Journal of Research of the National Bureau of Standards
Volume 90, Number 6, November-December, 1985

The Regression Analysis of Collinear Data

John Mandel
National Bureau of Standards, Gaithersburg, MD 20899

Accepted: July 1, 1985

This paper presents a technique based on the intuitively-simple concepts of Sample Domain and Effective Prediction Domain, for dealing with linear regression situations involving collinearity of any degree of severity. The Effective Prediction Domain (EPD) clarifies the concept of collinearity, and leads to conclusions that are quantitative and practically useful. The method allows for the presence of expansion terms among the regressors, and requires no changes when dealing with such situations.

Key words: collinearity; efficient prediction domain; ill-conditioning; multicollinearity; regression analysis.

Introduction

The scientists' search for relations between measurable properties of materials or physical systems can be effectively helped by the statistical technique known as multiple regression. Even when limited to linear regression, the technique is often of great value, as we shall see below. Often, however, difficulties in interpretation arise because of a condition called collinearity. This condition, which is inherent in the structure of the design points (the X space) of the regression experiment, is often treated, at least implicitly, as a sort of disease of the data that is to be remedied by special mathematical manipulations of the data.

We consider collinearity not as a disease but rather as additional information provided by the data to the data analyst, warning him to limit the use of the regression equation as a prediction tool to specific subspaces of the X space, and telling him precisely what these subspaces are. Thus, collinearity is an indication of limitations inherent in the data. The statistician's task is to detect these limitations and to express them in a useful manner. If this viewpoint is adopted, there is no need for remedial techniques. All that is required is a method for extracting the additional information from the data. We will present such a method.

About the Author: John Mandel is a statistical consultant serving with NBS' National Measurement Laboratory.

The Model

We assume that measurements y have been made at a number of "x-points," each point being characterized by the numerical values of a number of "regressor-variables" x_j. We also assume that y is a linear function of the x-variables. The mathematical model, for p regressors, is:

$$y=\beta_1x_1+\beta_2x_2+\ldots+\beta_jx_j+\ldots+\beta_px_p+\epsilon \qquad (1)$$

where ϵ is the error in the y measurement. We denote by N the number of points, or "design points", i.e., the combinations of the x's at which y is measured.

Usually, the variable x_1 *is identically equal to "one"* for all N points, to allow for the presence of a constant term. Then the expected value of y, denoted $E(y)$, is equal to β_1 when all the other x's are zero. This point, called the origin, is seldom one of the design points and is, in fact, quite often far removed from all design points. In many cases this point is even devoid of physical meaning.

First Example: Firefly Data

We present the problem in terms of two examples of real data. The first data set (Buck [1][1]) is shown in table 1. It consists of 17 points and has two regressors, in addition to

[1] Figures in brackets indicate literature references.

Table 1. Data for firefly study.

x_1	x_2	x_3	y
1	26	21.1	45
1	35	23.9	40
1	40	17.8	58
1	41	22.0	50
1	45	22.3	31
1	55	23.3	52
1	55	20.5	54
1	56	25.5	38
1	70	21.7	40
1	75	26.7	28
1	79	25.0	38
1	87	24.4	36
1	100	22.3	36
1	100	25.5	46
1	110	26.7	40
1	130	25.5	31
1	140	26.7	40

Definition of Variables
y=time of first flash (number of minutes after 6:30 p.m.)
x_2=light intensity (in metercandles, mc)
x_3=temperature (°C)

a constant term ($x_1 \equiv 1$). The measurement is the time of the first flash of a firefly, after 6:30 p.m. It is studied as a function of ambient light intensity (x_2) and temperature (x_3).

Figure 1 is a plot of x_3 versus x_2. There is obviously a trend: x_3 increases as x_2 increases. The existence of a relation of this type between some of the regressor variables often causes difficulties in the interpretation of the regression analysis. To deal with the problem in a general way we propose a method based on two concepts. The first of these we shall call the "sample domain."

For our data, the sample domain consists of the rectangle formed by the vertical straight lines going through the lowest and highest x_2 of the experiment, respectively, and by the horizontal straight lines going through the lowest and highest x_3, respectively (See Fig. 1). The concept is readily generalized to an X space of any number of dimensions, and becomes a hypercube in such a space. Note that the vertex B of the sample domain is relatively far from any of the design points. This has important consequences.

The regression equation

$$\hat{y}=\hat{\beta}_1\cdot x_1+\hat{\beta}_2\cdot x_2+\hat{\beta}_3\cdot x_3 \qquad (2)$$

allows us to estimate y at any point (x_1, x_2, x_3) (we recall that $x_1 = 1$) and to estimate the variance of $\hat{y}$ at this point. The point can be inside or outside the sample domain. Obviously the variance of $\hat{y}$, which we denote by Var ($\hat{y}$), will tend to become larger as the point for which the prediction is made is further away from the cluster of points involved in the experiment. Therefore Var ($\hat{y}$) at the point B may be considerably larger than at points A, C, and D. Such a condition is associated with the concept of "*collinearity*." We define collinearity, in a semi-quantitative way, as the condition that arises when for at least one of the vertices of the sample

