Multivariate Calibration With More Variables Than Observations

Rolf Sundberg
Institute of Actuarial Mathematics
and Mathematical Statistics
University of Stockholm
S-113 85 Stockholm
Sweden

Philip J. Brown
Department of Statistics and
Computational Mathematics
University of Liverpool
Liverpool L69 3BX
United Kingdom

Multivariate calibration involves the use of an estimated (linear) relationship between a q-variate response vector $Y$ and a p-dimensional explanatory vector $X$ to estimate or predict future unknown $X$ from further observed $Y$. In some applications—for example, when $Y$ represents automatic intensity measurements at $q$ different wavelengths on $n$ chemical standard samples—the feasible sample size $n$ may be restricted, whereas the dimension $q$ can be chosen quite large. In this article, the singular cases appearing when $n < p + q + 1$ are investigated. It is shown that, if $n > q$, the traditional solutions to the estimation and prediction problems—that is, the generalized least squares estimator and the estimated best linear predictor—are both unique, whereas for smaller $n$, $(q - n + 1)$-dimensional hyperplanes of solutions are obtained, the same in both problems. The properties of the predictor are also empirically studied in an example with $p = 1$, $q = 6$, and varying $n$.

KEY WORDS: Best linear predictor; Collinearity; Generalized least squares; Multivariate regression model; Singular sample covariance matrix.

1. INTRODUCTION

Calibration is needed when an expensive or laborious but accurate measurement method is replaced by a cheap and quick but indirect or less accurate method. As an example, for determination of concentrations in analytical chemistry, a spectrophotometer may give quick and cheap intensity measurements, linearly related to the concentrations of substances present. In multivariate calibration, $q$ different measurements, represented by $y_1, \ldots, y_q$, are used to determine $p \leq q$ different $x$ variables, $x_1, \ldots, x_p$. They may, for instance, be intensity measurements at $q$ different wavelengths and concentrations of $p$ different substances, respectively. We will discuss such an example with $q = 6$ and $p = 1$.

We assume a standard multivariate linear regression model; that is, each $y$ variate is assumed to have a multiple linear regression on the $p$ $x$ variables with mutually correlated error terms. In the preceding chemical example, the linearity can be motivated by Beer's law from chemical theory. In other applications, a linear relationship may be of a more empirical character.

For the calibration, we have a multivariate sample of size $n$ with $x$ values directly observed. In matrix notations, we express the calibration data as

$$Y = \mathbf{1}a' + XB + \mathbf{E}, \quad (1.1)$$

where $Y = (Y_1 \cdots Y_q)'$, size $(n \times q)$, and $X = (X_1 \cdots X_p)'$, $(n \times p)$, represent data, $a$ is a $q$ vector, $B$ is a $(p \times q)$ matrix of regression coefficients, and $\mathbf{1}$ is an $n$ vector of ones. The error matrix $\mathbf{E}$ consists of $n$ mutually uncorrelated rows $e_i'$, $\mathbf{E} = (e_1', \ldots, e_n')$, in which each vector $e_i$ has mean 0 and variance-covariance matrix $\Gamma = E(e_i e_i')$ $(i = 1, \ldots, n)$. We assume $X$ to be of full rank and $1 \leq p \leq q$.

After calibration, we want to infer about unknown $X$ vectors from (future) observed $Y$ vectors. To distinguish from calibration data, we denote the future random $Y$ vector by $Z$ and the unknown $X$ vector by $\xi$. For $Z$ ($q \times 1$), the model corresponding to (1.1) is

$$Z = \alpha + B'\xi + \mathbf{e}, \quad (1.2)$$

with the same distribution of $\mathbf{e}$ as of $e_i$ previously and with independence.

In controlled calibration, the vectors $X_i$ are regarded as fixed (controlled). In natural calibration, the $X_i$'s are regarded as random; we then have a joint $(p + q)$-dimensional distribution of $(X_i', Y_i')'$. A corresponding distinction can be made about $\xi$. 

