Teacher's Corner

An Inequality for a Measure of Deviation in Linear Models

Thomas MATHEW and Kenneth NORDSTRÖM

A matrix inequality is established that provides an upper bound for a quadratic form that involves the difference between two linear unbiased estimators of the same linear parametric function in a general linear model. Various special cases of the inequality are discussed. Certain inequalities that arise in the problem of outlier detection and prediction of observations come out as special cases. In addition, some extensions of Samuelson's inequality are also obtained.

KEY WORDS: Outliers; Samuelson's inequality; Studentized residuals.

1. INTRODUCTION

Several articles have recently appeared, all dealing with the maximum of, and bounds for, some version of standardized residuals in a linear regression model. The related question of how much an observation in a random sample can differ from the sample mean has similarly continued to receive constant attention in the literature. An example of an article dealing with the former problem is Gray and Woodall (1994), where several further recent contributions are cited, while the latter problem has been extensively (and excellently) reviewed both from the historical as well as the technical point of view by Olkin (1992).

This paper provides yet another perspective on the above-mentioned problems. It will be argued that several inequalities of the above type, including their vector-valued extensions, can be obtained directly from a single elementary matrix inequality (inequality between two orthogonal projections). This matrix inequality will further be shown to yield a number of inequalities and distributional results useful for regression diagnostics and outlier detection. For the most part the derived results can already be found in the literature, but some vector-valued extensions and inequalities appear to be new. However, the main purpose of the paper is to show that a single inequality underpins a multitude of apparently unrelated inequalities, thus allowing for a unified derivation of these inequalities.

Throughout the paper we shall be concerned with various versions of the standard linear regression model for an $n \times 1$ vector $\mathbf{y}$ of responses, given by

$$\mathbf{y} = X\beta + \varepsilon, \qquad E(\varepsilon) = 0,\ \operatorname{cov}(\varepsilon) = \sigma^2 I_n \qquad (1.1)$$

where X is a known $n \times m$ matrix, β is the unknown $m \times 1$ vector of regression parameters, and $\sigma^2 > 0$ is the unknown error variance. The model matrix X will be allowed to be rank-deficient at no real extra cost, so as to include, for example, various models for designed experiments. The vector of predicted (fitted) values $\hat{\mathbf{y}}$ is therefore given by $\hat{\mathbf{y}} = P\mathbf{y} = X\hat{\beta}$, where $P = X(X'X)^{-}X'$ is the projection matrix onto the regression space (the "hat matrix") and $\hat{\beta}$ denotes any solution to the set of normal equations of the model (1.1). If the reader prefers to think in terms of a full-rank model, then the generalized inverse, appearing above in the expression for P, should be replaced by the true inverse of $X'X$, and $\hat{\beta}$ is then the least squares estimator of β; see, for example, Seber (1977, sec. 3.8). In the illustrative examples given below we assume X to be of full rank for simplicity.

The general inequality, alluded to above, involves the deviation between two linear functions of the observation vector $\mathbf{y}$ having the same expected value under the model (1.1). More specifically, given two linear functions $A_1\mathbf{y}$ and $A_2\mathbf{y}$ of $\mathbf{y}$ satisfying

$$A_1X - A_2X = \mathbf{0}, \qquad (1.2)$$

we shall be concerned with the problem of constructing a suitable quadratic form in the vector of deviations $A_1\mathbf{y} - A_2\mathbf{y}$ in order to obtain a useful upper bound on this form. For this purpose a matrix inequality is first established. The required inequality for the quadratic form in $A_1\mathbf{y} - A_2\mathbf{y}$, which is the "omnibus inequality" of the paper, will follow from this matrix inequality.

One may well ask why it is worthwhile to study the case of linear functions $A_1\mathbf{y}$ and $A_2\mathbf{y}$ satisfying (1.2). It turns out that quite a few problems, for example in regression diagnostics and outlier detection, can indeed be cast in such a form. Below we shall outline several examples of this; further details and examples (including multidimensional extensions) will be given in Section 3.

Thomas Mathew is Professor, Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250 (E-mail: mathew@umbc2.umbc.edu). Kenneth Nordström is Research Associate, Department of Statistics, University of Helsinki, FIN-00100 Helsinki, Finland. The first author's research was supported by Grant AFOSR F49620-93-1-0001, and the second author's research was supported by a grant from the Ella and Georg Ehrnrooth Foundation. This work was initiated while T. Mathew was visiting the Department of Statistics, University of Helsinki, Finland, and was completed while K. Nordström was a Fulbright Research Scholar at the Department of Mathematics and Statistics, University of Maryland Baltimore County. The authors thank two referees for several valuable suggestions that substantially improved the exposition.