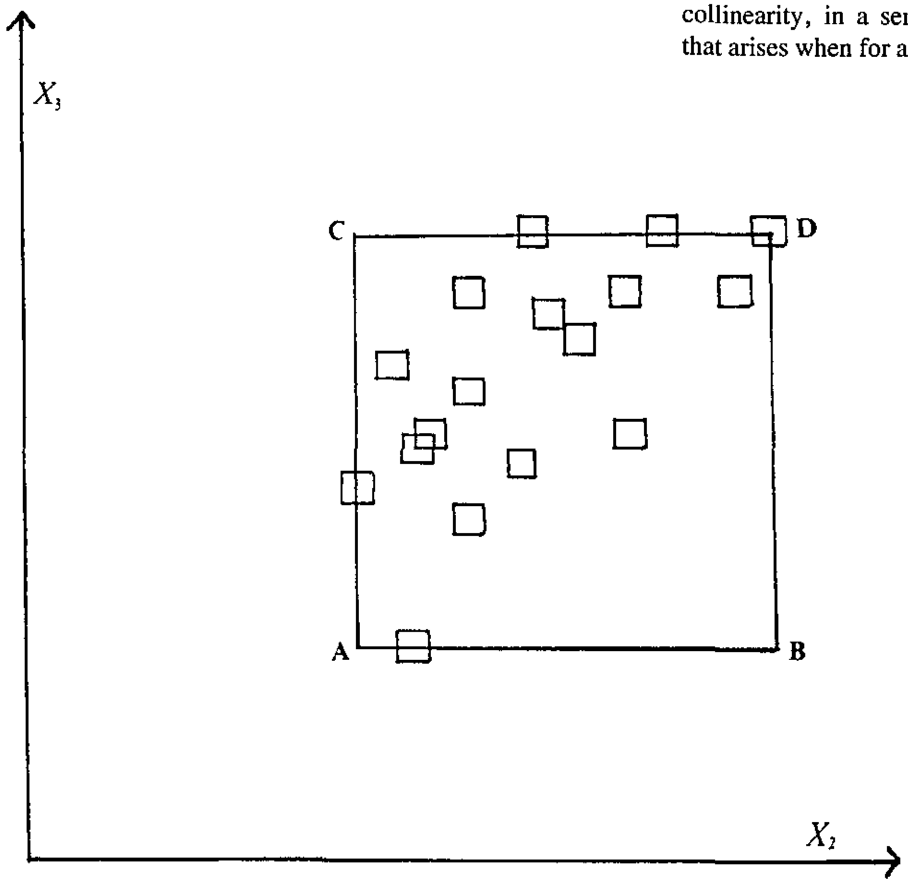


Figure 1—Sample domain.

domain, Var ($\hat{y}$) is considerably larger than for the other vertices. The concept will become clearer as we proceed.

At any rate, the larger variance at one of the vertices of the sample domain is generally the lesser of two concerns, the other being that the regression equation, for which validity may have been reasonably firmly established in the vicinity of the cluster of experimental points, may no longer be valid at a more distant point. It is important to note that the evidence from the data alone cannot justify inferences at such distant points. In order to validate prediction at such points, it is necessary to introduce either additional data or additional assumptions.

For these reasons, we seek to establish a region in the X-space for which prediction is reasonably safe *on the basis of the experiment alone*. We call this the *Effective Prediction Domain*, or EPD.

The EPD is the second concept required for our treatment of collinear data. It is closely related to the first concept, the sample domain, as will be shown below.

Establishing the EPD

Our procedure consists of two steps, involving two successive transformations of the coordinate system. The original coordinate system in which the x-regressors are expressed is referred to as the *X-system*.

1. The Z System

The first step consists in a *translation* of the X-system (parallel to itself) to a different origin, located centrally within the cluster of experimental points (*centering*); and simultaneously by a *rescaling* of each x to a standard scale. The new system, called the *Z-system*, is given by the equations[2]

$$\text{For } j=1\text{: } z_1=K \text{ (a constant)} \tag{3a}$$

$$\text{For } j>1\text{: } z_j=\frac{x_j-C_j}{R_j} \tag{3b}$$

For C_j and R_j we consider two choices, which we call the Correlation Scale Transformation (CST) and the Range Midrange Transformation (RMT). We discuss first the Correlation Scale Transformation defined by the choice

$$C_j=\bar{x}_j,\ R_j=\sqrt{\sum_i (x_{ij}-\bar{x}_j)^2} \tag{4}$$

where $i=1$ to N.

It easily follows from (3b) that

$$\bar{z}_j=0,\ \sum_i z_{ij}^2=1 \tag{5}$$

It is then reasonable to choose a value K in (3a) equal to

$$K=1/\sqrt{N} \tag{6}$$

so as to make $\sum_i z_{i1}^2=1$

The values of C_j and R_j for the firefly data are given in table 2. Contrary to statements found in the literature (see discussion at end of this paper), the centering and rescaling defined by the Correlation Scale Transformation have no effect whatsever on collinearity. The location of the sample domain relative to the design points remains unchanged, though it is expressed in different coordinates.

To arrive at an EPD, a second operation is necessary, viz. a *rotation* of the Z-coordinate system to a new coordinate system, which we shall call the *W-system* (of coordinates).

2. The W-System

The rotation from Z to W is accomplished by the method of Principal Components, or its equivalent, the Singular Value Decomposition (SVD). For a discussion of this method the reader is referred to Mandel [2]. Here we merely recall a few facts. Each w-coordinate is a linear combination of all z-coordinates given by the matrix equation:

$$W=Z\,V \tag{7}$$

where V is an orthogonal matrix.

In algebraic notation, eq (7) becomes

$$w_{ik}=\sum_j z_{ij}v_{kj} \qquad \begin{array}{l} i=1 \text{ to } N \\ j=1 \text{ to } p \end{array} \tag{8}$$

where the v_{kj} are the elements of the V matrix. The v_{kj}, for a given k, are simply the direction cosines of the w_k axis with respect to the Z-system. Consequently,

$$\sum_j v_{kj}^2=1 \tag{9}$$

Table 2. Firefly data—parameters for correlation scale transformation.

j	C	R
1	0	4.123106
2	73.176471	135.264447
3	23.582353	10.073962

[2] We assume that in the X-system, the regressor x_1 is identically equal to *unity*, to allow for an independent term.

Since the rotation is orthogonal, any two distinct w-axes, say w_k and $w_{k'}$, are orthogonal and consequently:

$$\sum_j v_{kj} \cdot v_{k'j} = 0 \qquad \text{for } k \neq k' \qquad (10)$$

For the firefly data, the V matrix is shown in table 3, and the complete set of z and w coordinates is given in table 4.

Note that row 2, as well as column 1, in table 3 consists of the element "one" in one cell and zeros in all others cells. This is a consequence of the orthogonality of z_1 with respect to all z_j with $j > 1$. This orthogonality is in turn due to the nature of the Correlation Scale Transformation, as expressed by eq (4).