365
If $\xi$ is regarded as a parameter to be estimated, the extended "classical" estimator $\hat{\xi}$ is obtained by generalized least squares (GLS) applied to (1.2), with estimates $\hat{\alpha}$, $\hat{B}$, and $\hat{\Gamma}$ inserted for the unknown parameters in (1.1); that is,

$$\hat{\xi} - \bar{X} = (\hat{B}'^{-1}\hat{B})^{-1}\hat{B}'^{-1}(Z - \bar{Y}).$$

(1.3)

Here $\bar{X}$ and $\bar{Y}$ are the averages of the calibration data vectors $X_i$ and $Y_i$; that is, $\bar{X} = X'1/n$ and correspondingly for $\bar{Y}$, $\hat{B}$ is the least squares (LS) estimate of $B$, $\hat{\Gamma}$ is the normed residual sum of squares and products matrix (residual covariance matrix),

$$\hat{B} = S_{xx}^{-1}S_{xy},$$

(1.4)

and

$$\hat{\Gamma} = (Y - 1\bar{Y}' - (X - 1\bar{X}')\hat{B})'\times(Y - 1\bar{Y}' - (X - 1\bar{X}')\hat{B})/(n - 1),$$

$$= S_{yy} - S_{sx}S_{xx}^{-1}S_{sy},$$

(1.5)

where $S_{xx}$, $S_{sy}$, $S_{sx}$, and $S_{yy}$ are taken from the sample covariance matrix $S$ of $(X', Y')$,

$$S = \begin{pmatrix} S_{xx} & S_{sx} \\ S_{sy} & S_{yy} \end{pmatrix}.$$  

(1.6)

The more usual unbiased estimator of $\Gamma$ differs from (1.5) by having the scalar factor $n - p - 1$ instead of $n - 1$, but this difference has no effect on $\hat{\xi}$. Even under normality, $\hat{\xi}$ is not the maximum likelihood estimator, unless $p = q$, but the difference is small except in less usual circumstances (see Brown and Sundberg 1987).

An alternative is a natural calibration situation, with $\xi$ treated as random and sampled from the same population as the $X_i$'s. The best linear predictor (BLP) $\hat{\xi}$, based on estimated parameters from the calibration, is obtained by regressing $\xi$ on $Z$ and inserting LS estimates for the parameters of this regression, that is, we use the "inverse regression" estimator $\hat{\xi}$, given by

$$\hat{\xi} - \bar{X} = S_{xy}S_{xx}^{-1}(Z - \bar{Y}).$$

(1.7)

For a general discussion of the properties of these two estimators/predictors, (1.3) and (1.7), see Brown (1982).

The predictor $\hat{\xi}$ is based on the stronger assumption of natural calibration and natural $\xi$. Even for estimation in controlled multivariate calibration, $\hat{\xi}$ may be justified from mean squared error comparisons (see Sundberg 1985) or by a Bayesian argument, which assumes that $\xi, X_1, \ldots, X_n$ are exchangeable (see Brown 1982, theorem 3). Other prior assumptions on $\xi$ may justify intervening estimators between $\hat{\xi}$ and $\xi$.

To avoid singularity of the residual covariance matrix $\hat{\Gamma}$, (1.5), a standard assumption in multivariate-linear-models analysis is that $n \geq p + q + 1$ (e.g., see Seber 1984, theorem 8.4). In some applications of the calibration model (1.1), $q - \dim(Y)$ may easily be chosen quite large, whereas the feasible number $n$ of observation vectors (the calibration sample size) is more restricted. As an example, $Y$ may represent cheap intensity measurements at $q$ different spectral wavelengths on each of $n$ expensive chemical standard samples. In an illustration in Sjöström, Wold, Lindberg, Persson, and Martens (1983), $n = 16$, $q = 27$, and $p = 3$. In near infrared (NIR) reflectance spectroscopy, very large $q$ values, such as $q = 500$ (Martens and Naes 1984) and even larger values (Osborne, Fearn, Miller, and Douglas 1984) have been discussed. In Section 6, we will study a set of NIR data from Fearn (1983) with $q = 6$ and $p = 1$. Although the $q$ value is not very high, we might ask if a calibration sample of size $n = 4$, say, would be possible, since univariate calibration is often performed with such small sample sizes.

When the assumption $n \geq p + q + 1$ is not satisfied, $\hat{\Gamma}$ is singular. A singular $\hat{\Gamma}$ invalidates Formula (1.3) for $\hat{\xi}$, but not necessarily (1.7) for $\hat{\xi}$. If $n \equiv q$, however, $S_{yy}$ will be singular and (1.7) is invalidated. In what follows, we will investigate the relationship between the two estimation methods in the two cases (a) $q < n \leq p + q$ and (b) $p < n \leq q$. In particular, the apparent lack of equality between the uniqueness conditions for $\hat{\xi}$ and $\hat{\xi}$ seen previously will be shown to disappear. We will also comment on a factor-analysis method proposed by Naes (1985) that apparently avoids singularities.

