent of p and n). He showed that

$$\hat{y}_i = (1 - h_{ii})\mathbf{x}_i\,\mathbf{b}(i) + h_{ii}y_i$$

where (here only) $\mathbf{x}_i$ is the ith *row* of $\mathbf{X}$ and $\mathbf{b}(i)$ is the estimate of β when the ith observation is omitted. This expresses $\hat{y}_i$ as a weighted average of y_i and a quantity which does not depend on y_i (it depends on the other y values since they are involved in computing $\mathbf{b}(i)$). It is easily seen from this that a point with $h_{ii} = 1$ completely determines its predicted value. It is also clear that $\partial\hat{y}_i/\partial y_i = h_{ii}$, i.e. h_{ii} is a measure of the rate of change of $\hat{y}_i$ with respect to y_i. Box and Draper [27] showed that the effect of one or more outliers on the vector of predicted values was proportional to Σh_{ii}^2 and that this is minimized when $h_{ii} = p/n$ for all i. This reinforces the suggestion that one should take $h_{ii} \approx p/n$ for all i, and also shows that p/n itself should not be large, i.e. that n should be large relative to p. D-optimal designs have all $h_{ii} = p/n$.

As an example, for linear regression with a single explanatory variable,

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$$

This shows that the further x_i is from its mean, the more influence it exerts. Conversely, a large h_{ii} signals an observation which is distant from the center of the data.

Finally, we note that studying the h_{ii} can be useful even if the experiment has been performed and the residuals, e_i, are available for scrutiny. Emphasis is properly placed on the examination of residuals (Ref. 2, Chap. 3) when the data have been analyzed by multiple linear regression, but this study should incorporate a consideration of leverage. We have seen that large h_{ii} can cause problems, but large h_{ii} are accompanied by small e_i (large h_{ii} 'forces' the regression surface to be close to y_i) so this aspect of the conditioning problem cannot be discovered through plotting

residuals alone. The h_{ii} are functions of the x values only, and h_{ii} measures the role of the x values in determining how important y_i is in affecting the fitted value $\hat{y}_i$.

(IV) In Meeker et al. [11], considerable emphasis was placed on the eigenvalues and eigenvectors of $\mathbf{X}'\mathbf{X}$ and $\mathbf{R}$. Multicollinearity was said to be indicated by the presence of a 'small' eigenvalue in either or both these matrices, and the ratio of the smallest to the largest eigenvalue of $\mathbf{R}$ was compared to the value of 1 for a completely orthogonal design. For non-orthogonal designs, this ratio is less than 1, sometimes considerably less, and thus the difference between this ratio and 1 was proposed as a measure the degree of non-orthogonality of the design under consideration. For reasons made explicit in Belsley et al. [4], we choose to approach these questions using the *singular value decomposition* (SVD) of $\mathbf{X}$. This will provide information that encompasses that given by the eigensystem of $\mathbf{X}'\mathbf{X}$, and provide additional measures which should prove useful in evaluating and comparing proposed designs. Other uses and properties of the SVD are given in the literature [28–31]. We follow (Ref. 4, p. 98) Any $n \times p$ matrix $\mathbf{X}$ can be written as

$$\mathbf{X} = \mathbf{UDV}' \tag{13}$$

where the $n \times p$ matrix $\mathbf{U}$ and the $p \times p$ matrix $\mathbf{V}$ are orthogonal ($\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$) and $\mathbf{D}$ is a $p \times p$ diagonal with non-negative diagonal elements $d_1, d_2, \ldots, d_p$ called the *singular values* of $\mathbf{X}$. It turns out that the d_i^2 are the eigenvalues of $\mathbf{X}'\mathbf{X}$ and the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{X}'\mathbf{X}$, so we do provide the same information as given in Meeker et al. [11]. However, the approach based on Eqn. 13 which produces the singular values d_i allows the definition of the *condition number* of $\mathbf{X}$ which in turn can be used to remove the subjectivity associated with deciding whether or not $\mathbf{X}'\mathbf{X}$ has a 'small' eigenvalue (Ref. 4, p. 96)

The condition number of $\mathbf{X}$ is defined as

$$\kappa(\mathbf{X}) = d_{\max}/d_{\min} \geq 1 \tag{14}$$

i.e. the ratio of the largest singular value of $\mathbf{X}$ to the smallest. The condition number of any matrix with orthonormal columns is unity, so that it reaches its lower bound in this 'cleanest of all possible cases' (Ref. 4, p. 104) It can also be shown that, for a given $\mathbf{X}$ (not necessarily orthogonal), the maximum VIF of $\mathbf{X}$ is a lower bound on the condition number. The condition number provides a measure of the potential sensitivity of the estimated standard errors of the regression coefficients to small changes in the data. The *elasticity* of the variance of any least squares estimate is bounded by twice the condition number of $\mathbf{X}$, i.e. $2\kappa(\mathbf{X})$ provides an upper bound to the sensitivity of the parameter variances to changes in $\mathbf{X}$. If, for example, a condition number were 100, a 1% change in any element of $\mathbf{X}$ could result in a $2 \times 100\%$ change in the variance of any estimate (Ref. 4, p. 177). We also define and compute the kth *condition index* of $\mathbf{X}$

$$\eta_k = \frac{d_{\max}}{d_k} \text{ for } k = 1, 2, \cdots, p-1 \tag{15}$$

There are as many near-dependencies in the **X** matrix as there are 'large' condition indices, and studies have shown that weak dependencies are associated with condition indices of about 5 or 10, while moderate to strong relationships are indicated for indices of 30–100. Condition indices of 100 or more can cause substantial variance inflation and great potential harm to regression estimates (Ref. 4, p. 153). Note that these rules-of-thumb apply to the condition number and condition indices of **X**. It can be shown that $\kappa(\mathbf{X}'\mathbf{X}) = \kappa^2(\mathbf{X})$, so that ill-conditioning in **X** is greatly compounded in **X'X** (Ref. 4, p. 114).