At the bottom of the w columns we find values labeled λ_j. They are simply the sums of squares of all w-values in that column.

$$\lambda_j = \sum_i w_{ij}^2 \qquad (11)$$

Table 3. Firefly data—V matrix.

	j		
k	*1*	*2*	*3*
1	0	.7071	.7071
2	1.000	0	0
3	0	−.7071	.7071

Table 4. Firefly data—z and w coordinates (CST).[1]

Point	z_2	z_3	w_1	w_3
1	−.3488	−.2464	−.4216	.0724
2	−.2822	.0315	−.1780	.2219
3	−.2453	−.5740	−.5800	−.2324
4	−.2379	−.1571	−.2800	.0572
5	−.2083	−.1273	−.2381	.0573
6	−.1344	−.0280	−.1156	.0753
7	−.1344	−.3060	−.3121	−.1213
8	−.1270	.1904	.0440	.2245
9	−.0235	−.1869	−.1495	−.1155
10	.0135	.3095	.2276	.2094
11	.0431	.1407	.1292	.0691
12	.1022	.0812	.1289	−.0148
13	.1983	−.1273	.0495	−.2302
14	.1983	.1904	.2741	−.0055
15	.2722	.3095	.4106	.0264
16	.4201	.1904	.4309	−.1624
17	.4940	.3095	.5674	−.1304
			$\lambda_1 = 1.6549$	$\lambda_3 =$ .3451

[1] $z_1 = 1/\sqrt{17} = .2425$ for all i
$w_2 = 1/\sqrt{17} = .2425$ for all i, $\lambda_2 = 1.0000$

The λ_j are also the *eigenvalues* of the Z'Z matrix which, for our choice of C_j and R_j, is the correlation matrix of the regressors x. Note that w_2 is the constant$=1/\sqrt{N}$. Consequently

$$\lambda_2 = N\left(\frac{1}{\sqrt{N}}\right)^2 = 1 \ .$$

We need to consider w_1 and w_3 only. A similar situation applied to the z coordinates, where $z_1 = 1/\sqrt{N}$ for all i. Figure 2 shows both the z-coordinates (z_2 and z_3) and the w-coordinates (w_1 and w_3) for the firefly data. The order of the w-coordinates (w_1, w_2, w_3) is that of the corresponding λ-values, in decreasing order.

3. The Effective Prediction Domain (EPD)

The EPD is simply the sample domain corresponding to the W-system of coordinates. Thus, straight lines parallel to the w_3-axis are drawn through the smallest and largest w_1, respectively, and lines parallel to the w_1-axis are drawn through the smallest and largest w_3. Here again generalization is readily made to a p-dimensional W-space. The EPD for the firefly data is also shown in figure 2.

The interpretation of EPD is straightforward. Unlike the sample domain in either the X-system or the Z-system, the EPD excludes points that are distant from the cluster of regressor points. This has two advantages. In the first place, the use of the regression equation is justified for all points inside, and on the periphery of the EPD. And accordingly, the variance of the predicted value $\hat{y}$ for any such point will not be unduly large. These statements require more detailed treatment. To this effect we introduce the concept of *variance factor* (VF).

4. The Variance Factor (VF)

From regression theory we know that the variance of any linear functon, say L, of the coefficient estimates $\hat{\beta}_j$ is of the form:

$$\text{Var}\ (L) = f(X) \cdot \sigma_\epsilon^2 \qquad (12)$$

where σ_ϵ^2 is the variance of the experimental errors ϵ of the y measurements. The multiplier $f(X)$ is independent of the y and depends only on the X matrix and on the coefficients in the L function. We call this multiplier the *variance factor*, VF.

Thus, we have:

$$\text{Var}\ (\hat{\beta}_j) = \text{VF}(\hat{\beta}_j) \cdot \sigma_\epsilon^2 \qquad (13)$$

and

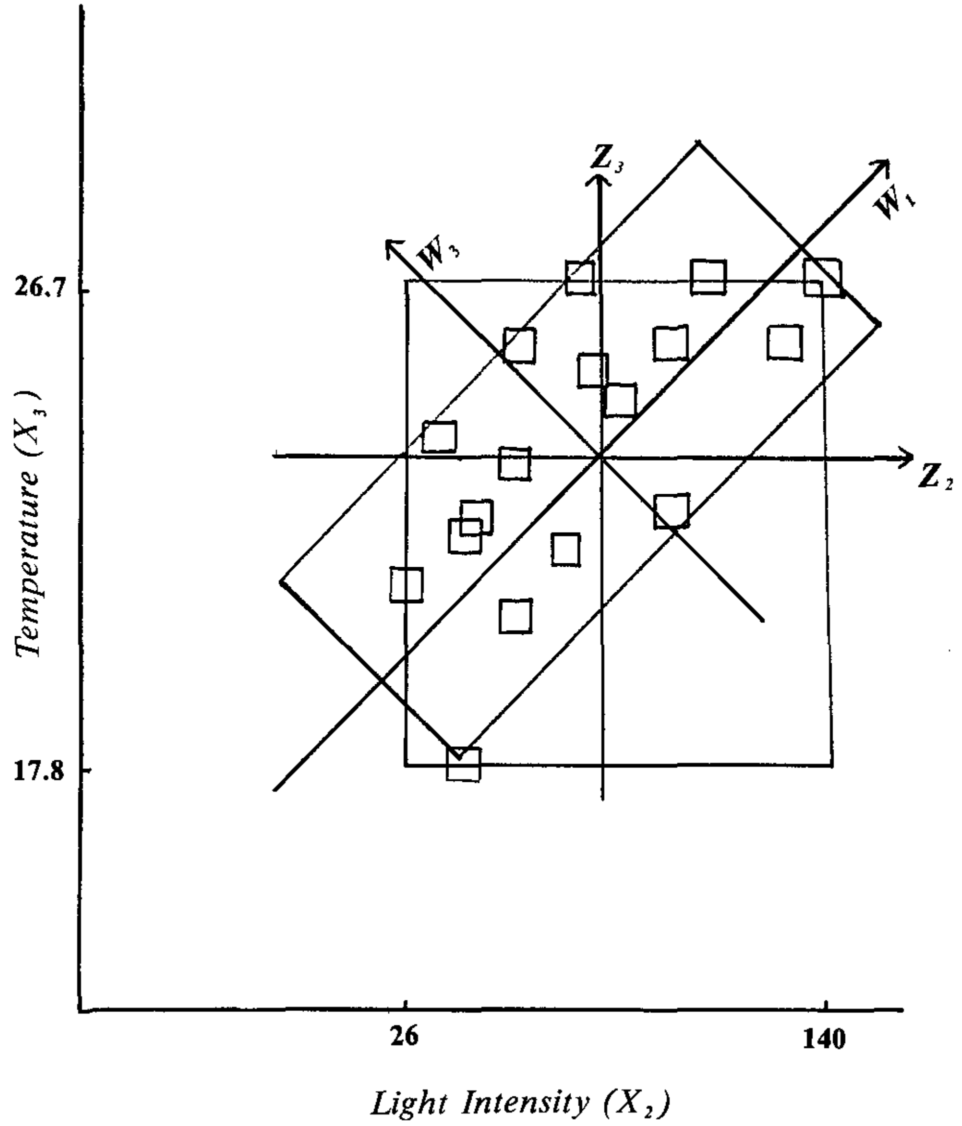


Figure 2—EPD for firefly data.

$$\text{Var}\ (\hat{y})=\text{VF}(\hat{y})\cdot\sigma_\epsilon^2 \qquad (14)$$

In eq (14), $\hat{y}$ is the estimated, or *predicted* y value at any chosen point in X-space. VF $(\hat{y})$ is of course a function of the location of this point.