Note that within a Bayesian framework problems with singularities in $\Gamma$ would disappear under a proper prior distribution for the parameters. Suitable priors for the estimation of a covariance matrix were discussed by Lindley (1978, sec. 6).

2. A CANONICAL REPRESENTATION

We will proceed by introducing the sample canonical representation for the calibration data $(X, Y)$, as a simple setup in which to discuss the GLS $\hat{\xi}$ and the BLP $\hat{\xi}$. Note that GLS and BLP could both be defined in a coordinate-free way as solutions to minimization problems invariant under linear coordinate transformations. The particular canonical transformation yields a most simple structure to the sample covariance matrix.

First, we make the natural basic assumptions $n > p$ and $S_{xx}$ nonsingular. If $X_1, \ldots, X_n$ are randomly generated from a continuous distribution, the latter requirement is satisfied with probability 1. Likewise, we assume the distribution of $\xi_0$ to be continuous. The statements to follow about ranks of covariance matrices then hold with probability 1 (see Seber 1984, sec. A5).
We use the same notation \((X, Y)\) for the calibration data after the canonical transformation has been undertaken. In standard literature on multivariate analysis, \(n \geq p + q + 1\) is required, and then the canonical covariance matrix has the form

\[
S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} = \begin{pmatrix} I_p & R \\ R^T & I_q \end{pmatrix},
\]

(2.1)

where \(I_p\) and \(I_q\) are unit matrices \((p \leq q)\) and \(R\) diagonal,

\[
R = \begin{pmatrix} r_1 & 0 \\ \vdots & \ddots \\ 0 & r_p \end{pmatrix},
\]

(2.2)

with \(1 > r_1 > \cdots > r_p > 0\) (e.g., see Rao 1973, chap. 8f). The quantities \(r_j\) \((j = 1, \ldots, p)\) are the canonical correlations of the sample.

For the description of the canonical covariance matrix \(S\) in the present case of smaller \(n\) values, we distinguish two cases: In case a, \(q < n \leq p + q\), both \(S_{xx}\) and \(S_{yy}\) will be nonsingular but \(\text{rank}(S) = n - 1 < p + q\). This implies that the canonical \(S\) still has the same form as (2.1) and (2.2), except that now \(r_j = 1\) for \(j \leq p + q - n + 1\) and \(r_j < 1\) for larger values of \(j\) \((j \leq p)\). In case b, \(p < n \leq q\), \(S_{yy}\) is not of full rank:

\[
S_{yy} = \begin{pmatrix} I_{n-1} & 0 \\ 0 & 0_{q-n+1} \end{pmatrix},
\]

(2.3)

Since \(\text{rank}(S) = n - 1\), it follows that

\[
R = I_p, (S_{xx}).
\]

(2.4)

In this case, the canonical transformation is no longer unique. Since the last \(q - n + 1\) canonical components are constant \((= 0)\) over the sample, any canonical component is unique only up to an arbitrary added contribution from these last ones. This could be used to choose the canonical transformation partially orthogonal if desired, but for most of the sequel, we do not need any such property.

For Fearn’s (1983) data, mentioned in Section 1 and to be studied in Section 6, \(q = 6\) and \(p = 1\). Then case a corresponds to a calibration sample of size \(n = 7\) and case b to sizes in the interval \(2 \leq n \leq 6\).

### 3. The Generalized Least Squares Estimator \(\hat{\xi}\)

From the canonical representation of \(S\), we obtain the following form for the residual covariance matrix \(\Gamma\), by insertion in (1.5): In case a, \(q < n \leq p + q\),

\[
\Gamma = \begin{pmatrix} I_p - R^2 & 0 \\ 0 & I_{q-p} \end{pmatrix}.
\]

(3.1)

Here, when \(n < p + q + 1\), the first \(p + q + 1 - n\) diagonal elements of \(\Gamma\) are, in fact, 0. In case b, \(p < n \leq q\);

\[
\hat{\Gamma} = \begin{pmatrix} 0_p & 0 \\ 0 & I_{q-p} \end{pmatrix}.
\]

(3.2)

From now on, for simplicity of notation only, we assume \(X\) and \(Y\) centered, so that \(\bar{X} = \bar{Y} = 0\). We introduce the empirical model for \((Z, \xi)\), based on the estimates [Eqs. (1.4)–(1.5)] of \(B\) and \(\Gamma\); that is,

\[
Z = \hat{B}\xi + e = \begin{pmatrix} R \\ 0 \end{pmatrix} \xi + e,
\]

(3.3)

\(E(e) = 0, \text{var}(e) = \hat{\Gamma}\).