Returning to the eigenvalues and eigenvectors of **X'X**, these may be useful in determining which linear combinations of **b** may be estimated with precision. Consider the linear combination **a'b** and let $\delta_1, \delta_2, \ldots, \delta_p$ be the eigenvectors corresponding to the eigenvalues $d_1^2, d_2^2, \ldots, d_p^2$. Then relatively precise estimation is possible in the directions of the δ_i corresponding to large eigenvalues, but relatively imprecise estimation is obtained in the directions corresponding to small eigenvalues. In cases where ill-conditioning is a problem, taking additional observations in the directions of eigenvectors corresponding to small eigenvalues can help circumvent the problem. See Seber (Ref. 3, p. 80) for a discussion and further references.

The matrix **U** in Eqn. 13 is related to the hat matrix by $\mathbf{H} = \mathbf{U}\mathbf{U}'$ and this is the preferred way of computing **H** [25]. The SVD is related to the variances of the estimators by

$$V(\mathbf{b}) = \sigma^2 \mathbf{V}\mathbf{D}^{-2}\mathbf{V}'$$

so that for a given b_i

$$V(b_i) = \sigma^2 \sum_{j=1}^{p} \frac{v_{ij}^2}{d_j^2}$$

Note that this decomposes $V(b_i)$ into into a sum of components, each associated with one of the singular values of **X**. Small values of d_j will cause large variances: $v_{ij}^2 = 0$ if columns i and j are orthogonal (Ref. 4, p. 106).

The SVD is also used in our program to compute other quantities, e.g. $(\mathbf{X}'\mathbf{X})^{-1}$. If our program is to be useful, it must be able to accomodate ill-conditioned design matrices; it goes without saying that we must employ the most accurate of the available computational techniques. While GAUSS is, in general, an extremely accurate program, certain of its functions are more accurate than others (at the cost of increased computing time). In particular, the inverse of a matrix can be computed in various ways; among those available are the INVPD and INV functions, the QR algorithm, and the SVD. The INVPD command is based on the Cholesky or square-root factorization in which $\mathbf{X}'\mathbf{X} = \mathbf{T}'\mathbf{T}$ where **T** is a $p \times p$ upper-triangular matrix. INV is based on the Crout decomposition (with partial row pivoting) in which $\mathbf{X}'\mathbf{X} = \mathbf{L}\mathbf{T}$, where **L**(**T**) is lower (upper) triangular. GAUSS uses a tolerance of 10^{-14} for these functions. In the *QR* algorithm, $\mathbf{X} = \mathbf{Q}\mathbf{T}$, where $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ and **T** is upper triangular. Then $\mathbf{X}'\mathbf{X} = \mathbf{T}'\mathbf{Q}'\mathbf{Q}\mathbf{T} = \mathbf{T}'\mathbf{T}$ and $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{T}^{-1}(\mathbf{T}^{-1})'$. In the SVD, $(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{V}^{-1})'\mathbf{D}^{-2}\mathbf{V}^{-1}$. These methods are described, and their accuracies

compared in Seber (Ref. 3, Chap. 11). It is concluded that the SVD is at least as accurate as the other methods and is particularly useful when **X** defines a polynomial regression model or is ill-conditioned in an unpredictable way. We also informally compared these procedures by computing $(\mathbf{X}'\mathbf{X})^{-1}{}^{*}\ \mathbf{X}'\mathbf{X}$ for several highly ill-conditioned matrices. For the 10×8 **X** polynomial regression design matrix corresponding to fitting a 7th degree polynomial to 10 equally spaced points 1(1)10 (cf. the example to follow, viz., the **X** matrix in Eqn. 19), off-diagonal elements were as large as 0.0133 when INV was used. We decided, therefore, to employ the SVD in all applicable situations, in particular in the computation of $(\mathbf{X}'\mathbf{X})^{-1}$ and all quantities involving this matrix.

(V) $\det\{(\mathbf{X}'\mathbf{X})\}$ and $\mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\}$. When comparing two designs, the one with the larger value of $\det\{(\mathbf{X}'\mathbf{X})\}$ is preferred using the *D*-criterion [9]; while the one with the smaller value of $\mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\}$ is preferred if the *A*-criterion is used [32]. The trace criterion for optimality leads to **X** matrices which have orthogonal columns; however, in the general case of multiple regression, this has several shortcomings (e.g., a dependence on the scaling of the x-variables) so that there is a general preference for *D*-optimallty. There are algorithms available for producing *D*-optimal designs [33]. For more details and references see Ref. 3 (p. 92). We consider the special case of polynomial regression, where we claim A-optimality may be more useful, later.

In any case, the quantity $\mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\}$ can be used to gain considerable insight into a number of issues regarding the 'goodness' of **b** in estimating β as a function of the design. Several of these are considered below. It can be shown [34,35] that the expected squared distance between **b** and β can be written

$$E[(\mathbf{b}-\beta)'(\mathbf{b}-\beta)] = \sigma^2\mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\} = \sigma^2 \sum_{i=1}^{p} \frac{1}{d_i^2} > \frac{\sigma^2}{d_{\min}^2} \tag{16}$$

where $d_1^2, d_2^2, \ldots, d_p^2$ are the eigenvalues of $\mathbf{X}'\mathbf{X}$. Thus when $\mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\}$ is large and/or $\mathbf{X}'\mathbf{X}$ has a small eigenvalue, this distance can be expected to be large. The inequality in [16] shows how $d_{\min}^2$ can be used to provide a quick rule of thumb for how far **b** will be from β.

