COMMUN. STATIST.-THEORY METH., 16(9), 2517-2540 (1987)

ON THE N-DIMENSIONAL GEOMETRY OF REGRESSION DIAGNOSTICS

R.W. Oldford

Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada

Keywords and Phrases: collinearity, effective dimension, influence, leverage, linear model, minimax estimation, numerical ranks, ridge regression

ABSTRACT

The n-dimensional geometry of collinearity and data that are influential in least-squares linear regression is explored. A generalization of vector space dimensionality is introduced to provide an intuitive description of these problems. It is also noted that this new measure of dimensionality plays the role of the usual dimension in a James-Stein like result. Some common regression diagnostics are critically examined in this geometric framework.

1. Introduction

Consider the linear model

$$y = \mathbf{X}b + e \tag{1.1}$$

where $b \epsilon \mathbb{R}^m$, $\mathbf{X} = (X_1, \ldots, X_m)$ and y, e, $X_1, \ldots, X_m \epsilon \mathbb{R}^m$. The geometry of this model is well-known. The vector y is to be fitted by some vector, $\mathbf{X}b$, lying in the subspace generated by the vectors $X_1, \ldots, X_m$. The vector e lies in a complementary subspace of $\mathbb{R}^n$ and gives the error of the fit $\mathbf{X}b$.

Standard statistical methods like least-squares regression, the analysis of variance and the analysis of covariance, are easily understood using this n-dimensional geometry. The purpose of the present paper is to incorporate newer techniques, like regression diagnostics, into this n-dimensional framework so that these techniques might also enjoy a more intuitive understanding in light of their geometry.

For this purpose, the primary strength of the usual geometry, namely its reliance on the theory of vector spaces, becomes a major weakness. In practice, individual vectors are observed, not entire vector spaces. Considering only vector spaces ignores information specific to the individual vectors that generated them. This is unfortunate since diagnostic issues typically depend upon precisely that kind of information: either individual observation vectors or individual variate vectors.

As an example, consider determining the rank of a given matrix $\mathbf{X}$. Geometrically, the rank is a function only of the column space of $\mathbf{X}$, namely its dimension. However, if $\mathbf{X}$ consists of data subject to error, or if the rank is to be determined numerically, then each element of $\mathbf{X}$ is known only up to its measurement accuracy (which at best is no worse than the precision of the machine used to compute the rank). If only the vector space generated by the columns of $\mathbf{X}$ is considered, then this information is missed and the resulting rank may be an overestimate. Consequently, classic vector space description of linear modelling does not easily accommodate a practical problem like collinearity: the problem disappears with an appropriate reparameterization (i.e. for a non-singular matrix $\mathbf{A}$, $\mathbf{X}$ and b in (1.1) may be replaced by $\mathbf{XA}$ and $\mathbf{A}^{-1}b$, respectively, without changing the subspace containing the fitted vector).

The geometric descriptions which follow supplement the classic n-dimensional theory by focussing on the particular vectors which define the matrix (e.g. the vectors $X_1, \dots, X_m$ rather than their span). A class of simple functions of these vectors, which are known to have many properties usually associated with the dimension (except that positive fractional values are also obtainable), is introduced in section 2. These functions (d_a-effective dimensions) and their constituent parts are used throughout the remaining sections to give an intuitive geometric description of regression diagnostics.

Section 3 records a minimax result by Thisted and Morris [1980] where the d_1-effective dimension plays a role completely analogous to that played by the vector space dimension in James-Stein estimation. The effective dimension from this theorem has been suggested as a useful collinearity diagnostic by Thisted [1982]. This suggestion is considered and criticized in section 4. There the effective dimension is used to describe the difference between the minimaxity result and the diagnosis of collinearity. Section 5 considers the n-dimensional geometry of influential data in least-squares regression and section 6 critically examines two common diagnostic statistics: Cook's distance and the Andrews-Pregibon statistic. Finally, some concluding remarks are made in section 7.