Here the notation \(e\) is chosen for the \(e\) of (1.2) to signify that the “error” covariance matrix of model (3.3) is regarded to be \(\hat{\Gamma}\). The motivation for this “model” is that the GLS and the BLP are defined as optimal choices under the assumption that calibration data yield accurate parameters. Note that, if data are precise, \(\hat{B}\) will be close to the true \(B\) even if \(n\) is small.

To express how \(\xi\) is best estimated from \(Z\) under Model (3.3), we introduce notations for subvectors of \(Z\) and \(e\). In case b, we partition in analogy with (3.2) and let \(Z_1\) and \(e_1\) be the subvectors of the first \(p\) components and \(Z_2\) and \(e_2\) those of the next \(n - p - 1\) components. Finally, \(Z_3\) and \(e_3\) represent the last \(q - n + 1\) components, which are of zero variance under Model (3.3). In case a, \(Z_3\) and \(e_3\) are void and \(Z_2\) and \(e_2\) are of dimension \(q - p - 1\) [see (3.1)].

Any estimator of \(\xi\) linear in \(Z\) may now be written

\[
CZ = C_1Z_1 + C_2Z_2 + C_3Z_3,
\]

(3.4)

where \(C = (C_1, C_2, C_3)\) is a \((p \times q)\) coefficient matrix, to be chosen. Inserting Model (3.3), we get

\[
CZ = C_1R\xi + C_1e_1 + C_2\xi_2 + C_3e_3,
\]

(3.5)

The GLS estimator should be a minimum variance unbiased linear estimator of \(\xi\) under model (3.3). Obviously, unbiasedness of (3.5) under (3.3) requires \(C_1 = R^{-1}\), but nothing more. To minimize the variance of (3.5), note that under (3.3) \(e_1\) and \(e_2\) are uncorrelated, that \(\text{var}(e_2)\) is strictly positive, and that \(e_3\) has zero variance. Hence we must take \(C_3 = 0\), whereas \(C_1\) can be chosen arbitrarily. Then we arrive at the following estimators [Eq. (3.6) for case a and Eq. (3.7) for case b; \(R = I_p\) in (3.7)]. For \(q < n \leq p + q\),

\[
\hat{\xi} = R^{-1}Z_1.
\]

(3.6)
and for \( p < n \leq q \),
\[
\hat{\xi} = R^{-1}Z_1 + C_3Z_3
\\ = Z_1 + C_3Z_3, \quad C_3 \text{ arbitrary.} \tag{3.7}
\]

In principle, since the (estimated) distribution of \( Z_3 \) was concentrated at 0, the expressions (3.7) for different \( C_3 \) should be identical with probability 1. The true variance of \( Z_3 \) is not 0, however, so the actually observed \( Z_3 \) will be almost surely nonzero. It follows that, for each component of \( \hat{\xi} \), (3.7) represents a \((q - n + 1)\)-dimensional hyperplane of estimators, as expressed in the calibration sample canonical representation \( Z \).

A particular choice of \( C_3 \) is the one that makes \( \hat{\xi} \) of minimum length among estimators of form (3.7). Suppose that the canonical transformation is chosen such that \( Z_1 \) is orthogonal to \( Z_3 \), which is always possible (see Sec. 2). Among all estimators of form (3.7), the particular choice
\[
\hat{\xi} = Z_1 \tag{3.8}
\]
will then have for each component a coefficient vector of minimum Euclidean length when expressed in the original data. Note, however, that the minimum-length solution is not transformation-invariant. For instance, if data are first scaled to unit variances, as is often done, so that the Euclidean metric is defined in terms of variance-standardized variates, the minimum-length estimator will change correspondingly.

Alternatively, Formula (3.7) could have been derived by inserting a g-inverse \( \Gamma^{-1} \) of \( \Gamma \) in (1.3). Not all g-inverses could be allowed, however (note that many g-inverses of \( \Gamma \) are also proper inverses of other covariance matrices, which could differ essentially from \( \Gamma \)), and this makes such a derivation more tricky, besides being less intuitive than the one given previously.