It can also be shown that the estimated coefficients tend to be *too large* in absolute value, i.e. $|\mathbf{b}| > |\beta|$, viz.,

$$E(\mathbf{b}'\mathbf{b}) = \beta'\beta + \sigma^2\mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\} > \beta'\beta + \frac{\sigma^2}{d_{\min}} \tag{17}$$

and it is possible that some will even have the wrong sign [34]. The more $\mathbf{X}'\mathbf{X}$ is ill-conditioned, the more **b** can be expected to be too large in magnitude. A 'real world' example is given previously [35] where $\mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\} = 33.825$ so that the expected squared distance of **b** from β is $33.825\sigma^2$, which is more than three times what it would be for an orthogonal system (for an orthogonal **X**, $\mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\} = p$ and $p = 10$ in their example). This has lead some to consider 'shrunken estimators' of the form $\lambda\mathbf{b}$ $(0 < \lambda \leq 1)$. See Ref. 3 (p. 90) for a description and references.

Finally, Ref. 3 (p. 88) shows that

$$\sum_{j=0}^{p-1} \mathrm{Var}(b_j) = \sigma^2 \mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\} = \sigma^2 \sum_{j=1}^{p} d_j^{-2} > \frac{\sigma^2}{d_{\min}^2} \tag{18}$$

which shows that the 'total variance' may be high if $\mathbf{X}'\mathbf{X}$ has a small eigenvalue.

(VI) The standard errors (SEs) and half-lengths of the confidence intervals (HLCIs) for each of the elements of **b**. The SE of b_i, the ith element of **b**, is σ times times the square root of the ith diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. The corresponding HLCI is this value times $t\ (1 - \alpha/2;\ n - p)$, i.e. the $(1 - \alpha/2) \times 100\%$ percentile of the t-distribution with $n - p$ degrees of freedom.

(VII) The SEs and HLCIs for the predicted mean value of **y** at the design points, **X**. The fitted or predicted values of **y** are obtained from $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{y}$. The variance of $\hat{\mathbf{y}}$ is $\alpha^2\mathbf{H}$, so that $\mathrm{SE}(\hat{y}_i)$ is σ times the square root of the ith diagonal element of **H**. The corresponding HLCI is this value times $t\ (1 - \alpha/2;\ n - p)$.

The above summarizes the output which is obtained from our program. The output is shown on the screen and automatically written into a file called DESIGN.OUT so that it can be modified using a word processor and printed.

An Example

Consider the polynomial regression model with $\sigma^2 = 1$ where a quadratic equation is to be fit to the $n = 5$ points 1, 2, 3, 4, 5 so that (Ref. 2, p. 260 and Ref. 9)

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \tag{19}$$

We compute

$$\mathbf{R} = \begin{bmatrix} 1 & 0.98110491 \\ & 1 \end{bmatrix}$$

$$\det(\mathbf{R}) = 0.03743315$$

$$\mathbf{R}^{-1} = \begin{bmatrix} 26.714286 & -26.209517 \\ & 26.714286 \end{bmatrix}$$

$$\mathrm{SQRT}(\mathbf{R}) = \begin{bmatrix} 1 & 0.98110491 \\ 0 & 0.19347650 \end{bmatrix}$$

$$\mathbf{X'X} = \begin{bmatrix} 5 & 15 & 55 \\ & 55 & 225 \\ & & 979 \end{bmatrix}$$

and

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} 4.6 & -3.3 & 0.5 \\ & 2.671486 & -0.42857143 \\ & & 0.071428571 \end{bmatrix}$$

where we do not repeat the lower elements of symmetric matrices. From this it is seen that $VIF_1 = VIF_2 = 26.714286$ ($R^2 = 0.96256684$). Substantial variance inflation is in evidence. In this example, with $p = 3$, SQRT(**R**) adds no information to that already contained in R^2, viz., $1 - R^2_{2.1} = (0.19347650)^2$ gives $R_{2.1} = R^2$, as it should. For larger values of p, recall that successive squared diagonal elements of SQRT(**R**) refer to the squared multiple correlations with the *preceding* variables.

The catcher and hat matrices are given by

$$\mathbf{C} = \begin{bmatrix} 1.8 & 0 & -0.8 & -0.6 & 0.6 \\ -1.0571429 & 0.32857143 & 0.85714286 & 0.52857143 & -0.65714286 \\ 0.14285714 & -0.071428571 & -0.14285714 & -0.071428571 & 0.14285714 \end{bmatrix}$$

and

$$\mathbf{H} = \begin{bmatrix} 0.88571429 & 0.25714286 & -0.085714286 & -0.14285714 & 0.085714286 \\ & 0.37142857 & 0.34285714 & 0.17142857 & -0.14285714 \\ & & 0.48571429 & 0.34285714 & -0.085714286 \\ & & & 0.37142857 & 0.25714286 \\ & & & & 0.88571429 \end{bmatrix}$$

It is seen from Eqn. 7 e.g., that

$$\hat{b}_0 = 1.8y_1 + 0y_1 - 0.8y_3 - 0.6y_4 + 0.6y_5$$

Similarly, from Eqn. 11

$$\hat{y}_1 = 0.8857y_1 + 0.2571y_2 - 0.0857y_3 - 0.1428y_4 + 0.0857y_5$$

One can also see that the h_{ii} vary somewhat from the optimal (constant) value $p/n = 3/5 = 0.6$ suggested in literature [26,27], but that none exceeds $2p/n = 1.2$, so that the while the proposed design may be sub-optimal, none of the h_{ii} is unduly large.