Example 1. Suppose that we wish to compare the ith observation y_i with its predicted value $\hat{y}_i$, predicted from the model (1.1) using the whole data. Such a comparison would naturally be of interest if it is suspected that the ith observation is an outlying observation.

Let $\mathbf{u}_i$ denote the ith standard unit (column) vector with 1 in the ith position and zeros elsewhere, and partition the model matrix X and the projection matrix P row-wise as

$$X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)' \text{ and } P = (\mathbf{p}_1, \ldots, \mathbf{p}_n)'. \qquad (1.3)$$

If we now take $A_1 = \mathbf{u}_i'$ and $A_2 = \mathbf{u}_i'P = \mathbf{p}_i'$, then $A_1\mathbf{y} = y_i$ and $A_2\mathbf{y} = \hat{y}_i$, and both $A_1\mathbf{y}$ and $A_2\mathbf{y}$ clearly have the same expected value $\mathbf{x}_i'\beta$. Therefore, the deviation $A_1\mathbf{y} - A_2\mathbf{y} = y_i - \hat{y}_i$ is of the form considered above; it is, of course, equal to the ordinary residual corresponding to the ith observation in the model (1.1).

Example 2. Consider the case when the model matrix X in (1.1) is the $n \times 1$ vector $\mathbf{1}_n$ of ones, implying that (1.1) is simply the general mean model. Then choosing A_1 and A_2 as in Example 1 above will give us $A_1\mathbf{y} - A_2\mathbf{y} = y_i - \bar{y}$, that is, the deviation between the ith observation and the sample mean. This is the quantity of interest in numerous papers, as reviewed by Olkin (1992).

Example 3. Suppose that we would like to compare an observed response with its predicted value, obtained from fitting the rest of the data. Assume, without loss of generality, that y_1 is the observed response to be compared. Its predicted value, predicted from the data with the first observation excluded, is then given by $\mathbf{x}_1'\hat{\beta}_{(1)}$, where $\mathbf{x}_1'$ is the first row of the matrix X in the partition (1.3) and $\hat{\beta}_{(1)} = (X_{(1)}'X_{(1)})^{-1}X_{(1)}'\mathbf{y}_{(1)}$ is obtained from the partitions

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \mathbf{y}_{(1)} \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} \mathbf{x}_1' \\ X_{(1)} \end{pmatrix}. \qquad (1.4)$$

If we now choose $A_1 = \mathbf{u}_1'$ and $A_2 = (0{:}\,\mathbf{x}_1'(X_{(1)}'X_{(1)})^{-1}X_{(1)}')$, we obtain $A_1\mathbf{y} = y_1$ and $A_2\mathbf{y} = \mathbf{x}_1'\hat{\beta}_{(1)}$; moreover, both $A_1\mathbf{y}$ and $A_2\mathbf{y}$ have the same expected value $\mathbf{x}_1'\beta$. Therefore, the deviation $A_1\mathbf{y} - A_2\mathbf{y} = y_1 - \mathbf{x}_1'\hat{\beta}_{(1)}$ is of the form considered above.

The above comparison could be of interest for several reasons. The resulting predicted residual $e_{(1)} = y_1 - \mathbf{x}_1'\hat{\beta}_{(1)}$ is useful for diagnostic purposes [see, for example, Cook and Weisberg (1982, sec. 2.2.3)], and occurs also as part of the PRESS criterion for selection of models. On the other hand, one might be interested in the predictive capability of the model. Employing a data-splitting technique, y_1 and $\mathbf{x}_1'$ would then correspond to (univariate) validation data, with $y_1 - \mathbf{x}_1'\hat{\beta}_{(1)}$ as a validation residual vector estimating the prediction error; see, for example, Picard and Berk (1990, sec. 2).

Example 4. Assume that we would like to assess the influence of an observation, say the first observation, on the least squares estimator of the vector of regression parameters β in the model (1.1). Choosing $A_1 = (X'X)^{-1}X'$ and $A_2 = (0{:}\,(X_{(1)}'X_{(1)})^{-1}X_{(1)}')$, corresponding to the split (1.4) above, will give us $A_1\mathbf{y} = \hat{\beta}$ and $A_2\mathbf{y} = \hat{\beta}_{(1)}$. Clearly, $A_1\mathbf{y}$ and $A_2\mathbf{y}$ have the same expected value ($= \beta$), and the deviation $A_1\mathbf{y} - A_2\mathbf{y} = \hat{\beta} - \hat{\beta}_{(1)}$ is again of the form considered above.