Returning now to our statements above, it is well-known that a regression equation can show excellent (very small) residuals and yet be very poor for certain prediction purposes. The small residuals merely mean that a good fit has been obtained *at the points used in the experiment*. This is no guarantee that the fit is good at other points. However, if the regression equation is scientifically reasonable, it is likely that the experimental situation underlying it will also be valid for points *that are close* to the cluster of the regressor points used in the experiment. Every point in the EPD satisfies this requirement.

Furthermore, the variance of prediction, measured by the VF, will also be reasonably small for all points of the EPD, simply because they are geometrically close to the design points.

The calculation of VF $(\hat{y})$ is quite simple, once the V-matrix and the λ values have been calculated. It is based on the equation

$$\text{VF}(\hat{y})=\sum_k u_k^2 \qquad (15)$$

where u_k is defined as:

$$u_k=\frac{w_k}{\sqrt{\lambda_k}} \qquad (16)$$

Combining eqs (8) and (16), we obtain

$$u_{ik}=\sum_j z_{ij}\frac{v_{kj}}{\sqrt{\lambda_k}} \qquad (17)$$

and hence:

$$\mathrm{VF}(\hat{y}) = \sum_{k} \frac{\left(\sum_{j} z_{ij} v_{kj}\right)^2}{\lambda_k} \tag{18}$$

Figure 3 shows the VF values at the vertices of the original sample domain and of the EPD. Interpreting these results, we see that the collinearity of our data is reflected in the rejection of an appreciable portion of the sample domain for purposes of safe prediction. This does *not* mean that prediction outside the EPD is impossible, or unacceptable. It merely means that such prediction cannot be justified on the basis of the data alone. Of course, the risk of predicting outside the EPD increases with the distance from the EPD. It will generally be reasonably safe to use the regression equation even outside the EPD, as long as the point for which prediction is made is reasonably close to the borders of the EPD. Using eq (18), the VF for any contemplated prediction point is readily calculated and can serve as a basis for decision.

Second Example: Calibration for Protein Determination

The instructive and intuitively satisfying graphical display of the EPD becomes impossible when the number of regressors, including the independent term, exceeds 3. We must then replace the graphical procedure by an analytical one, as will now be shown in the treatment of our second example.

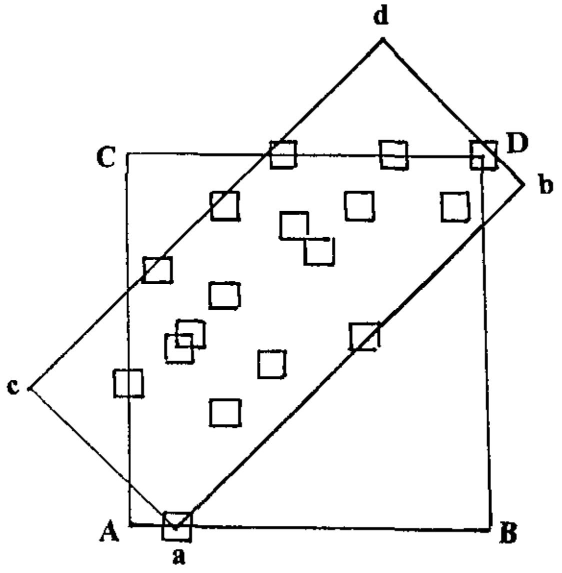
The data were presented by Fearn [3], in a discussion of Ridge Regression. They represent the linear regression of percent protein, in ground wheat samples, on near-infrared reflectance at six different wavelengths.

For reasons of simplicity in presentation, we include here only three of the six wavelengths, a change that has a rather small effect on the final outcome of the analysis: it turns out that the regression equation based on these 3 wavelengths is very nearly as precise as that based on 6 wavelengths.

The data, displayed in table 5, are a very good example of the use of regression equations: the regression equation is indeed to be used as a "calibration curve" for the analysis of protein, using the rapid spectrometry instead of the far more time-consuming Kjeldahl nitrogen determination. Our data have an N value of 24, and p (including the independent term) is 4.

Table 6 exhibits the correlation matrix of the 24 design points. It is very apparent that the x values at all three wavelengths are highly correlated with each other, thus indicating a high degree of collinearity. At a first glance one would be very skeptical about such a set of data, and suspect that the X matrix shows such a high degree of redundancy as to make the regression useless for prediction purposes. Fearn explains that the correlations are more a reflection of particle size variability than of protein content. Our analysis will confirm that, properly interpreted, the data lead to a very satisfactory calibration procedure.

We will find it useful to introduce a slightly different Z transformation, which we call the *Range-Midrange Transformation*.



Sample Domain

Vertex	VF
A	.39
B	1.71
C	.69
D	.30

EPD

Vertex	VF
a	.41
b	.41
c	.41
d	.40

Figure 3—VF at vertices of sample domain and of EPD.

Table 5. Protein Calibration Data(*)

Point	Reflectance			% Protein
	x_2	x_3	x_4	y
1	246	374	386	9.23
2	236	386	383	8.01
3	240	359	353	10.95
4	236	352	340	11.67
5	243	366	371	10.41
6	273	404	433	9.51
7	242	370	377	8.67
8	238	370	353	7.75
9	258	393	377	8.05
10	264	384	398	11.39
11	243	367	378	9.95
12	233	365	365	8.25
13	288	415	443	10.57
14	293	421	450	10.23
15	324	448	467	11.87
16	271	407	451	8.09
17	360	484	524	12.55
18	274	406	407	8.38
19	260	385	374	9.64
20	269	389	391	11.35
21	242	366	353	9.70
22	285	410	445	10.75
23	255	376	383	10.75
24	276	396	404	11.47

(*)$x_1 \equiv 1$

Table 6. Protein calibration data—correlation matrix of x_1 through x_4.