The singular value decomposition $\mathbf{X} = \mathbf{UDV}'$ yields

$$\mathbf{U} = \begin{bmatrix} -0.038954489 & -0.52790324 & 0.77814845 \\ -0.13670189 & -0.58903782 & -0.075997413 \\ -0.29496108 & -0.45745271 & -0.43525771 \\ -0.51373205 & -0.13314788 & -0.29963243 \\ -0.79301481 & 0.38387664 & 0.33087841 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} 32.156334 & 0 & 0 \\ 0 & 2.1977332 & 0 \\ 0 & 0 & 0.37437558 \end{bmatrix}$$

and

$$\mathbf{V}' = \begin{bmatrix} -0.05527260 & -0.22444237 & -0.97291858 \\ -0.62286470 & -0.76967727 & 0.21177325 \\ 0.79636420 & -0.59768096 & 0.09263653 \end{bmatrix}$$

As mentioned earlier, the d_i^2 are the eigenvalues of $\mathbf{X}'\mathbf{X}$ and the columns of $\mathbf{V}$ (rows of $\mathbf{V}'$) are the corresponding eigenvectors.

From these we compute

$$\kappa(\mathbf{X}) = d_{\max}/d_{\min} = 32.156334/0.37437558 = 85.89324656$$

and it is again seen that serious variance inflation may exist. Were an element of $\mathbf{X}$ to be changed by 1%, the variance of an estimator could be changed by as much as $2 \times 86 \times 100\% = 172\%$. The condition indices are

$$\eta_1 = 1$$

$$\eta_2 = 32.156334/2.1977332 = 14.63159131$$

and

$$\eta_3 = 32.156334/0.37437558 = 85.89324656$$

and we see that there is one moderate to strong relationship among the columns of $\mathbf{X}$.

The values relevant to the assessment of the A- and D-criteria are

$$\mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\} = 7.3428571 \text{ and } \det(\mathbf{X}'\mathbf{X}) = 700$$

The standard errors of the elements of **b** and the corresponding HLCIs for 95% confidence intervals (t (0.975; 2) = 4.3027)) are

SE (b_0) = 2.1447611, HLCI = 9.2280

SE (b_1) = 1.6344506, HLCI = 7.0324

SE (b_2) = 0.26726124, HLCI = 1.1499

The standard errors of the elements of $\hat{\mathbf{y}}$ and the corresponding HLCIs are

SE ($\hat{y}_1$) = 0.94112395, HLCI = 4.0493

SE ($\hat{y}_2$) = 0.60944940, HLCI = 2.6222

SE ($\hat{y}_3$) = 0.69693206, HLCI = 2.9986

SE ($\hat{y}_4$) = 0.60944940, HLCI = 2.6222

SE ($\hat{y}_5$) = 0.94112395, HLCI = 4.0493

This shows how **DESIGN** can be used to *evaluate* a given experimental design. It can also be used to *compare* designs. In the context of polynomial regression, one might envision using our program to compare designs in two distinct ways, e.g. to assess the effects of centering and/or scaling on the conditioning of the design matrix (computational considerations), and to see how a change in the design points might effect leverage, etc. (design considerations).

Consider first the effect of centering the data points, i.e. of using (Ref. 1, p. 300)

$$\dot{\mathbf{X}} = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}$$

instead of **X**. It should be noted here that 'centering' is used in (at least) two different ways in the literature. The first, appropriate for polynomial models, is illustrated above. There the values 1–5 are mean-centered (the mean of the column, 3, is subtracted from each entry) to produce the second column of **X**; these values are squared to obtain the third column. In the second kind of centering, each of the columns is centered by subtracting from each value in column j *the mean for that column*, viz., $z_{ij} = x_{ij} - \bar{x}_j$. In the context of the above example, this would result in the centered Z matrix (the mean of the third column is 11) (Ref. 2, p. 262)

$$\mathbf{Z} = \begin{bmatrix} -2 & -10 \\ -1 & -7 \\ 0 & -2 \\ 1 & 5 \\ 2 & 14 \end{bmatrix}$$

Note that $\mathbf{Z}$ is 5×2. The reason for this is that the intercept is no longer included in the model: The y-values are also centered when this approach is used, and this leads to $b_0 = \bar{y}$. See Ref. 2 (p. 260) for details.

Using the first (polynomial) form of centering, we have $\mathbf{R} = \mathbf{I}$ (columns 2 and 3 are orthogonal), $\mathbf{R}^{-1} = \mathbf{I}$ and so $\mathrm{VIF}_1 = \mathrm{VIF}_2 = 1$.

$$(\dot{\mathbf{X}}' \dot{\mathbf{X}})^{-1} = \begin{bmatrix} 0.48571429 & 0 & -0.14285714 \\ & 0.10 & 0 \\ & & 0.071428571 \end{bmatrix}$$

The condition number is reduced from $\kappa(\mathbf{X}) = 85.89$ to $\kappa(\dot{\mathbf{X}}) = 4.44$. The determinant is unchanged, but the trace is reduced to 0.657. The hat matrix is unchanged.

Scaling for equal column length of unity has been claimed to be 'nearly optimal' in the sense of minimizing the condition number of a matrix by (Ref. 4, p. 184). We again need to note that 'scaling' can have different meanings. Most often scaling refers to dividing by the standard deviation, but scaling for equal column length is also often used. To scale $\mathbf{X}$ to have column length 1, compute the sums of squares of each column and multiply $\mathbf{X}$ by the diagonal matrix with diagonal entries the reciprocals of the square roots of these numbers, i.e. compute

$$\ddot{\mathbf{X}} = \mathbf{X} \begin{bmatrix} \frac{1}{\sqrt{5}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{55}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{979}} \end{bmatrix}$$

This results in the scaled design matrix

$$\ddot{\mathbf{X}} = \begin{bmatrix} 0.44721360 & 0.13483997 & 0.031960139 \\ 0.44721360 & 0.26967994 & 0.12784056 \\ 0.44721360 & 0.40451991 & 0.28764125 \\ 0.44721360 & 0.53935988 & 0.51136222 \\ 0.44721360 & 0.67419985 & 0.79900347 \end{bmatrix}$$