From the above examples it should be plain that differences of the type $A_1\mathbf{y} - A_2\mathbf{y}$, with A_1 and A_2 satisfying the condition (1.2), are commonplace in problems in linear regression theory. In particular, the quantities that measure the effect of adding or deleting observations (or regressors) in the model (1.1), and that are central in regression diagnostics and outlier detection, can almost all be expressed as differences of the above type. Therefore, it is imperative to assess the magnitude of such differences and to obtain bounds on them. The main result in this article provides not only upper bounds for differences of the type $A_1\mathbf{y} - A_2\mathbf{y}$, but also yields the various F tests that arise for instance in regression diagnostics.

The paper is organized as follows. The basic matrix inequality is given by the lemma in the next section. Some general consequences of this inequality for linear models are given in Section 3.1. The rest of the paper deals with a series of applications to specific problems in linear models, presented in Sections 3.2–3.5. The applications include extensions of Samuelson's inequality, inequalities and tests relating to outlier detection—both single and multiple outliers—and prediction of observations and residuals. Some brief concluding remarks are made in Section 4.

2. A MATRIX INEQUALITY

In this section we establish the desired matrix inequality, that will then be applied in the next sections in order to derive deviation inequalities for linear models. For two real nonnegative definite matrices A and B of the same order, the notation $A \leq B$ denotes that $B - A$ is nonnegative definite, that is, the usual nonnegative partial ordering of matrices.

Lemma. Let X be an $n \times m$ matrix, and let P denote the orthogonal projection matrix onto the range space $\mathcal{R}(X)$ of X, that is, $P = X(X'X)^{-}X'$, with superscript "$-$" denoting generalized inverse.

1. Let A be a $k \times n$ matrix satisfying $AX = 0$. Then

$$A'(AA')^{-}A \leq (I_n - P). \qquad (2.1)$$

2. Let A_1 and A_2 be $k \times n$ matrices satisfying $A_1X = A_2X$. Then

$$(A_1 - A_2)'[(A_1 - A_2)(A_1 - A_2)']^{-}(A_1 - A_2) \leq (I_n - P). \qquad (2.2)$$

Proof. By assumption, $\mathcal{R}(A')$ is orthogonal to $\mathcal{R}(X)$, and because $(I_n - P)$ is the orthogonal projection matrix onto the orthogonal complement of $\mathcal{R}(X)$ [see, for example, Seber (1977, app. B1)], we obtain $\mathcal{R}(A') \subset \mathcal{R}(I_n - P)$. On the other hand, the matrix on the left-hand side of (2.1) is the orthogonal projection matrix onto $\mathcal{R}(A')$, and therefore $(I_n - P) - A'(AA')^{-}A$ is a projection matrix and hence is nonnegative definite [see Result 2 of Seber (1977, app. B3)]. Taking $A = A_1 - A_2$, (2.2) follows directly from (2.1). This completes the proof of the lemma.

Remark 1. If $\mathcal{S}$ is any subspace of R^n and if we use the usual inner product $(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{y}$ (for $\mathbf{x}, \mathbf{y} \in R^n$), then it is well known that projection onto any subspace of $\mathcal{S} \leq$ projection onto $\mathcal{S}$. The inequality (2.1) is simply a reformulation of this fact, as should be clear from the proof of the lemma. The applications in the next section use the inequality in the form stated in (2.2). We also note that the inequality (2.1) is sharp in the sense that equality in (2.1) can be achieved, and this is indeed the case if and only if A satisfies $\mathcal{R}(A') = \mathcal{R}(I_n - P)$.

3. APPLICATIONS TO LINEAR MODELS

3.1 General Results

Consider the standard linear regression model (1.1), and assume that A is a $k \times n$ matrix satisfying $\mathcal{R}(A') \subset \mathcal{R}(X')$. Then the parameter function $A\beta$ is estimable, with least squares estimator (LSE) given by

$$A\hat{\beta} = A(X'X)^{-}X'\mathbf{y}.$$

The residual vector, resulting from the least squares fit $\hat{\mathbf{y}} = P\mathbf{y}$ to the data, is given by $\mathbf{e} = (I_n - P)\mathbf{y}$. Using the standard notation RSS for $\mathbf{e}'\mathbf{e}$, the residual sum of squares, the corresponding estimator of σ^2 is

$$s^2 = \text{RSS}/(n-r)$$

where $r = \text{rank}(X)(\leq m)$.