1	0	0	0
	1	.9843	.9337
		1	.9545
			1

The Range-Midrange Transformation

The Range-Midrange Transformation (RMT) is defined as follows:

$$\text{For } j=1: \quad z_1 = 1 \tag{19a}$$

$$\text{For } j>1: \quad z_j = \frac{x_j - C_j}{R_j} \tag{19b}$$

but now C_j is defined as the *midrange* of the N values of x_j and R_j is *one-half the range* of these values. With these definitions, it is clear that the smallest z-value, for any regressor, is (-1) and the largest z-value is $(+1)$. It is because of this -1 to $+1$ scale that this transformation was introduced. The benefits of this scale will become apparent in the following section.

EPD for the Protein Data

The EPD resulting from the Singular Value Decomposition based on the Range-Midrange Transformaton will not be he same as the EPD we would have obtained using the Correlation Scale Transformation, but we will see that those features of the EPD that are of importance for us, in establishing the limitations of the regression equation, are practically unaffected.

Table 7 shows the C and R values for the four regressors and table 8 exhibits the V matrix and the λ values obtained from the Singular Value Decomposition. The latter, it may be recalled, simply expresses the rotation of the Z coordinate system to the W system.

For each w_k coordinate, there are 24 values, corresponding to the 24 regressor points.

Table 9 shows the smallest and the largest w_k value, for each of the four k.

According to table 9, we must have, in the EPD:

$$-1.9282 \leqq w_1 \leqq .6181 \tag{20}$$

with similar statements for w_2, w_3, and w_4. Applying now eq

Table 7. Protein calibration data—parameters for Z transformation (RMT).

j	C	R
1	0	1
2	296.5	63.5
3	418.0	66.0
4	432.0	92.0

Table 8. Protein calibration data—V matrix and λ values (RMT).

k	1	2	3	4	λ
1	−.6665	.4845	.4217	.3784	43.7810
2	.7365	.3299	.3797	.4523	8.3782
3	−.1096	−.5491	−.2509	.7896	.3758
4	−.0332	−.5958	.7843	−.1698	.06624

Table 9. Protein calibration data—limits defining the EPD.

Coordinate (k)	*Smallest w*	*Largest w*
1	−1.9282	.6181
2	−.4097	1.8989
3	−.1669	.3158
4	−.0801	.1324

(8), this double inequality can be written:

$$-1.9282 \leqq -.6665\ z_1 + .4845\ z_2 + .4217\ z_3 + .3784\ z_4 \leqq .6181$$

Since z_1 is constant and $=1$, this double inequality becomes:

$$-1.2617 \leqq .4845\ z_2 + .4217\ z_3 + .3784\ z_4 \leqq 1.2846\ . \quad (21a)$$

With the RMT, the value of any z_k is, for any $k>1$, between (-1) and $(+1)$. Thus the expression in the middle has, for all design points, a value between -1.2846 and 1.2846, where 1.2846 is the sum of the absolute values of the three coefficients. Therefore, the double inequality expressed by eq (21a) holds, essentially, for every point in the original sample domain. Thus, w_1, the first coordinate of the EPD, which represents its largest dimension, imposes essentially no restrictions on the sample domain.

Doing the same calculations for the three other w-coordinates (see table 9), we obtain, respectively:

$$-1.1462 \leqq .3299\ z_2 + .3797\ z_3 + .4523\ z_4 \leqq 1.1619 \quad (21b)$$

$$-0.568 \leqq -.5491\ z_2 - .2509\ z_3 + .7896\ z_4 \leqq .4254 \quad (21c)$$

$$-.0469 \leqq -.5958\ z_2 + .7843\ z_3 - .1698\ z_4 \leqq .1656. \quad (21d)$$

We see that w_2 too, imposes only very light restrictions on the sample domain. On the other hand, w_3 and w_4 do imply limitations that eliminate appreciable portions of the sample domain from the EPD.

We could readily convert eqs (21c) and (21d) to x coordinates by means of table 7 and eqs (19a) and (19b), but the z-coordinates, using the Range-Midrange Transformation, are more readily interpreted in terms of the severity of collinearity than the x-coordinates.

Thus, the sum of the absolute values of the coefficients in the middle terms of (21c) and (21d) are 1.5896 and 1.5499, respectively. Points for which these linear combinations take the valves ± 1.5896 and ± 1.5499 exist in the original sample domain. The EPD, on the other hand, limits these functions to intervals with much narrower limits.

Effect of Type of Z Transformation

We have used two different Z transformations, the Correlation Scale, and the Range-Midrange. It is proper to ask how our results would have been affected in the Protein Calibration Data, had we used Correlation Scale, instead of the Range-Midrange Transformation. We show the comparison in table 10. Let us recall that with the CST, one of the w coordinates yields a λ-value of unity, and a constant w value for all points. Therefore, we obtain for CST, only three sets of inequalities, as compared to the four sets for RMT. To allow the comparison between the two transformation to be made, we have multiplied eqs (21a) through (21d) by positive constants, so as to make the coefficient of z_4 equal to ± 1. The same was done for the corresponding inequalities obtained by the Correlation Scale Transformation.

Table 10. Protein calibration data—effect of Z transformation.[1]

w coordinate	Z Transf.	Inequalities
1	CST	$-3.034 \leq 1.021\ z_2 + 1.061\ z_3 + z_4 \leq 3.082$
	RMT	$-3.334 \leq 1.280\ z_2 + 1.114\ z_3 + z_4 \leq 3.395$
2	CST	
	RMT	$-2.534 \leq .729\ z_2 + .840\ z_3 + z_4 \leq 2.569$
3	CST	$-.075 \leq -.686\ z_2 - .321\ z_3 + z_4\ .535$
	RMT	$-.072 \leq -.695\ z_2 - .318\ z_3 + z_4 \leq .539$
4	CST	$-.278 \leq -3.531\ z_2 + 4.640\ z_3 - z_4 \leq .980$
	RMT	$-.276 \leq -3.509\ z_2 + 4.619\ z_3 - z_4 \leq .975$

[1] All inequalities are expressed in RMT z coordinates.