This has $\kappa(\ddot{\mathbf{X}}) = 25.537210$, which is considerably smaller than $\kappa(\mathbf{X})$. The **H** matrix is the same as for **X**. The diagonals of the **D** matrix are $d_{11} = 1.660296$, $d_{22} = 0.46913664$ and $d_{33} = 0.065239296$. Note that **X** matrices that differ from one another only by the scale assigned the columns are essentially equivalent model structures (Ref. 4, p. 120); yet scaling can drastically effect the conditioning of **X** and result in very different SVDs and condition numbers. Thus, it is necessary to standardize to equal length before condition indices can be meaningfully compared. This scaling transforms a matrix with mutually orthogonal columns, 'the standard of ideal data', into a matrix whose condition indices are all unity (Ref. 4, p. 120). It is also true that when all condition indices are unity, the matrix is orthogonal.

If one wished to consider altering the design by using

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 9 \\ 1 & 3.5 & 12.25 \\ 1 & 4.5 & 20.25 \\ 1 & 5 & 25 \end{bmatrix}$$

which places more emphasis on the higher values in the range from 1 to 5 than did the original **X**, the diagonals of the hat matrix are

$$h_{11} = 0.99228,\ h_{22} = 0.51814,\ h_{33} = 0.42705,\ h_{44} = 0.32824,\ h_{55} = 0.73428$$

and it is seen that, relative to $\mathbf{X}, \tilde{\mathbf{X}}$ has larger values of h_{11} and h_{22}, smaller values of h_{33}, h_{44} and h_{55}. Shifting to larger x-values has moved the smaller values ($x = 1$ and $x = 3$) further from the mean and increased their importance in predicting y_1 and y_2.

We also have $\det(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}) = 728.68750$ and $\mathrm{tr}\{(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\} = 6.8476713$, both of which are slightly worse than for **X**.

Discussion

We have concentrated on polynomial regression in illustrating our program. This emphasis was deliberate. While **DESIGN** can accomodate a wide variety of linear models, polynomial regression, where multicollinearity is in a sense self-induced, provides a ready supply of examples, and continues to be actively discussed — and debated — in the literature. It is hoped that our program, by allowing the direct comparison of possible reparameterizations of the original design matrix (e.g. centering and/or scaling) will provide some insight into these issues. However, our choice of examples should not be construed as indicating that this is the only area where controversy exists. Active debate about all forms of multiple regression analysis continues. An appreciation for some of the points of contention in this debate is available in the literature [23,36,37]. All three papers are accompanied by critiques

from a number of discussants and it is clear that sharp differences of opinion exist with respect to how best to approach multiple regression in general and, in particular, polynomial regression. These points of view are briefly summarized below.

Hocking [23] gives a review of developments in linear regression during the period 1959–1982. He considers that most of the progress has been in regression diagnostics, and in developing strategies for dealing with serious multicollinearities once discovered, including transformations [38,39], ridge regression [34,35], robust estimation [40–42], and variable elimination [43,44]. He recognized the special problems presented by polynomial models, and suggested that these might often be resolved by simply centering the data. He thought that, 'there is little disagreement as to its value in polynomial regression'. Snee seconded the importance of centering both so that the regression coefficients would be estimated within the range of the data and would be easy to interpret, but noted, 'This important characteristic of good statistical practice does not appear to be widely recognized'. He recommended both centering *and* scaling: Centering reduces the correlation between the terms in the model, and scaling makes the estimated regression coefficients directly comparable. He noted that several forms of centering and scaling were often used, viz.,

$$z_i = \frac{x_i - \bar{x}}{S_i} \tag{20}$$

$$z_i = \frac{x_i - (x_{i\,\min} + x_{i\,\max})/2}{(x_{i\,\max} - x_{i\,\min})/2} \tag{21}$$

and

$$z_i = \frac{x_i - \bar{x}}{\sqrt{n-1}S_i} \tag{22}$$

the latter being the form routinely used and recommended by Snee. In this form, $\mathbf{Z}'\mathbf{Z} = \mathbf{R}$, the correlation matrix, and the regression coefficients are often called *beta weights*. The transformation (Eqn. 22) is sometimes called the *correlation transformation* (Ref. 1, p. 378) Finally, Snee noted that irregular experimental conditions can produce multicollinearity, a point of interest to potential users of our program. Welsch, on the other hand, argued that centering could mask problems associated with ill-conditioning: 'If we wish to diagnose when the constant term is collinear with any of the others, we must use an uncentered variance inflation factor'.

This point was reiterated and expanded upon in Belsley [36]. There Belsley focused on centering regression data by taking deviations from the mean. While he conceded that centering might have certain uses, he argued that assessing the conditioning of the data was not included among these. He did, on the other hand, reiterate his earlier position [4] that the (uncentered) columns of $\mathbf{X}$ should be scaled to have unit length to obtain the most meaningful measure of the conditioning of the basic data. He noted that centered $\mathbf{X}$ matrices, $\dot{\mathbf{X}}$, will have $\kappa(\dot{\mathbf{X}}) < \kappa(\mathbf{X})$ and

indeed, in practice one expects $\kappa(\dot{\mathbf{X}}) \ll \kappa(\mathbf{X})$, so that mean-centering apparently 'helps'. He claimed, however, that $\kappa(\dot{\mathbf{X}})$ 'gives us information about the wrong problem... mean-centering typically removes from the data the interpretabitity that makes conditioning diagnostics meaningful'. Snee and Marquardt commented that centering removed only 'nonessential ill-conditioning', viz., that tied up with collinearites associated with the constant term (β_0), which they view as a nuisance parameter. They also stated that the importance of centering becomes especially clear when one considers fitting polynomial models.