2. Effective Dimension

Suppose we have two matrices $\mathbf{S}$ and $\mathbf{P}$ say, whose column vectors are related as follows. Each column vector of $\mathbf{S}$ equals a constant in the interval $[0,1]$ times some column vector in $\mathbf{P}$, and no vector in $\mathbf{P}$ appears more than once in $\mathbf{S}$ (e.g., if $\mathbf{P} = (P_1,P_2)$ then $\mathbf{S} = (.5P_1)$, $\mathbf{S} = (.5P_1,.8P_2)$ and $\mathbf{S} = \mathbf{P}$ are all allowed, but $\mathbf{S} = (.5P_1,.8P_1)$ is not). For two such matrices, the d_a-*effective dimension* of $\mathbf{S}$, with respect to $\mathbf{P}$, is defined to be

$$d_a(\mathbf{S};\mathbf{P}) = (\lambda_1(\mathbf{P}))^{-a} \sum_i (\lambda_i(\mathbf{S}))^a$$

where $\lambda_i(.)$ denotes the i^{th} largest singular value of its matrix argument and $a>0$. This class of functions can be shown to share many properties with the usual dimension. Indeed, the class can be extended to include the usual dimension by allowing $a = 0$. Formal mathematical theory exists which motivates and describes properties of the d_a-effective dimension (see Oldford [1983, 1985] for details). Here we note only two major differences between the d_a-effective dimension and the dimension of the entire column space of **S**.

First, if we regard the set of column vectors in **S** as providing uncertain information about the intended column space, then some directions of the column space will be better established than others by these vectors. The contribution to the effective dimension is discounted more for poorly established directions than for others. Thus, the ratio $\eta_i \equiv \lambda_i(\mathbf{S})/\lambda_1(\mathbf{P})$, which we call the i^{th} *dimension index* of **S**, with respect to **P**, represents the amount the i^{th} principal direction of **S** actually has to contribute, relative to **P**, to the dimensionality of the column space of **S**. Each d_a-effective dimension simply raises these contributions to a fixed power a and computes their sum.

Compare these dimension indices to the statistics commonly used to numerically determine the rank of **S**. The condition number $\lambda_1(\mathbf{S})/\lambda_{\min}(\mathbf{S})$, for example (Wilkinson [1965]), is simply the inverse of $\eta_{\min}$ of **S** (with respect to itself). **S** is declared rank-deficient if $\eta_{\min}$ is too small. Similarly, Chambers [1977] has suggested the numerical rank of $\mathbf{S}$ be taken to be the number of dimension indices greater than some cutoff. The d_a-effective dimension instead sums each dimension-index raised to a.

The second major difference is that the d_a-effective dimension has two arguments, $\mathbf{S}$ and $\mathbf{P}$, so that the d_a-effective dimension of $\mathbf{S}$ is measured with respect to $\mathbf{P}$. This second argument simply provides a context for scaling the relative importance of each principal direction. Clearly for each $\mathbf{P}$, many $\mathbf{S}$s are possible and vice versa. In this framework, the usual dimension measures the rank of $\mathbf{S}$ with respect to $\mathbf{S}$.

The next section gives an example where the d_1-effective dimension plays the role of the dimension in a James-Stein like result.

3. A Minimax Result

Given the model (1.1), assume that the errors, e_i, are independent and identically distributed as $N(0,\sigma^2)$ with known $\sigma^2>0$ so that b is now a distributional parameter to be estimated. An alternative to the least squares estimator, b_{LS}, is the adaptive Ridge-A estimator (Thisted [1982]) given by (when shrinking b_{LS} to 0)

$$b_A = (\mathbf{X}^T\mathbf{X} + k^2\mathbf{I}_m)^{-1}\mathbf{X}^T y \tag{3.1}$$

where $k^2 = A\sigma^2/(b_{LS}^T\mathbf{V}\,\mathbf{D}_w\mathbf{V}^T b_{LS})$. Here $A>0$, $\mathbf{D}_w$ equals some diagonal matrix of weights, and $\mathbf{V}$ is the matrix having the eigenvectors of $\mathbf{X}^T\mathbf{X}$ as columns. Further, suppose that the expected loss of an estimator δ of b can be given by $E[(\delta-b)^T\mathbf{L}(\delta-b)]$ for some positive semidefinite matrix of **L**. Thisted [1982] has shown that the d_a-effective dimension plays much the same role in this setting as does the dimension in the well-known James-Stein result. Letting

$$\mathbf{S} = \mathbf{D}_w^{1/2}\mathbf{V}^T Var(b_{LS})\mathbf{L}\ Var(b_{LS})\mathbf{V}\mathbf{D}_w^{1/2}, \tag{3.2}$$

the following result is proved by Thisted and Morris [1980] and may be found in Thisted [1982].

Proposition 1: For suitable choices of $A \geq 0$, Ridge-A estimators given by (3.1), are minimax with respect to the above loss function, if and only if $d_1(\mathbf{S};\mathbf{S})>2$.

Note that whereas the James-Stein result required a dimension larger than 2, the above result requires a $d_1>2$. For this reason, Thisted [1982] has called $d_1(\mathbf{S};\mathbf{S})$ *the* effective dimension. Since d_a shares many geometric properties with the usual dimension for values of a other than $a = 1$, the term d_1-effective dimension is preferred here.