4. THE BEST LINEAR PREDICTOR \( \hat{\xi} \)

Now \( \hat{\xi} \) is regarded as a random variable to be predicted from \( Z \). In case a, \( q < n \leq p + q \), Formula (1.7) for the BLP \( \hat{\xi} \) is without complications. From (2.1), we find (still assuming \( X \) and \( Y \) centered) that
\[
\hat{\xi} = S_{xy}S_{yy}^{-1}Z = (R | 0)l_qZ - RZ_1. \tag{4.1}
\]

This result corresponds to (3.6), \( \hat{\xi} = R^{-1}Z_1 \), precisely as we would expect, illustrating the difference between “regress \( y \) on \( x' \)” and “regress \( x \) on \( y \).”

When \( n \leq q \), \( S_{xy} \) is singular and (1.7) is invalidated, so in case b we go to the definition of a BLP to find its form. The estimated model on which to base the BLP is now characterized by the zero mean vector and the covariance matrix for \((\xi, Z_1, Z_2, Z_3)\),
\[
S = \begin{pmatrix}
I_p & R & 0 & 0 \\
R & I_q & 0 & 0 \\
0 & 0 & I_{n-1-p} & 0 \\
0 & 0 & 0 & 0_{q-n+1}
\end{pmatrix}, \tag{4.2}
\]
where actually \( R = I_p \). From this it follows that the BLP for \( n \leq q \) is
\[
\hat{\xi} = RZ_1 + C_3Z_3 - Z_1 + C_3Z_3, \tag{4.3}
\]
with \( C_3 \) an arbitrary coefficient matrix. In particular, we conclude that \( \hat{\xi} = \xi \) for \( n \leq q \), by comparison with (3.7).

As with \( \xi \), we could make \( \hat{\xi} \) unique by choosing the minimum-length version so that (3.8) holds for \( \hat{\xi} \) as well. The discussion of dependence on the metric behind is equally valid for \( \xi \), so this uniqueness is somewhat illusory.

Alternatively (4.3) could have been obtained by inserting the general form of a g-inverse, \( S_{yy}^{-1} \) of \( S_{yy} \), for \( S_{yy}^{-1} \) in (1.7). A particular form of \( S_{yy}^{-1} \) will yield the minimum-length version. In practice, this choice could be computationally more convenient than to derive explicitly the canonical representation (which was used previously to make theory simple). An orthogonal transformation to principal components in \( Y \) space will make \( S_{yy} \), diagonal of rank \( n - 1 \). Now define \( S_{yy}^{-1} \) by inverting the \( n - 1 \) nonzero elements of the diagonal \( S_{yy} \) and preserving all zeros. This will yield the minimum-length version used in the calculations for the example of Section 6.

Substitutes for \( S_{yy}^{-1} \) in (1.7) other than a g-inverse \( S_{yy}^{-1} \) will lead to alternative predictors. One proposed alternative is discussed in Section 5 [see (5.2)]. Other alternatives derive from methods for treating collinearity in multiple regression, such as ridge regression, principal component regression (PCR), and partial least squares (PLS).

5. FACTOR STRUCTURE ON THE RESIDUAL COVARIANCE MATRIX \( \Gamma \)

Here we discuss, in the light of previous sections, an approach by Naes (1985, 1986) to represent the residual covariance matrix \( \Gamma \) by a factor-analysis model,
\[
\Gamma = PP' + F, \tag{5.1}
\]
where \( P \) is a \((q \times k)\) matrix of loadings with one column per factor (so preferably it should have few columns) and \( F \) is a diagonal variance matrix representing pure noise. In the chemical context, we may think of the factors as representing known or unknown substances present in concentrations that are varying but not measured. Both \( P \) and \( F \), and usually even \( k \), are a priori unknown and must be estimated from the calibration data.