Now suppose that A_1 and A_2 are two $k \times n$ matrices such that $E(A_1\mathbf{y}) = E(A_2\mathbf{y})$, that is, the matrices A_1 and A_2 satisfy condition (1.2). Then (2.2) gives us the inequality

$$(A_1\mathbf{y} - A_2\mathbf{y})'[(A_1 - A_2)(A_1 - A_2)']^{-}(A_1\mathbf{y} - A_2\mathbf{y}) \leq \text{RSS}. \quad (3.1)$$

The left-hand side of (3.1) can be considered a measure of the deviation between the two estimators $A_1\mathbf{y}$ and $A_2\mathbf{y}$. We note that, apart from the scalar multiple σ^2, the matrix $(A_1 - A_2)(A_1 - A_2)'$, whose generalized inverse appears in (3.1), is the covariance matrix of $(A_1\mathbf{y} - A_2\mathbf{y})$. This shows that the quadratic form

$$Q = (A_1\mathbf{y} - A_2\mathbf{y})'[(A_1 - A_2)(A_1 - A_2)']^{-}(A_1\mathbf{y} - A_2\mathbf{y}) \quad (3.2)$$

is indeed proportional to the squared Mahalanobis (pseudo-)distance between the random variables $A_1\mathbf{y}$ and $A_2\mathbf{y}$; see, for example, Rao (1973, p. 595) or Rousseeuw and Leroy (1987, pp. 223–224).

If rank $(A_1 - A_2) = p$, then, assuming a normal distribution for $\mathbf{y}$, the quadratic forms $(1/\sigma^2)Q$ and $(1/\sigma^2)$RSS have χ^2 distributions with p and $n - r$ degrees of freedom, respectively. Furthermore, because $Q \leq$ RSS, in view of (3.1), the difference $(1/\sigma^2)(\text{RSS} - Q)$ also has a χ^2 distribution with $n - r - p$ degrees of freedom and is distributed independently of Q [see, for example, Rao (1973, p. 187) or Seber (1977, Theorem 2.9)]. Hence

$$F = \frac{Q}{p} \Big/ \frac{(\text{RSS} - Q)}{(n-r-p)} \quad (3.3)$$

follows a central F distribution with degrees of freedom $(p, n-r-p)$. This general fact can be used to test hypotheses in a number of different situations.

We now proceed to describe a series of applications of (3.1) and (3.3).

3.2 Samuelson's Inequality—Some Extensions

Samuelson's (1968) inequality states that for n numbers $z_1, z_2, \ldots, z_n$, the inequality

$$(z_j - \bar{z})^2 \leq \frac{n-1}{n} \sum_{i=1}^{n} (z_i - \bar{z})^2 \quad (3.4)$$

holds for $j = 1, \ldots, n$, where $\bar{z}$ denotes the ordinary arithmetic mean of the z_i's. This inequality thus gives an upper bound on how deviant an observation can be with respect to the mean. The inequality (3.4) was actually known long before the appearance of the article by Samuelson (1968). This is pointed out, for example, in the review article by Olkin (1992), which contains an excellent survey of this literature, along with a matrix proof of (3.4). Below we shall present both a scalar and a vector generalization of (3.4), both following as special cases of (3.1), for appropriate choices of A_1 and A_2.

Consider the problem of comparing the ith observation y_i with its predicted value, predicted from the linear regression model (1.1). This is the problem outlined in Example 1 of Section 1. Let $A_1 = \mathbf{u}_i'$ and $A_2 = \mathbf{u}_i'P$ as in Example 1, and observe that

$$\begin{aligned}(A_1 - A_2)(A_1 - A_2)' &= (\mathbf{u}_i' - \mathbf{u}_i'P)(\mathbf{u}_i' - \mathbf{u}_i'P)' \\ &= \mathbf{u}_i'\mathbf{u}_i - \mathbf{u}_i'P\mathbf{u}_i = 1 - p_{ii}\end{aligned}$$

where p_{ii} is the ith diagonal element of P. Note also that $A_2\mathbf{y} = \mathbf{u}_i'X\hat{\beta} = \mathbf{x}_i'\hat{\beta}$, the LSE of $\mathbf{x}_i'\beta$. The inequality (3.1) thus takes the form

$$(y_i - \mathbf{x}_i'\hat{\beta})^2 \leq (1 - p_{ii})\text{RSS}. \quad (3.5)$$

Inequality (3.5) is clearly a generalization of Samuelson's (1968) inequality (3.4). Indeed, the latter can be obtained from (3.5) by considering the special case of the regression model (1.1) corresponding to the general mean model, that is, by considering $X = \mathbf{1}_n$ (see also Example 2).

Next suppose that $\mathbf{y}$ and X are partitioned as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \quad (3.6)$$

where $\mathbf{y}_1$ and $\mathbf{y}_2$ are, respectively, $n_1 \times 1$ and $n_2 \times 1$ vectors, and X_1 and X_2 are, respectively, $n_1 \times m$ and $n_2 \times m$ matrices. Let

$$A_1 = (I_{n_1} \colon 0) \quad \text{and} \quad A_2 = X_1(X'X)^{-}X'.$$

Then $A_1X = X_1 = A_2X, A_1\mathbf{y} = \mathbf{y}_1$, and $A_2\mathbf{y} = X_1\hat{\beta}$. Also, $(A_1 - A_2)(A_1 - A_2)' = I_{n_1} - P_{11}$, where $P_{11} = X_1(X'X)^{-}X_1'$ is the $n_1 \times n_1$ top left-hand corner submatrix of P. From (3.1) we thus obtain the inequality

$$(\mathbf{y}_1 - X_1\hat{\beta})'(I_{n_1} - P_{11})^{-}(\mathbf{y}_1 - X_1\hat{\beta}) \leq \text{RSS}. \quad (3.7)$$

The inequality (3.7) provides a vector generalization of (3.5) [and (3.4)].