Of course, since the z coordinates are different for the two transformations, the inequalities for the CST, expressed in the CST z-units, had to be converted to RMT z-units, for a meaningful comparison. As can be seen from table 10, the two smallest dimensions of the EPD are practically the same for the two transformations. Thus, even though the method of principal components is not invariant with respect to linear transformations of scale, our analysis leads, in this case, to very similar results for the small dimensions of the EPD. We believe that this is generally true for all situations in which collinearity is noticeable, i.e., for all situations in which the EPD eliminates considerable portions of the original sample domain. For situations in which this does not apply, i.e., totally non-collinear cases, the inequalities do not matter, since they impose no restrictions on the sample domain.

It is interesting to contrast the remarkable similarity between the inequalities for w_3 and w_4 for the two transformations in table 10, with the behavior of a commonly advocated measure of collinearity (Belsley, Kuh, and Welsch [4], the *condition-number*.

The SVD resulting from the CST yields the following eigenvalues: 2.9151, 1.0000, .07176, .01312. The condition number is defined as the ratio of the largest to the smallest eigenvalue. In this case:

$$\text{condition number} = 2.9151/.01312 = 222.2$$

On the other hand, the SVD resulting from the RMT on the same data yields the eigenvalues: 43.7810, 8.3782, .37575, .066244. This time we have:

condition number=43.7810/.066244=660.9 .

Thus the condition number varies considerably when the data are subjected to different standardizing transformations. It is not clear what useful information can be derived from the condition number.

By contrast, the treatment of collinearity we advocate has a useful and readily understood interpretation: the EPD is that part of the X space in which, and near which, prediction is safe. It also indicates what portions of the original sample domain are inappropriate for prediction *on the basis of the given data alone*. It fulfills this function in a way which is practically invariant with respect to *intermediate* transformations of scale. We use the qualifier "intermediate" because collinearity has meaning only in terms of a given original coordinate system (the X system). This system, which determines the original sample domain, must be considered fixed. On the other hand, transformations of this system prior to calculating the EPD can be defined in different ways without affecting the practical inferences drawn from the data on the basis of the final EPD derived form the standardizing transformation.

Cross-Validation

We can take advantage of the availability of a second set of protein calibration data, also given in Fearn [3]; to verify the correctness of our approach. Fearn lists 26 additional points for which the reflectance measurements, as well as the Kjeldahl nitrogen determination, were made. We applied the Z transformation obtained above (RMT on first set of 24 points) to each of these 26 points, and noted every point for which at least one of the four sets of inequalities (21a) through (21d) failed to be satisfied. We found 14 such points. This means that 14 "future points" obtained under the same test conditions were outside the EPD established on the basis of the original 24 points. However, as we observed above, as long as the point is not far from the EPD, prediction at that point is likely to be valid. We tested "predictability" at these 14 points by calculating the VF value for each of them, and by comparing the predicted protein value with the measured one. The results are shown in table 11. It is apparent that all VF are relatively small, indicating that even though these 14 points are outside the EPD calculated from the original set, they are not far from that EPD. This is confirmed by the good agreement between the observed and predicted values. The standard deviation of fit for the original set of 24 points was 0.23; the standard deviation for a single measurement derived from the 14 differences in table 11 is 0.30.

Expansion Terms

Quite frequently, a regression equation contains x variables that are non-linear functions of one or more of the other x variables, such as x_2^2, $x_2 \cdot x_3$, etc. Polynomial regressions are necessarily of this type. Since the x variables are non-stochastic in the usual regression models, the least squares solution for the regression equation is not affected by the presence of such "expansion terms." On the other hand, collinearity can be introduced, or removed, or modified by them.

Table 11. Protein calibration data—cross-validation of analysis.

Point[1]	% Protein Observed	 Predicted	VF
1	8.66	9.53	.281
4	11.77	11.97	.416
6	10.46	10.96	.193
9	12.03	11.47	.212
10	9.43	9.54	.762
11	8.66	8.15	.454
12	14.44	13.99	.881
14	10.41	10.17	.468
16	11.69	11.24	.472
17	12.19	11.83	.390
18	11.59	11.39	.314
20	8.60	8.39	.201
22	9.34	8.93	.151
26	10.89	10.94	.741

[1] Point in additional set (Fearn [3]) with its number designation in that set.

In our treatment the expansion terms cause no additional problems. Consider for example, the regression

$$y=\beta_1x_1+\beta_2x_2+\beta_3x_2^2+\epsilon \tag{22}$$

with $x_1 \equiv 1$.

Here we have $p=3$. Using RMT, followed by a singular value decomposition, we obtain an EPD of three dimensions, leading to the inequalities.

$$A_1 \leq w_1 \leq B_1,\ A_2 \leq w_2 \leq B_2,\ A_3 \leq w_3 \leq B_3\ . \tag{23}$$

Expressing the w as functions of the z, this leads to three double inequalities governing the z, of the form

$$A_1 \leq f_1(z) \leq B_1,\ A_2 \leq f_2(z) \leq B_2,\ A_3 \leq f_3(z) \leq B_3, \tag{23}$$

Now, since $x_3=x_2^2$, we have

$$z_3=\frac{x_3-C_3}{R_3}=\frac{x_2^2-C_3}{R_3}=\frac{(R_2z_2+C_2)^2-C_3}{R_3}\ .$$

Hence:

$$z_3=\frac{C_2^2-C_3}{R_3}\,z_1+\frac{2\,C_2R_2}{R_3}\,z_2+\frac{R_2^2}{R_3}\,z_2^2\ . \tag{24}$$

Because of this relation the functions $f_1(z), f_2(z), f_3(z)$ become functions of z_1, z_2 (and z_2^2) only. Using this fact, we interpret the three sets of inequalities (23) exactly as we have interpreted eqs (21a) through (21d) by determining which of these inequalities, if any, impose restrictions on the use of the original sample domain.

To illustrate this procedure, consider the small set of artificial data shown in table 12, for which the model is given at the bottom of the table. The term $x_3=x_2^2$ introduces a high correlation between x_2 and x_3 and consequently also considerable collinearity.