TECHNOMETRICS, AUGUST 1989, VOL. 31, NO. 3
When an estimate $\hat{f}$ of $f$ has been obtained, it can be used directly for $\hat{f}$ in the GLS estimator (1.3) (see Naes 1986) or used indirectly in the BLP (Naes 1985). In the latter case, $S_{yy}$ in (1.7) is replaced by

$$\hat{f} + S_{xy} S_{xx}^{-1} S_{yx}$$  (5.2)

[see (1.5)]. Another argument for (5.1) and (5.2) given by Naes (1985) was that his predictor, but not (1.7), is applicable when $n \leq q$. As shown in Section 4, however, there is a natural extension of (1.7) to this case, given by (4.3), and by choosing, for instance, the minimum-length solution, (4.3) is made unique. Moreover, note that estimators/predictors based on (5.1) are not transformation invariant but depend on the choice of metric behind, for instance, whether data are first variance-standardized or not.

The reason that the factor structure (5.1) yields a nonsingular estimate of $\Sigma_{yy}$ is that typically the matrix difference between the residuals covariance matrix $\hat{f}$ and the estimated $PP'$ will have all diagonal elements positive, even though it is only semidefinite. The diagonal matrix $F$ is next estimated as strictly positive. This will be so unless the number $k$ of factors is high enough to allow $\hat{f}$ to be perfectly modeled through $PP'$—that is, unless $k \geq n - p - 1$ [ = rank($\hat{f}$); see (3.2)]. This observation is confirmed by looking at example 6.1 (meat data) in Naes (1985), where he reported matrix inversion problems for $k \geq 13$ when $n = 15$ and $p = 1$; that is, $n - p - 1 = 13$.

6. AN EXAMPLE

Fearn (1983) published a calibration example of highly collinear NIR reflectance data. An NIR vector of dimension $q = 6$ was used to predict the protein content in wheat, a scalar $\xi$ ($p = 1$). Fearn had a calibration sample of size $n = 24$ and tried the calibration relationship on a prediction sample of size 26. He used these data in a criticism of ridge regression. His challenge was answered by Hoerl, Kennard, and Hoerl (1985). Fearn's data were also used by Naes and Martens (1985) for comparison with the BLP (1.7) of predictors derived from PCR and PLS regression. They used only the first half of Fearn's original calibration sample, for tendencies to be more clearly seen. We will use the same data to illustrate the theory of the present article, with calibration samples of sizes $n \leq q = 6$. We use the same prediction sample and the same root mean squared error (RMSE) criterion as Fearn (1983) and Naes and Martens (1985). To illustrate the lack of transformation invariance of the GLS/BLP estimators/predictors for $n \leq q$, they are applied not only on the original log values used by Fearn and by Naes and Martens but also on linearly transformed data along the lines of Hoerl et al. (1985). Specification will follow.

As a first observation, the old controversy in the statistical theory of calibration on whether to use estimator (GLS) or predictor (BLP) versions was seen to be of no significance. Their difference in prediction RMSE was negligible for the larger sample sizes, $n = 24$ and $n = 12$, that were tried, and from the theory in previous sections we know that they are identical for $n \leq q = 6$. Hence we need not distinguish GLS and BLP even for larger sample sizes, and mostly we will refer only to BLP.

The BLP/GLS Predictor for $n = 6$. In Table 1, the first $n = 6$ observations from Fearn's calibration data set are reproduced. We will use these data specifically to illustrate the dependence on choice of metric for the minimum length BLP/GLS.

For the data of Table 1, with $n = q$, BLP/GLS spans a line in six-dimensional coefficient space [see (3.7) and (4.3)]. Figure 1 illustrates the RMSE for this set of predictors as a function of the coefficient of $y_1$. Minimum-length BLP/GLS corresponding to some particular choices of metric are indicated in the figure. All of these metrics (transformations) represent variance-standardized data. A few attempts with nonstandardized variates gave much worse predictions than with standardized variates, so the standardization to unit variances typically employed in transformation-dependent methods like PCR, PLS, and ridge regression appeared wise also in this case.

The original data are denoted data $(n, 0)$ to indicate $n$ observations and no transformation. The transformations undertaken before variance standardization followed the proposal by Hoerl et al. (1985). The original components $y_1, \ldots, y_6$ were replaced by their average $\bar{y}$ together with the deviations $y_j - \bar{y}$ from the average for five selected $j$ values (out of the six). After variance standardization, six different alternative metrics were generated in this way, corresponding to data sets called data $(n, j)$ when deviation $y_j - \bar{y}$ was excluded ($j = 1, \ldots, 6$). It was found natural to include also the metric corresponding to the singular data set data $(n, 7)$, in which all deviations $y_j - \bar{y}$ were retained, in addition