4. Collinearity

In a collinearity analysis† it helps to distinguish between these procedures used to detect the presence of collinearity and those used to ascertain its effect on the problem of interest.

For detection, consider the matrix $\mathbf{X} = (X_1, \ldots, X_m)$, where each $X_i \epsilon \mathbb{R}^n$. Given that the X_i's are in a structurally interpretable form

† Recently, formal definitions have beeb proposed by Gunst [1984], and by Belsley and Oldford [1986].

(see Belsley [1984] and Belsley and Oldford [1986] for discussion), collinearity is judged to be present if at least one of the dimension indices, η_i, of $d_1(\mathbf{X};\mathbf{X})$ is small†† and inestimability occurs if one or more are zero. Thus, collinearity is present if at least one orthogonal direction is not well determined. Further, the η_i's, whose inverses are called "condition" indices by Belsley, Kuh, and Welsch [1980], are used to assess the extent of the collinearity. The greater the number of poorly determined orthogonal directions of $\mathbf{X}$, the more extensive is the collinearity.

Consider now the effect on the ridge-A estimator. Thisted [1980, 1982] has suggested that the d_1-effective dimension of Proposition 1 be used to assess the effect of collinearity on the minimax property of the ridge-A estimators. As will be demonstrated, the statistic itself is not at all related to the presence or absence of collinearity. However, since ridge-A estimators are often suggested in place of the least-squares estimator when collinearity is present, it is of interest in this case to see when, and why, minimaxity is obtained.

In particular, let $\mathbf{S}_1$ and $\mathbf{S}_2$ denote the matrix of (3.2) with $\mathbf{D}_w = \mathbf{I}$, when $\mathbf{L} = \mathbf{I}$ and when $\mathbf{L} = \mathbf{X}^T\mathbf{X}$, respectively. It can be shown that

†† Based on experimental evidence, Belsley, Kuk, and Welsch [1980] suggest that those η_i's less than 0.033 be regarded as small.

$$d_1(\mathbf{S}_1;\mathbf{S}_1) = \lambda_1(\mathbf{S}_1)^{-1}\sum\lambda_i(\mathbf{S}_1)$$

$$= \lambda_m(\mathbf{X})^4\sum\lambda_i(\mathbf{X})^{-4} \tag{4.1}$$

and

$$d_1(\mathbf{S}_2;\mathbf{S}_2) = \lambda_1(\mathbf{S}_2)^{-1}\sum\lambda_i(\mathbf{S}_2)$$

$$= \lambda_m(\mathbf{X})^2\sum\lambda_i(\mathbf{X})^{-2}. \tag{4.2}$$

Thisted [1980, 1982] has called $d_1(\mathbf{S}_1;\mathbf{S}_1)$ and $d_1(\mathbf{S}_2;\mathbf{S}_2)$ the multicollinearity index (*mci*) and the predictive multicollinearity index (*pmci*), respectively. From Proposition 1, each quantity is related to the minimaxity of a particular $(\mathbf{D}_w = \mathbf{I})$ ridge-A estimator, first when the expected loss is that of the mean-square-error of the estimator $(\mathbf{L} = \mathbf{I})$ and second when the expected loss is that of the mean-square-error of the predicted response at the observed $\mathbf{X}\,(\mathbf{L} = \mathbf{X}^T\mathbf{X})$. Values of *mci* or *pmci* less than two indicate that the corresponding minimax property is lost.

That *mci* and *pmci* bear no relationship to the presence or absence of collinearity, as assessed by the dimension indices η_i of $d_1(\mathbf{X};\mathbf{X})$, is easily demonstrated by an example. Let $\mathbf{X}$ be of full mathematical rank $m = 4$ and denote by Λ the row-vector of ordered singular values of $\mathbf{X}$, written as $\Lambda = (\lambda_1(\mathbf{X}), \ldots, \lambda_4(\mathbf{X}))$. Now consider the following three possibilities for Λ,