The inequalities characterizing the EPD based on a Range-Midrange Transformation and converted to the z-scales, are shown in table 13. Applying eq (24) to express z_3 in terms of z_2, the three double-inequalities become:

$$\begin{aligned}&\text{for } w_1{:}-.8431\leq 1.1284\ z_2+.2853\ z_2^2\leq 1.4137\\ &\text{for } w_2{:}-.6928\leq .8541\ z_2+.1612\ z_2^2\leq 1.0153\\ &\text{for } w_3{:}\quad .0077\leq .0003\ z_2+.3218\ z_2^2\leq .3221.\end{aligned}$$

It is readily verified that of these six inequalities, all but one are satisfied for all z_2 values between -1 and $+1$. The last one, involving the left side of the third set, is satisfied for all z_2 values except for the interval: $-.156\leq z_2\leq .155$. This corresponds to an x_2 interval between 2.1 and 2.8, or between the design points $x_2=2.1$ and $x_2=3.6$ (see table 12). The interpretation of this finding is that while all design points are of course inside the EPD, a small portion of the curve x_2^2 versus x_2 falls slightly outside the EPD. This is of no practical significance since the VF for these points, even though they are outside the EPD, does not exceed 0.58. By comparison, the smallest VF value along the curve, for the range $x_2=.2$ to $x_2=4.7$, is of the order of 0.26. Thus we see that the serious collinearity in this data set is merely a consequence of the presence of the expansion term $x_3=x_2^2$.

Table 12. An artificial quadratic example[1].

Point	x_2	x_3	y
1	.2	.04	28.3
2	.4	.16	27.5
3	1	1.00	25.6
4	2.1	4.41	28.7
5	3.6	12.96	46.4
6	4.7	22.09	69.8

[1] $y=\beta_1x_1+\beta_2x_2+\beta_3x_3+\epsilon$; $\beta_1=30$, $\beta_2=8$, $\beta_3=3.5$, $\sigma_\epsilon=0.2$ $x_1=1$.
Note that $x_3=x_2^2$.

Table 13. Quadratic example—inequalities for EPD.

w-coordinate	Inequalities
w_1	$-1.1281\leq -.5070\ z_2+.6214\ z_3\leq 1.1284$
w_2	$-.8541\leq .5029\ z_2+.3511\ z_3\leq .8541$
w_3	$-.3141\leq -.7005\ z_2+.7008\ z_3\leq .0003$

Any point in X space, in order to be acceptable, must lie on the curve $x_3=x_2^2$. An x_3 with any other value is obviously not valid and our analysis of the data, through the EPD, calls attention to this fact: in the direction of w_3, the width of the EPD is only .31 as compared with widths of 2.26 and 1.71 for w_1 and w_2.

Discussion

The common mathematical definition of collinearity is the existence of at least one linear relation between the x's, of the form

$$\sum_j c_j x_{ij}=0 \qquad j=1 \text{ to } p \tag{25}$$

where the c_j are not all zero, and such that eq (25) holds with the same c_j values, for all i. This defines what we shall call "exact collinearity." Geometrically, it means that all design points lie in an hyperplane of the x-space, going through the origin of the coordinate system. Equation (25) also implies that the matrix $X'X$ is singular, and consequently that the estimates of the β coefficients are not uniquely defined.

Exact collinearity seldom occurs in real experimental situations; indeed, if the X matrix is not the result of a designed experiment, it is highly improbable that a relation such as eq (25) would hold exactly. If, on the other hand, the experiment is designed, care would generally have been taken to avoid a situation of exact collinearity.

While *exact* collinearity is practically of little concern, near-collinearity is a frequent occurrence in real-life data. This occurs when an equation such as (25) is "approximately" true for all i. Many attempts have been made to define more closely the concept of near-collinearity, but while these endeavors have led to a number of proposals for *measuring* collinearity, they are of little practical use to the experimenter confronted with the task of interpreting his data.

It is not our intention to discuss here the pros and cons of the various attempts made by a number of authors to "remedy" a near-collinear situation. The best-known of these remedial procedures is Ridge Regression. We merely repeat what we have said in the body of the paper: any attempt to remedy collinearity must necessarily be based on additional assumptions, unless it consists of making additional measurements. The latter alternative is of course logical and valid, but the making of assumptions invented specifically for the purpose of removing collinearity does not appear to us to be a recommendable policy in data analysis.

One easily recognizable condition leading to collinearity is the existence of at least one high correlation coefficient among the non-diagonal elements of the correlation matrix of the x's. This has given rise to the *concept of the Variance Inflation Factor* (VIF). The VIF for $\hat{\beta}_j$ is defined (Draper and Smith [5]), as:

$$\mathrm{VIF}(\hat{\beta}_j)=\frac{1}{1-R_j^2} \qquad (26)$$

where R_j is the multiple correlation coefficient of x_j on all other regressors. If d represents a residual in this regression, the usual formula for R_j is given by

$$R_j^2=1-\frac{\Sigma d^2}{\sum_i (x_{ij}-\bar{x}_j)^2} \quad . \qquad (27)$$

Now, Snee and Marquardt (Belsley [6], "comments") make, implicitly, a distinction between the two "models":

$$y=\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\epsilon \qquad (28a)$$

with $x_1\equiv 1$, and

$$y-\bar{y}=\beta_2(x_2-\bar{x}_2)+\cdots+\beta_p(x_p-\bar{x}_p)+\epsilon \qquad (28b)$$

where (28b) is called the "centered" model. For (28b), Snee and Marquardt use eq. (27), but for (28a) they appear to use the definition:

$$R_j^2=1-\frac{\Sigma d^2}{\sum_i x_{ij}^2} \quad . \qquad (29)$$

Equation 29, in which the denominator of the last term is not centered, is not explicitly given by Snee and Marquardt, but is implied by their statement:

> "If the domain of prediction includes the full range from the natural origin through the range of the data, then collinearity diagnostics should not be mean-centered," and confirmed by the VIF values given in their table 1. In this table, "no centering" results in VIF values of 200,000 and 400,000, while the VIF for the "centered" data are unity. The quoted statement occurs in a section entitled "Model building must consider the intended or implied domain of prediction." The basic idea underlying the section in question is that the analysis of the data, based on the "collinearity diagnostics" (specifically: the VIF values), is goverened by the location of the points were one *wishes* to make predictions and, more specifically, on whether the origin ($x_1=1$, $x_2=x_3\cdots=0$) is such a point. The VIF values which, according to Snee and Marquardt's formulas, depend heavily on whether or not this origin is included, will then indicate the quality of the predicted values.