3.3 Residuals and Outlier Detection—Single-Case Results

In this subsection we show that the inequality (3.5) yields as immediate special cases several bounds on standardized residuals in the linear regression model (1.1). Moreover, from the independence of the quadratic forms Q and (RSS $-$ Q) as well as from the corresponding general F statistic (3.3), several distributional results will also be shown to follow at once.

Let us first look at the problem of obtaining a bound on the standardized residuals; this problem is considered, for example, in the recent article by Gray and Woodall (1994). Upon noting that $e_i = y_i - \mathbf{x}_i'\hat{\beta}$, the ordinary residual corresponding to the ith observation, and using the fact that RSS $= s^2(n-r)$, we can rewrite (3.5) in the form

$$e_i^2 \leq (1-p_{ii})s^2(n-r) \tag{3.8}$$

which yields

$$\left|\frac{e_i}{s}\right| \leq \sqrt{(1-p_{ii})(n-r)} \tag{3.9}$$

[where $r = \text{rank}(X)$]. From (3.9) we also obtain directly

$$\max_i \left|\frac{e_i}{s}\right| \leq \sqrt{(1-\min_i p_{ii})(n-r)} \tag{3.10}$$

which gives a common upper bound for all the standardized residuals.

The bound in (3.9) may be contrasted with

$$\left|\frac{e_i}{s}\right| \leq \sqrt{\frac{(n-1)(n-r)}{n}} \tag{3.11}$$

which is the bound given as (1) in Gray and Woodall (1994) (in our notation). It transpires that, although the bound in (3.9) depends on the model matrix X (through the ith diagonal element p_{ii} of the "hat matrix"), it is generally substantially tighter than the bound given by Gray and Woodall (1994) for it is easy to show that $1/n$ is only a lower bound for p_{ii} [see, for example, Cook and Weisberg (1982, p. 12)]. Actually, if the model (1.1) does not include a constant term, then this lower bound will have to be replaced by 0, as pointed out by Cook and Weisberg (1982, p. 13) [see also Rousseeuw and Leroy (1987, p. 220)]. Hence the bound (3.11) is valid only for models with a constant term included.

From (3.8) we also obtain directly the corresponding bound for internally Studentized residuals. Defining the ith internally Studentized residual in the usual way by

$$r_i = e_i/(s\sqrt{1-p_{ii}})$$

the inequality (3.8) reads

$$r_i^2 \leq n-r \tag{3.12}$$

which is the bound given, for example, in Cook and Weisberg (1982, sec. 2.2.1). Furthermore, choosing Q as in (3.2), we have $Q = r_i^2 s^2$, that is, $r_i^2/(n-r) = Q/\text{RSS}$. Because RSS $= Q +$ (RSS $-$ Q) with independent χ^2-distributed summands (see Section 3.1), we also obtain at once that $r_i^2/(n-r)$ follows a Beta distribution, with parameters $((1/2),(1/2)(n-r-1))$ [see Cook and Weisberg (1982, sec. 2.2.1)].

The F ratio (3.3) can now be written as

$$t_i^2 = \frac{e_i^2/(1-p_{ii})}{[(n-r)s^2 - e_i^2/(1-p_{ii})]/(n-r-1)} \tag{3.13}$$

which follows an F distribution with $(1, n-r-1)$ degrees of freedom. The expression in the denominator of (3.13) is easily seen to be $s_{(i)}^2$, the estimator of σ^2 obtained by deleting the ith observation from the estimation data. Therefore, (3.13) can be rewritten as

$$t_i = \frac{e_i}{s_{(i)}(1-p_{ii})^{1/2}} \tag{3.14}$$

which is recognized as the ith externally Studentized (or jackknifed) residual. In view of (3.13), t_i follows a t distribution with $n-r-1$ degrees of freedom.

The t statistic (3.14) is usually employed in the context of a labeled mean-shift outlier model for the purpose of testing whether the ith observation is an outlier. In practice, one is, of course, generally compelled to work with the corresponding unlabeled model, wherein the potentially outlying observation is not specified. This forces consideration of the maximum of t_i, or rather t_i^2, over all i, and leads naturally to standard simultaneous testing procedures; see, for example, Cook and Weisberg (1982, sec. 2.2.2) and Beckman and Cook (1983, sec. 4.2).