Another set of problems may be described by noting that the model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ can always be rewritten in the form

$$\mathbf{y} = \mathbf{X}\beta + \epsilon = (\mathbf{XA})\,(\mathbf{A}^{-1}\beta) + \epsilon = \mathbf{Z}\gamma + \epsilon \tag{23}$$

where $\mathbf{A}$ is any non-singular matrix. One can then either estimate β directly from [4] or estimate γ ($\hat{\gamma} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$) and use $\mathbf{b} = \mathbf{A}\hat{\gamma}$. This opens up the possibility that $\mathbf{A}$ can be chosen so that $\mathbf{Z}$ is 'stable'. Smith and Campbell [37] argue that the resulting estimators should be equal and thereby criticize the use of *ridge regression* [34,35] — one of the alternatives available to investigators with highly collinear data — since this property does not hold. They note further that $\mathbf{A}$ can always be chosen such that $\mathbf{Z}'\mathbf{Z} = \mathbf{I}$, but argue that this cannot make the data 'more informative', and characterize the use of such an $\mathbf{A}$ as, 'the cosmetic extreme of orthogonalizing the data'. They suggest that the high correlations among the variables have simply been transformed into low variances on linear combinations of these variables. In the ensuing discussion, Thisted questioned the 'obvious' need for the invariance of $\mathbf{b}$. He noted that choosing $\mathbf{b}$ to minimize the mean squared error, $\mathrm{MSE}(\mathbf{b},\beta) = E[(\mathbf{b} - \beta)'(\mathbf{b} - \beta)]$ (cf. Eqn. 15), is not equivalent to choosing $\hat{\gamma}$ to minimize $\mathrm{MSE}(\hat{\gamma}, \gamma)$; rather the former is equivalent to choosing $\hat{\gamma}$ to minimize $E[(\hat{\gamma} - \gamma)'(\mathbf{A}'\mathbf{A})(\hat{\gamma} - \gamma)]$. These coincide only if $\mathbf{A}$ is orthogonal. He also showed that taking $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1/2}$ results in each γ_i being estimated with equal precision. Marquardt argued in favor of standardizing the predictor variables, claiming that polynomial regression coefficients are interpretable only when the predictor variables are centered. He also noted that centering reflects the prior belief that the effects of the predictor variables, over their actual ranges in the data, are of comparable magnitude.

Centering, however, may not be *completely* effective in reducing multicollinearity in polynomial regression models. It is shown in Bradley and Srivastava [45] that, for centered x-values, $r(x^a, x^b) = 0$ when $a + b$ is odd, but this correlation may remain large when $a + b$ is even. They therefore suggest the use of orthogonal polynomials, which ensures that *all* such correlations are zero. Recall that using orthogonal polynomials is A-optimal. Seber (Ref. 3, p. 58) shows that for x_{ij} centered and scaled so that

$$\sum_i x_{ij} = 0 \text{ and } \sum_i x_{ij}^2 = c$$

$$\frac{1}{p} \sum_{j=0}^{p-1} V(b_j)$$

is minimized when the columns of **X** are orthogonal. Orthogonal design matrices for a given **X** (equally spaced x-values or not) can be generated by our program ORPOL [16].

It is even true that centering can affect tests of the form H: $\beta_k = 0$ [46]: The t-ratios for the coefficients of lower-order terms *change* when the data are centered (the t-ratio for the highest order term is unaffected). This argues in favor of the use of 'well-formulated' regression models [47], i.e. if a Dth degree polynomial is fit, the model includes terms for powers $1,2, \ldots, D-1$. Evidence that testing for lower order terms persists is given in Bernhardt and Jung [48]. A text book example where the square and cross-product terms are included in a model, but the first order term is dropped because it is not significant is cited in Griepentrog et al. [46]. It should also be realized that measures of the goodness-of-fit of a not-well-formulated polynomial regression model can be artificially raised or lowered by centering [47].

It is seen, then, that there is not complete agreement in how best to assess the goodness-of-fit of the general linear model, nor how one should proceed when multicollinearity is a problem. Some claim that centering is enough others advocate both centering and scaling; still others recommend scaling, but not centering. Our contention is that a program such as **DESIGN** may prove useful not only in comparing alternative choices in this context, but also in making decisions about such matters as where to position observations to satisfy certain requirements when resources are limited. It needs to be realized that many, if not all of the above notions are tied to the problem being solved, i.e. a matrix can be ill-conditioned with respect to one problem, but well-conditioned with respect to another. Our program may be of assistance in balancing the aims of a given study with what is possible from the practical standpoint.

It should be noted, however, that the program has some definite limitations. In particular, we do not provide for direct assessment of the goodness-of-fit of the model. That this is an important consideration is demonstrated below; however, most measures of the adequacy of the model require that the y-values be available and this is beyond the scope of the present paper. Consider, for example, the fact that $E(\mathbf{b}) = \beta$ holds only if the postulated model is correct. If the model is not *correct,* than the estimates are *biased.* The extent of the bias depends not only on the postulated and true models, but also on the values of the **X** variables. In the case of a designed experiment, then, the bias depends on the design.