A more reasonable approach, and one more consistent with the procedures commonly used by scientists, is to limit prediction to the vicinity of where one made the measurements, *unless additional information is available* that justifies extrapolation of the regression equation to more distant points of the samples space. The vicinity of the measured points is determined by the EPD which, in the case of collinearity, may be considerably smaller than the sample domain. In this view, it is the location of the *design points*, rather than that of the *intended points of prediction*, that determines predictability. The latter is measured, not by VIF values, but rather by the more concrete VF values, for any desired point of prediction.

The view advocated by Snee and Marquardt sometimes results in an enormous difference in the VIF values between the centered and non-centered forms. Equation 29 serves no useful purpose and is, in fact, unjustified and misleading. It is unjustified because it not only includes the origin ($x_1=1$, $x_k=0$ for $k>1$) in the correlation and VIF calculations, but moreover, gives this point infinite weight in these calculatons. Yet, no measurement was made at that point. Equation 29 is also misleading because it leads to very large VIF values for some non-centered regressions, implying that severe "ill-conditioning" exists, even when the X matrix is except for some trivial coding, completely orthogonal (cf. [6]).

The ill-conditioning exists only in terms of the large VIF value. It is an artifact arising from the desire to make the two forms of the regression equation into two distinct "models".

The two forms, eqs 28a and 28b lead to identical estimates for the β_j, including β_1, and for their standard errors. They also lead to identical values and variances for an estimated (predicted) $\hat{y}$, at any point of the X space. There seems to be no valid reason for the two distinct equations for the VIF. They only lead to the false impression that centering can reduce or even remove collinearity.

Our viewpoint in this paper is that the usefulness of a regression equation lies in its abilty to "predict" y for interesting *combinations of the x's*. We also take the position that inferences *from the data alone* should be confined to x points that are in the general geometric vicinity of the cluster of design points. An inference for points that are well outside this domain (i.e., outside a suitably defined EPD) is, in the absence of additional information, only a tentative conclusion, and not a valid scientific inference. Such conclusions may however, be very useful, provided their tentative character is recognized, and provided they are subsequently subjected to further experimental verification.

Daniel and Wood [7] discuss briefly the relation between the variance of $\hat{y}$ and the location of the point at which the prediction is made. However, their discussion is in the context of selecting the best subset of regressors from among the entire set of regressors, a subject different from the one dealt with in this paper.

Another publication that deals explicitly with predictability is a paper by Willan and Watts [8]. These authors define a "Region of Effective Predictability" (REP_A) as that portion of the X space in which the variance of the predicted $\hat{y}$ does not exceed twice the variance of $\hat{y}$ predicted at the centroid of the X matrix. The volume of the region is then compared with that of a similarly defined REP, denoted REP_0. The latter refers to a "fictitious orthogonal reference design" of "orthogonal data with the same N and the same rms values as the actual data." The ratio of the volume of REP_A to that of REP_0 is taken as "an overall measure of the loss of predictability volume due to collinearity".

This concept, apart from its artificial character, suffers from other shortcomings. Like so many other treatments, it attempts to provide a *measure of collinearity*. But the practitioner who is confronted with a collinear X matrix does not need a measure of collinearity: he needs a way to use the data for the purpose for which they were obtained. Furthermore, this measure loses its meaning when expansion variables are present. For example, for the artificial quadratic set of table 12, Willan and Watts' measure would indicate a high degree of collinearity which, while literally true, is totally misleading since the collinearity in no way reduces the usefulness and predicting power of the regression equation, as long as the meaning of the expansion term is taken into account. But even in cases without expansion terms, the measure in question may be misleading. Thus when applied to the protein calibration data of table 5, it may well lead the analyst to give up on these data as a hopelessly highly-collinear set, whereas, as we have seen, there is nothing wrong with this set and it can indeed be used very effectively for the calibration of a method for protein determination based on reflectance measurements.

Finally, a few words about estimating the β-coefficients considered as rates of change of y with changes in the individual x_j. As pointed out by Box [9], this is generally *not* a desirable use of regression equations. If, however, it is the major purpose of a particular experiment, then this experiment should be designed accordingly, which means: essentially with an orthogonal X matrix. A collinear X matrix leads to the ability to estimate certain linear combinations of the β's much better than the β's themselves. The experimenter can calculate the VF values, not only for any point of **X** space, but also for any **β** or combination of **β**'s, *and he can do this without making a single measurement*, i.e., in the planning stages of the experiment. If the experimenter does not take advantage of this opportunity, he may be in for considerable disappointment, after having spent time, money, and effort on inadequate experimentation. We believe that he advocacy of remedial techniques, such as Ridge Regression for collinear data is unwise. One of the most important tasks of a data analyst is to detect, and to call attention to, limitations in the use and interpretation of the data.

References

[1] Buck, J.B., Studies on the Firefly, Part I: The Effects of Light and Other Aspects on Flashing in Photinus Pyralic, with Special Reference to Periodicity and Diurnal Rhythm, Physiological Zoology, **10**, 45-58 (1937).

[2] Mandel, J., Use of the Singular Value Decomposition in Regression Analysis, The American Statistician, **36**, 15-24 (1982).

[3] Fearn, T., A Misuse of Ridge Regression in the Calibration of a Near Infrared Reflectance Instrument, Applied Statistics, **32**, 73-79 (1983).

[4] Belsley, D.A., E. Kuh, and R.E. Welsh, *Regression Diagnostics: Identifying Influential Observations and Sources of Collinearity*, Wiley, NY (1980).

[5] Draper, N.R. and H. Smith, *Applied Regression Analysis*, Wiley, 2nd Edition, NY (1981).

[6] Belsley, D.A., Demeaning Conditioning Diagnostics Through Centering, The American Statistician, **38**, 73-77, and "Comments", 78-93 (1984).

[7] Daniel, C., and F.S. Wood, *Fitting Equations to Data*, Wiley, 2nd Edition, NY (1980).

[8] Willan, A.R. and D.G. Watts, Meaningful Multicollinearity Measures, Technometrics, **20**, 407-12 (1978).

[9] Box, G.E.P., Use and Abuse of Regression, Technometrics **8** 625-29 (1966).