![Table 1. The First $n = 6$ Observations From Fearn's (1983) Calibration Set, With $x =$ Protein Percentage and $y_1, \ldots, y_6$ Referred to as "Log Values" (reflectance values on a $\log$ scale).](image-url)

<table>
<thead>
<tr>
<th>Observation</th>
<th>$x$</th>
<th>$y_1$</th>
<th>$y_2$</th>
<th>$y_3$</th>
<th>$y_4$</th>
<th>$y_5$</th>
<th>$y_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.23</td>
<td>468</td>
<td>123</td>
<td>246</td>
<td>374</td>
<td>386</td>
<td>-11</td>
</tr>
<tr>
<td>2</td>
<td>8.01</td>
<td>458</td>
<td>112</td>
<td>236</td>
<td>386</td>
<td>383</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>10.95</td>
<td>457</td>
<td>118</td>
<td>240</td>
<td>359</td>
<td>353</td>
<td>-16</td>
</tr>
<tr>
<td>4</td>
<td>11.84</td>
<td>450</td>
<td>116</td>
<td>238</td>
<td>362</td>
<td>340</td>
<td>-15</td>
</tr>
<tr>
<td>5</td>
<td>10.41</td>
<td>464</td>
<td>119</td>
<td>243</td>
<td>366</td>
<td>371</td>
<td>-16</td>
</tr>
<tr>
<td>6</td>
<td>9.51</td>
<td>499</td>
<td>147</td>
<td>273</td>
<td>404</td>
<td>433</td>
<td>5</td>
</tr>
</tbody>
</table>

TECHNOMETRICS, AUGUST 1989, VOL. 31, NO. 3
Figure 1. RMSE(BLP) as a Function of the Coefficient for $y$, n = 6, Quarter Sample I. Arrows 0, 1, . . . , 7 point at RMSE values for minimum length BLP under transformations (metrics) (0), (1), . . . , (7) of data (see the text). The minimum-length RMSE(BLP) attainable, .77, and RMSE($\bar{x}$) = 1.51 are also indicated.

The singularity of this data set (and of the corresponding metric) presented no extra problem.

Notation for transformed calibration data is as follows: data $(n, 0) = y_1, y_2, . . . , y_n;$ data $(n, j) = \tilde{y}_1, y_1^*, . . . , y_n^*$ where $y_j^* = y_j - \bar{y} = \sum_i y_i y_j/q.$

It is clear from the figure that the choice of metric was of great importance. The least RMSE possible for the BLP with these data was .77. The eight realized values of RMSE were between 1.06 and 4.55 (see also Table 2). An explanation for the extreme RMSE of (6, 4) is that $y_4$ happened to dominate in the other predictors based on this data set. The value for the pessimistic constant predictor $\bar{x} = \sum_i x_i/n$. RMSE = 1.51, is also indicated in the figure. (Note that this value almost precisely equals the least possible RMSE for a constant predictor.)

**BLP/GLS for Varying Sample Size.** The RMSE of the BLP and the GLS based on the whole calibration set of $n = 24$ was .24. For the two half-samples of $n = 12$, RMSE was substantially higher—.83 and .62, respectively. For quarter-samples, $n = 6$, the BLP is no longer unique, as discussed previously. Table 2 gives RMSE values for minimum-length BLP/GLS based on data $(6, j)$ $(j = 0, . . . , 7).$ Even if we choose one of the better metrics, we find that RMSE has increased further for three of the four quarter-samples and substantially so for two of them. As far as this goes, the results agree with the idea that smaller samples typically yield more imprecise calibration.

Table 3 gives RMSE values for the minimum length BLP based on four sets of calibration data of size $n = 4$, selected as the first four observations from each quarter-sample. Note that one of these calibration samples yields terribly bad predictors, but also that the other three samples yield predictors that are even better than for $n = 12$.