3.4 Residuals and Outlier Detection—Multiple-Case Extensions

Below we shall show how the inequality (3.7) can in turn be used to extend the results of Section 3.3 to the vector-valued case. Such an extension is of interest in multiple-case diagnostics for the linear regression model (1.1).

Thus suppose that $\mathbf{y}$ and X are partitioned as in (3.6), and let $\mathbf{e}_1$ denote the corresponding $n_1 \times 1$ residual vector, that is, $\mathbf{e}_1 = \mathbf{y}_1 - X_1\hat{\beta}$. In the labeled case there is clearly no loss of generality in considering the subvector $\mathbf{e}_1$ corresponding to the n_1 first observations because the more general case of an arbitrary index set I is simply obtained by rearranging the observations in the partitions (3.6). (This, of course, is no longer true in the unlabeled case.)

Generalizing the scalar residual r_1, we define the vector of internally Studentized residuals by

$$\mathbf{r}_1 = (I - P_{11})^{+1/2}\mathbf{e}_1/s \tag{3.15}$$

where $(I - P_{11})^{+1/2}$ denotes the Moore–Penrose generalized inverse of $(I - P_{11})^{1/2}$. [For basic properties of the Moore–Penrose inverse, see, for example, Seber (1977, sec. 3.8.1c).] For the case of nonsingular $(I - P_{11})$, the residual vector (3.15) has been defined and used in the literature; see, for example, (2.2.4) in Cook and Weisberg (1982). However, in the multiple-case setting it is not uncommon to have eigenvalues of P_{11} equal to 1, as noted by Cook and Weisberg (1982, p. 13). Therefore, the generalization (3.15) is not simply a theoretical one, but should be of some practical interest.

Now the left-hand side of (3.7) equals $Q = \|\mathbf{r}_1\|^2 s^2$, and therefore (3.7) takes the form

$$\|\mathbf{r}_1\|^2 \leq (n-r) \tag{3.16}$$

where $\|\mathbf{r}_1\|^2 = \mathbf{r}_1'\mathbf{r}_1$, the squared Euclidean norm of $\mathbf{r}_1$. The inequality (3.16) is a considerably stronger result than the corresponding scalar inequality (3.12). The fact that such a vector-valued version is possible, and yields generally stronger results, is related to Remark 1 of Section 2. There it was pointed out that the sharpness of the bound is indeed related to the dimension of $\mathcal{R}(A') = \mathcal{R}(A_1 - A_2)'$ as a subspace of $\mathcal{R}(I_n - P)$.

Assuming a normal distribution for $\mathbf{y}$, and using the independence of the quadratic forms Q and (RSS $-Q$), as done in the previous section, it is seen that $\|\mathbf{r}_1\|^2/(n-r)$ follows a Beta distribution with parameters $((1/2)n_1^*, (1/2)(n-r-n_1^*))$, where $n_1^* = \text{rank}(I_{n_1} - P_{11})(\leq n_1)$. Further, it is seen that the F statistic (3.3) now takes the form

$$F = \frac{\|\mathbf{r}_1\|^2(n-r-n_1^*)}{[(n-r)[3pt] - \|\mathbf{r}_1\|^2]n_1^*}. \tag{3.17}$$

In the case that $(I_{n_1} - P_{11})$ is nonsingular, the above F statistic can be found in Cook and Weisberg (1982, p. 30) [see also (4.9) in Beckman and Cook (1983)], although formulated in terms of the residual vector $\mathbf{e}_1$. Whether a genuine vector formulation such as (3.17), utilizing the vector $\mathbf{r}_1$ of internally Studentized residuals, is available in these sources is not clear to us [see, for example, (4.9) and (4.10) in Beckman and Cook (1983)].

The F statistic (3.17) is used in the labeled mean-shift model to test for multiple outlying observations. As in single-case testing, the unlabeled case leads to problems of simultaneous testing, but the computational effort now required increases dramatically.

3.5 Prediction of Observations and Residuals

Partitioning the data into two (or more) groups of observations, and predicting one (each) group using the other is a common procedure for evaluating the performance of a prediction procedure. The resulting vector of prediction errors is useful for assessing the accuracy of the prediction. Below we shall apply the inequality (3.1) in order to obtain an upper bound for a quadratic form involving the prediction errors in such a setting.