If we postulate $\mathbf{y} = \mathbf{X}_1\beta_1$, and thus compute $\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$, if the true model is $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2$ then $E(\mathbf{b}_1) = \beta_1 + \mathbf{A}\beta_2$ where

$$\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\,\mathbf{X}_1'\mathbf{X}_2 = \mathbf{C}_1\mathbf{X}_2$$

where $\mathbf{C}_1$ denotes the catcher matrix for the postulated design. It is seen that the bias depends on the design through $\mathbf{X}_1$ and $\mathbf{X}_2$ (the matrix **A** is known as the *alias matrix* (Ref. 2, p. 118), and that the catcher matrix can be used to evaluate potential bias in the regression estimates. When $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ is orthogonal, $\mathbf{A} = \mathbf{0}$, but it can be substantial otherwise. Note that in the context of polynomial regression, choosing $\mathbf{X}_1$ is tantamount to choosing the degree of the polynomial to be fit (if $\mathbf{X}_1$ is $n \times p$, this implies a degree, D, of $p-1$). Choosing D too small will cause the estimates

b to be biased. Draper and Smith (Ref. 2, p. 119) give an example where $E(y) = \beta_0 + \beta_1 x$ is postulated, but $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$ is true. Using the design matrix

$$\mathbf{X}_1 = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

they show that $E(b_0) = \beta_0 + 2\beta_2/3$ and $E(b_1) = \beta_1$ so that b_1 is unbiased but b_0 is not. If a linear model is postulated but a cubic is correct (observations at -3, -2, -1, 0, 1, 2, 3), $E(b_0) = b_0 + 4\beta_2$ and $E(b_1) = \beta_1 + 7\beta_3$ Here both estimators are biased. This does not, however, mean that one should tend to 'overfit' to avoid bias. While choosing too small a degree does lead to bias, choosing a degree too large will necessarily — unless **X** is orthogonal — increase the variances of the estimators which are properly included in the model. Indeed, Bock (Ref. 49, p. 199) shows that the variances of the estimators will increase as additional terms are added (except if the new columns are orthogonal to those already included) no matter what the correct model is, and he gives an example where $V(\beta_1) = 0.00499\sigma^2/16$ when a line is fit, but $V(\beta_1) = 3.63742\sigma^2/16$ for a cubic equation, an increase by a factor of over 1000. The variances of the estimated values of the y values are also adversely effected when additional terms are entered into the model: The variance of $\hat{y}$ cannot decrease when another regressor is added to the model (Ref. 3, p. 138). Walls and Weeks [50] give an example where the variance is increased tenfold when the model is enlarged from a line to a quadratic.

Finally, we provide a brief indication of how all this fits in with our work in the polynomial modeling of longitudinai growth processes. The present discussion has focused on the situation where β was estimated by ordinary least squares (OLS), viz., $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, which is appropriate only when $V(\mathbf{y}) = \sigma^2\mathbf{I}$. When the y values are measured over time, we do not expect them to be uncorrelated and generally have to use weighted or generalized least squares (WLS) to estimate the parameters, viz., $\mathbf{b} = (\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{S}^{-1}\mathbf{y}$, where **S** reflects the correlation structure of the repeated measurements [12,13,15,17,18]. However, there are two important cases in which, given longitudinal data, the parameters of the model are properly estimated by OLS. One is the two-stage polynomial growth curve model [14]; the other the Potthoff-Roy model when their so-called arbitrary matrix is taken to be the identity [17]. In these cases the conditioning of $\mathbf{X}'\mathbf{X}$ is important, especially so since recent authors, e.g. (Ref. 51, p. 93) have reverted to the use of the successive powers-of-time form of the time design matrix (e.g., the **X** matrix in Eq. 19) on grounds of interpretability. This can be very poorly conditioned, especially for moderate to high degree polynomials like those used, e.g. in Ref. 52.

Acknowledgment

Supported by DE08730 from the National Institute for Dental Research.

Appendix: Computer Implementation

A full set of PC programs for longitudinal data analysis, including this program, can be obtained on 5.25" or 3.5" diskettes (please request type) by sending $25 to defray the cost of handling and licensing fees. These programs require a 80386 or 80486 based personal computer (PC) running the MS-DOS operating system (version 5.0 or higher is recommended, although versions as low as 3.3 will suffice). 80386 computers must also be equipped with a 80387 math coprocessor. At least 4 mb of memory is required, and must be available to GAUSS386i, i.e. not in use by memory resident programs such as Windows. EGA or VGA graphic capabilities are required to display the color graphics; VGA or SVGA is suggested to display optimally the graphic results. Runtime modules are supplied with the programs so that no additional software (i.e., compiler or interpreter) is required to run these programs. One can create and edit ASCH data sets for use by these programs using the full screen editor supplied with MS-DOS version 5.0. The programs are written and compiled using GAUSS386i, version 3.0, require no additional installation or modification, and are run with a single command. When requesting the programs, address inquiries to the corresponding author and make checks payable to Baylor College of Dentistry.

References

1 Neter J, Wasserman W and Kutner MH: *Applied Linear Statistical Moaels*, 2nd edn., Irwin, Homewood IL, 1985.
2 Draper N and Smith H: *Applied Regression Analysis*, 2nd edn., Wiley, New York, 1981.
3 Seber GAF: *Linear Regression Analysis*, Wiley, New York, 1977.
4 Belsley DA, Kuh E and Welsch RE: *Regression Diagnostics*, Wiley, New York, 1980.
5 Rao CR: *Linear Statistical Inference and Its Applications*,Wiley, New York, 1965.
6 Swindel BF: Instability of regression coefficients illustrated, *Am Stat*, 28 (1974) 63–65.
7 Wannacott RJ and Wannacott TH: *Econometrics*, 2nd edn., Wiley, New York, 1979.
8 Timm NH: *Multivariate Analysis with Applications in Education and Psychology*. Brooks/Cole, Monterey, CA, 1975.
9 Box MJ and Draper NR: Factorial designs, the [X'X] criterion and some related matters, *Technometrics*, 13 (1971) 731–742.
10 Steinberg DM and Hunter WG: Experimental design: Review and comment, *Technometrics*, 26 (1984) 71–97.
11 Meeker WQ, Hahn GJ and Feder PI: A computer program for evaluating and comparing experimental designs and some applications, *Am Stat*, 29 (1975) 60–64.
12 Schneiderman ED and Kowalski CJ: Implementation of Rao's one-sample polynomial growth curve model using SAS, *Am J Phys Anthropol*, 67 (1985) 323–333.
13 Schneiderman ED and Kowalski CJ: Implementation of Hills' growth curve analysis for unequal-time intervals using GAUSS, *Am J Hum Biol*, 1 (1989) 31–42.
14 Ten Have TR, Kowalski CJ and Schneiderman ED: PC program for analyzing one sample longitudinal data sets which satisfy the two-stage polynomiai growth curve model, *Am J Hum Biol*, 3 (1991) 269–279.
15 Schneiderman ED, Willis SM, Ten Have TR and Kowalski CJ: Rao's polynomial growth curve model for unequal-time intervals: A menu-driven GAUSS program, *Int J Biomed Comput*, 29 (1991) 235–244.
16 Ten Have TR, Kowalski CJ and Schneiderman ED: A PC program for obtaining orthogonal polynomial regression coefficients for use in longitudinal data analysis, *Am J Hum Biol*, 4 (1992) 403–416.
17 Ten Have TR, Kowalski CJ, Schneiderman ED and Willis SM: A PC program for performing multigroup longitudinal comparisons using the Potthoff-Roy analysis and orthogonal polynomials, *Int J Biomed Comput*, 30 (1992) 103–112.