The reason behind these substantial improvements of the BLP when the sample size is reduced to $n = 4$ (excluding for a moment quarter-sample 3) can be found in the eigenvalue structure of $S$. For a larger $n$, all except a few eigenvalues are very small and correspond to nonpredictive near singularities. Replacing them by zeros and choosing a minimum-length solution will stabilize and improve the predictor. This is the beneficial effect of sample-size reduction below $n = q + 1$, which is in many cases apparently able to overtake the usual effect of sample-size reduction. With a natural sample as small as $n = 4$, however, we are more vulnerable to the outcome of the selection of calibration $x$ values and to the random noise in individual NIR data. As it turned out, the third quarter-sample generated a very bad calibration sample of size $n = 4$ that yielded only bad predictors, all (not only BLP's) worse than the constant predictor $\bar{x}$ (of RMSE = 1.51). The samples that we tried were not selected to be good or bad. For pos-

<table>
<thead>
<tr>
<th>Data transformation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data (6, 0)</td>
<td>.62</td>
<td>.63</td>
<td>.65</td>
<td>.66</td>
</tr>
<tr>
<td>Data (6, 1)</td>
<td>.61</td>
<td>.59</td>
<td>.60</td>
<td>.61</td>
</tr>
<tr>
<td>Data (6, 2)</td>
<td>.60</td>
<td>.58</td>
<td>.60</td>
<td>.61</td>
</tr>
<tr>
<td>Data (6, 3)</td>
<td>.59</td>
<td>.57</td>
<td>.59</td>
<td>.60</td>
</tr>
<tr>
<td>Data (6, 4)</td>
<td>.58</td>
<td>.56</td>
<td>.58</td>
<td>.60</td>
</tr>
<tr>
<td>Data (6, 5)</td>
<td>.57</td>
<td>.55</td>
<td>.57</td>
<td>.59</td>
</tr>
<tr>
<td>Data (6, 6)</td>
<td>.56</td>
<td>.54</td>
<td>.56</td>
<td>.58</td>
</tr>
<tr>
<td>Data (6, 7)</td>
<td>.55</td>
<td>.53</td>
<td>.55</td>
<td>.57</td>
</tr>
<tr>
<td>Constant predictor $\bar{x}$</td>
<td>1.51</td>
<td>1.51</td>
<td>1.51</td>
<td>1.51</td>
</tr>
</tbody>
</table>

**Table 3. RMSE for Minimum-Length BLP With Calibration Samples of Size $n = 4$**

<table>
<thead>
<tr>
<th>Data transformation</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data (4, 0)</td>
<td>1.21</td>
<td>3.41</td>
<td>3.56</td>
<td>.55</td>
</tr>
<tr>
<td>Data (4, 1)</td>
<td>.65</td>
<td>.46</td>
<td>.56</td>
<td>.53</td>
</tr>
<tr>
<td>Data (4, 2)</td>
<td>.75</td>
<td>.73</td>
<td>3.17</td>
<td>.50</td>
</tr>
<tr>
<td>Data (4, 3)</td>
<td>.74</td>
<td>.60</td>
<td>4.59</td>
<td>.58</td>
</tr>
<tr>
<td>Data (4, 4)</td>
<td>1.37</td>
<td>.66</td>
<td>2.83</td>
<td>.82</td>
</tr>
<tr>
<td>Data (4, 5)</td>
<td>.66</td>
<td>.89</td>
<td>2.07</td>
<td>.52</td>
</tr>
<tr>
<td>Data (4, 6)</td>
<td>.64</td>
<td>.76</td>
<td>4.47</td>
<td>.55</td>
</tr>
<tr>
<td>Data (4, 7)</td>
<td>.72</td>
<td>.58</td>
<td>3.54</td>
<td>.51</td>
</tr>
<tr>
<td>Constant predictor $\bar{x}$</td>
<td>1.51</td>
<td>1.91</td>
<td>1.51</td>
<td>1.52</td>
</tr>
</tbody>
</table>
possible use in routine practice, a calibration sample of such small size must be (to some extent at least) controlled to guarantee a high signal-to-noise ratio through a large enough spread in x values.

These conclusions are supported when looking at results for \( n = 5 \) and \( n = 3 \). RMSE values for \( n = 5 \) were typically between those for \( n = 6 \) and \( n = 4 \). For four samples of size \( n = 3 \), the results were less clear-cut, being a mixture of low and high values, with more of the latter type than for \( n = 4 \).

7. CONCLUSIONS

We have seen in this article that the simple GLS estimator and the simple BLP can be used in the singular case \( n \leq q \), in which they both span the same \((q - n + 1)\)-dimensional hyperplane in the space of linear estimators/predictors. We have also seen in a prediction test on real data, supported by theoretical arguments, that when only a small sample is available, the minimum-length BLP ( = GLS) can be a good predictor (estimator). It must be ensured, however, that the sample is informative, and it might be a problem to choose a good metric for the minimum-length version.

[Received November 1985. Revised November 1988.]

REFERENCES