Suppose that the observations from the linear regression model (1.1) are partitioned into two groups, corresponding to the partition (3.6). Suppose further that

$$\mathcal{R}(X_1') \subset \mathcal{R}(X_2') \tag{3.18}$$

so that $X_1\beta$ is estimable using $\mathbf{y}_2$ alone. Then the LSE of $X_1\beta$ based on $\mathbf{y}_2$, say $X_1\hat{\beta}_{(1)}$, is given by $X_1\hat{\beta}_{(1)} = X_1(X_2'X_2)^- X_2'\mathbf{y}_2$. Taking

$$A_1 = (I_{n_1} : 0) \quad \text{and} \quad A_2 = [0 : X_1(X_2'X_2)^- X_2']$$

we have $A_1\mathbf{y} = \mathbf{y}_1$ and $A_2\mathbf{y} = X_1\hat{\beta}_{(1)}$; moreover, both $A_1\mathbf{y}$ and $A_2\mathbf{y}$ have the same expected value $X_1\beta$. A direct computation shows that

$$(A_1 - A_2)(A_1 - A_2)' = I_{n_1} + X_1(X_2'X_2)^- X_1'$$

and hence (3.1) takes the form

$$(\mathbf{y}_1 - X_1\hat{\beta}_{(1)})'\{I_{n_1} + X_1(X_2'X_2)^- X_1'\}^{-1}(\mathbf{y}_1 - X_1\hat{\beta}_{(1)}) \leq \text{RSS}. \tag{3.19}$$

Under the assumption (3.18) it can actually be shown that $I_{n_1} - P_{11}$ is invertible, and that

$$I_{n_1} + X_1(X_2'X_2)^- X_1' = (I_{n_1} - P_{11})^{-1} \tag{3.20}$$

(see Appendix). Therefore, (3.19) can be recast into the form

$$(\mathbf{y}_1 - X_1\hat{\beta}_{(1)})'(I_{n_1} - P_{11})(\mathbf{y}_1 - X_1\hat{\beta}_{(1)}) \leq \text{RSS}. \tag{3.21}$$

The identity (3.20) has an interesting consequence in multiple-outlier detection. Let e_i and $e_{(i)}$ denote the scalar residuals

$$e_i = y_i - \mathbf{x}_i'\hat{\beta} \quad \text{and} \quad e_{(i)} = y_i - \mathbf{x}_i'\hat{\beta}_{(i)},$$

that is, e_i is the ith ordinary residual and $e_{(i)}$ is the corresponding predicted residual, predicted from the data with the ith observation excluded. Then we have

$$\text{var}(e_i) = \sigma^2(1 - p_{ii}) \quad \text{and} \quad \text{var}(e_{(i)}) = \sigma^2/(1 - p_{ii}), \tag{3.22}$$

the latter variance following directly from the relationship between e_i and $e_{(i)}$; see Cook and Weisberg (1982, sec. 2.2.3). The inverse relation between the variances, exhibited in (3.22), is useful for revealing the different roles of e_i and $e_{(i)}$ in identifying cases with large or small diagonal elements p_{ii}, as pointed out in Cook and Weisberg (1982, p. 34).

Now assuming (3.18), the ordinary vector of residuals $\mathbf{e}_1 = \mathbf{y}_1 - X_1\hat{\beta}$ will have a nonsingular covariance matrix given by

$$\text{cov}(\mathbf{e}_1) = \sigma^2(I_{n_1} - P_{11}); \tag{3.23}$$

see the derivation preceding (3.7). Defining the vector of predicted residuals by $\mathbf{e}_{(1)} = \mathbf{y}_1 - X_1\hat{\beta}_{(1)}$, the identity (3.20) shows indeed that

$$\text{cov}(\mathbf{e}_{(1)}) = \sigma^2(I_{n_1} - P_{11})^{-1}. \tag{3.24}$$

A comparison of (3.24) with (3.23) therefore reveals that the covariance matrices of $\mathbf{e}_1$ and $\mathbf{e}_{(1)}$ are inverses of each other, thus extending the relationship between the variances of e_i and $e_{(i)}$ in (3.22). Hence this shows that also the vectors $\mathbf{e}_1$ and $\mathbf{e}_{(1)}$ emphasize different cases, depending on whether the submatrix P_{11} is large or small, relative to I_{n_1} in the matrix ordering. This can be compared favorably with the scalar interpretation pointed out above.

The vector $\mathbf{y}_1 - X_1\hat{\beta}_{(1)}$, which is a generalization of the predicted residuals given in Cook and Weisberg (1982, sec. 2.2.3), also occurs naturally in other contexts. In the literature on cross validation it is often referred to as the validation residual vector; see, for example, Picard and Berk (1990, sec. 2.2).