18 Schneiderman ED, Willis SM, Kowalski CJ and Ten Have TR: A PC program for growth prediction in the context of Rao's polynomial growth curve model, *Comput Biol Med*, 22 (1992) 181–188.
19 Aptech: *The GAUSS System Version 2.0*, Aptech Systems, Inc, Kent WA, 1988.
20 Raveh A: On the use of the inverse of the correlation matrix in multivariate data analysis, *Am Stat*, 39 (1985) 39–42.
21 Snee RD: Discussion of [23], *Technometrics*, 25 (1983) 230–237.
22 Mansfield ER and Helms BP: Detecting multicollinearity, *Am Stat*, 36 (1982) 158–160.
23 Hocking RR: Developments in linear regression methodology: 1959–1982, *Technometrics*, 25 (1983) 219–230.
24 Velleman PF and Welsch RE: Efficient computing of regression diagnostics, *Am Stat*, 35 (1981) 234–242.
25 Hoaglin DC and Welsch RE: The hat matrix in regression and ANOVA, *Am Stat*, 32 (1978) 17–22.
26 Huber PJ: Robustness and designs. In: *A Survey of Statistical Design and Linear Models*, (Ed. JN Srivastava), North-Holland, New York, 1975.
27 Box GEP and Draper NR: Robust designs, *Biometrika*, 62 (1975) 347–352.
28 Good IJ: Some applications of the singular-decomposition of a matrix, *Technometrics*, 11 (1969) 823–831.
29 Mandel JM: Use of the singular value decomposition in regression analysis, *Am Stat*, 36 (1982) 15–24.
30 Nelder JA: An alternative interpretation of the singular-value decomposition in regression, *Am Stat*, 39 (1985) 63–64.
31 Eubank RL and Webster JT: The singular-value decomposition as a tool for solving estimability problems, *Am Stat*, 39 (1985) 64–66.
32 Chernoff H: Locally optimal designs for estimating parameters, *Ann Math Stat*, 24 (1953) 586–602.
33 St John RC and Draper NR: D-optimality for regression designs: A review, *Technometrics*, 17 (1975) 15–23.
34 Hoerl AE and Kennard RW: Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, 12 (1970) 55–67.
35 Hoerl AE and Kennard RW: Ridge regression: Applications to nonorthogonal problems, *Technometrics*, 12 (1970) 69–82.
36 Belsley DA: Demeaning conditioning diagnostics through centering, *Am Stat*, 38 (1984) 73–77.
37 Smith G and Campbell F: A critique of some ridge regression methods, *J Am Stat Assoc*, 75 (1980) 74–81.
38 Box GEP and Cox DR: An analysis of transformations, *J R Stat Soc*, Series B, 126 (1964) 211–246.
39 Box GEP and Tidwell PW: Transformations of the independent variables, *Technometrics*, 4 (1962) 47–67.
40 Andrews DF: A robust method for multiple linear regression, *Technometrics*, 16 (1974) 523–531.
41 Huber PJ: Robust statistics: A review, *Ann Math Stat*, 43 (1972) 1041–1067.
42 Huber PJ: Robust regresssion: Asymptotics, conjectures and Monte Carlo, *Ann Stat*, 1 (1973) 799–821.
43 Hocking RR: Criteria for selection of a subset regression: Which one should be used, *Technometrics*, 14 (1972) 967–970.
44 Hocking RR: The analysis and selection of variables in linear regression, *Technometrics*, 18 (1976) 425–438.
45 Bradley RA and Srivastava SS: Correlation in polynomial regression, *Am Stat*, 33 (1979) 11–14.
46 Griepentrog GL, Ryan JM and Smith LD: Linear transformations of polynomial regression models, *Am Stat*, 36 (1982) 171–174.
47 Peixoto JL: A property of well-formulated polynomial regression models, *Am Stat*, 44 (1990) 26–30.
48 Bernhardt I and Jung BS: The interpretation of least squares regression with interaction or polynomial terms, *Rev Econ Stat*, 61 (1979) 481–483.
49 Bock RD: *Multivariate Statistical Methods in Behavioral Research*, McGraw-Hill, New York, 1975.
50 Walls RC and Weeks DL: A note on the variance of a predicted response in regression, *Am Stat*, 23 (1969) 24–26.
51 Goldstein H: *The Design and Analysis of Longitudinal Studies*, Academic Press, New York, 1979.
52 Zerbe GO: A new nonparametric technique for constructing percentiles and normal ranges for growth curves determined from longitudinal data, *Growth*, 43 (1979) 263–272.