Remark 2. There are a number of other problems which can be formulated using differences of the type $A_1\mathbf{y} - A_2\mathbf{y}$ considered in this paper. As outlined in Example 4 of Section 1, the problem of assessing the influence of one observation (or a subset of observations) on the least squares estimator of β in the model (1.1) is an example of such a problem. Also, suppose that several independent linear models are available, containing a common parameter vector of interest and perhaps some model-dependent nuisance parameters. Then it may be of interest to assess how much the individual estimators of β differ from the combined (pooled) estimator, or to obtain bounds on the pairwise difference between individual estimators. These problems can indeed also be shown to fall under the general setup considered in this paper. Finally, the matrix results of Section 2 can be similarly used to derive results, parallel to those in Section 3.1, for a multivariate linear model. This extension is entirely straightforward due to the generality of the approach, and should yield directly the corresponding results for regression diagnostics and outlier detection in the multivariate regression model.

Remark 3. Throughout we have assumed that the covariance matrix of $\mathbf{y}$ is $\sigma^2 I_n$, as specified in (1.1). Generalization to the case when the covariance matrix is $\sigma^2 V$, with V positive definite, is straightforward. For a positive definite V we can establish the following generalization of (2.1):

$$A'(AVA')^- A \leq V^{-1}(I_n - P_{X,V^{-1}}) \qquad (3.25)$$

where $P_{X,V^{-1}} = X(X'V^{-1}X)^- X'V^{-1}$ and A is as in (2.1). The inequality (3.25) can be reduced to (2.1) by writing $A_* = AV^{1/2}, X_* = V^{-1/2}X$, and noting that (3.25) is equivalent to $A'_*(A_*A'_*)^- A_* \leq (I_n - P_*)$ [with $P_* = X_*(X'_*X_*)^- X'_*$], which is of the form (2.1). The inequality (2.2) can be generalized similarly.

4. CONCLUDING REMARKS

The main result established in this paper is the inequality (3.1). The derivation of this result is based on the matrix result (2.2). The inequality (3.1) also yields a general F statistic (3.3). As consequences of the single inequality (3.1) and 3), we have in Sections 3.2–3.5 rederived several known results, and established some new ones, mostly in the area of regression diagnostics and outlier detection.

The matrix-algebraic techniques used to arrive at the general results (3.1) and (3.3) are rather elementary, and should be familiar to anyone who has had a course in standard linear model theory. The inequality (3.1) and the F ratio (3.3) provide a considerable unification of a number of results in regression diagnostics, and should therefore be of interest in teaching courses on regression diagnostics as well as to researchers in this area.

APPENDIX

In order to establish (3.20), we need to show that under (3.18), that is, when $\mathcal{R}(X'_1) \subset \mathcal{R}(X'_2)$ [with respect to the partition (3.6)], the following is true:

$$[I_{n_1} + X_1(X'_2X_2)^- X'_1](I_{n_1} - P_{11}) = I_{n_1}. \qquad (A.1)$$

Noting that $P_{11} = X_1(X'X)^- X'_1$, (A.1) follows if we can show that

$$X_1(X'_2X_2)^- X'_1 - X_1(X'X)^- X'_1 \\ - X_1(X'_2X_2)^- X'_1X_1(X'X)^- X'_1 = \mathbf{0}. \qquad (A.2)$$

Observe that

$$\begin{aligned} &X_1(X'_2X_2)^- X'_1X_1(X'X)^- X'_1 \\ &\quad = X_1(X'_2X_2)^-(X'X - X'_2X_2)(X'X)^- X'_1 \\ &\quad = X_1(X'_2X_2)^- X'_1 - X_1(X'_2X_2)^- X'_2X_2(X'X)^- X'_1 \\ &\quad = X_1(X'_2X_2)^- X'_1 - X_1(X'X)^- X'_1 \text{ [using (3.18)].} \end{aligned} \qquad (A.3)$$

Thus (A.2) follows by substituting the left-hand side of (A.2) by the expression in (A.3). This concludes the proof of (3.20).

[Received August 1993. Revised March 1995.]

REFERENCES

Beckman, R. J., and Cook, R. D. (1983), "Outlier . . .s," *Technometrics*, 25, 119–149.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman and Hall.

Gray, J. B., and Woodall, W. H. (1994), "The Maximum Size of Standardized and Internally Studentized Residuals in Regression Analysis," *The American Statistician*, 48, 111–113.

Hawkins, D. M. (1980), *Identification of Outliers*, London: Chapman and Hall.

Olkin, I. (1992), "A Matrix Formulation on How Deviant an Observation Can Be," *The American Statistician*, 46, 205–209.

Picard, R. R., and Berk, K. N. (1990), "Data Splitting," *The American Statistician*, 44, 140–147.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: Wiley.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Samuelson, P. A. (1968), "How Deviant Can You Be?," *Journal of the American Statistical Association*, 63, 1522–1525.

Seber, G. A. F. (1977), *Linear Regression Analysis*, New York: Wiley.