(i) $\Lambda_1 = (1,1,1,\epsilon)$

(ii) $\Lambda_2 = (1,1,\epsilon,\epsilon)$

$$\text{(iii)} \quad \Lambda_3 = (1,\epsilon,\epsilon,\epsilon)$$

where $0<\epsilon<1$. Corresponding to each case are the values of $d_1(\mathbf{S}_1;\mathbf{S}_1)$ and $d_1(\mathbf{S}_2;\mathbf{S}_2)$,

$$\text{(i)} \quad (1+3\epsilon^4) \text{ and } (1+3\epsilon^2)$$

$$\text{(ii)} \quad (2+2\epsilon^4) \text{ and } (2+2\epsilon^2)$$

$$\text{(iii)} \quad (3+\epsilon^4) \text{ and } (3+\epsilon^2).$$

Suppose first that $\epsilon = 1/5$. In all cases, the condition number of the X-matrix which results is five and collinearity is not likely to be judged present. However, the values of $d_1(\mathbf{S};\mathbf{S}_1)$ are (i) 1.0048, (ii) 2.0032, and (iii) 3.0018 giving minimaxity of the ridge-A estimator in the last two cases but not in the first. Now suppose that $\epsilon = 10^{-5}$, yielding 100,000 as the condition number of $\mathbf{X}$. Most likely, collinearity will be judged to be present. But the minimaxity or not of the ridge-A estimator remains the same in each case as when $\epsilon = 1/5$. Indeed, when there are three out of four mutually orthogonal linear combinations of the parameters which are very nearly inestimable, as in case (iii), the minimax property of the estimator is assured, whereas in the case (i) of least extensive collinearity the minimaxity is lost.

Although *mci* and *pmci* and their dimension indices are of little use for the general diagnosis of collinearity, they do provide interesting geometric information about the minimaxity of the ridge-A estimator. The ridge-A estimator (3.1) can be thought of as an estimator which shrinks the least squares estimates toward zero. Those parameter estimates shrunk most are those which have the greatest variance. Letting $\Gamma = \mathbf{V}^T b$ with b and $\mathbf{V}$ as in (3.1), then the components γ_i of Γ have variances equal to $\sigma^2\lambda_i^{-2}(\mathbf{X})$. In the above examples, these correspond to the following row-vectors of variances,

(i) $(\sigma^2, \sigma^2, \sigma^2, \sigma^2/\epsilon^2)$

(ii) $(\sigma^2, \sigma^2, \sigma^2/\epsilon^2, \sigma^2/\epsilon^2)$

(iii) $(\sigma^2, \sigma^2/\epsilon^2, \sigma^2/\epsilon^2, \sigma^2/\epsilon^2)$.

For small $\epsilon>0$, the variance σ^2 is negligible when compared to σ^2/ϵ^2. In (i) this means that there is essentially only one least squares estimate, γ_4, with non-negligible variance, or equivalently, there is effectively only one random quantity to shrink. Not until ϵ^2 (or ϵ^4) is greater than 1/3 does $d_1(\mathbf{S}_2;\mathbf{S}_2)$ (or $d_1(\mathbf{S}_1;\mathbf{S}_1)$) produce a value larger than two. In terms of variance, then, as long as $Var(\gamma_4)\geq 9\ Var(\gamma_i)$ for $i \neq 4$, the γ_i for $i \neq 4$ act as fixed quantities compared to γ_4. This interpretation makes sense of the fact that in case (i) large values of ϵ, which might properly be ignored by a collinearity detection diagnosis, cannot be tolerated by the minimaxity property. Similar remarks and interpretations apply to cases (ii) and (iii).

The example shows that the minimaxity of the ridge-A estimator and collinearity of the explanatory variates are really two quite different problems. And hence it casts serious doubt on any recommendation based on minimax grounds for the ridge-A estimator as a panacea for collinearity.

In light of the above example and given its geometric interpretation, it is tempting to interpret the d_1-effective dimension as the effective number of explanatory variables in the regression model (1.1) for collinearity purposes, and as the effective number of least-squares parameter estimates which might reasonably be regarded as random quantities to determine the minimaxity of the ridge-A estimator. This agrees well with our intuition on these issues and hence has pedagogic value.

In the next section this kind of informal interpretation presents itself again.

5. Influential Observations in Least-squares Linear Regression

Suppose the model (1.1) is fitted by least-squares and that $\hat{y}$ and $\hat{e}$ are the fitted and residual vectors, respectively. If k observations are suspect, they can be eliminated from the least-squares fit by expanding the model to include k new parameters, one for each observation. The expanded model will be

$$y = \mathbf{X}b + \mathbf{X}^{*}\delta + e \tag{5.1}$$

where δ is $k \times 1$ and $\mathbf{X}^{*} = (\mathbf{0}, \mathbf{I}_k)^T$ is an $n \times k$ matrix of zeros and ones. Without loss of generality, we have taken the last k observations as the suspect ones. Many peculiarities of the data and fit can be incorporated in this manner (e.g. Andrews[1971]).

Let $\hat{y}^{*}$ and $\hat{e}^{*}$ denote the vectors of fitted values and residuals, respectively, from fitting (5.1) by least-squares. The influence the k

observations have on the original fit can be measured by the difference in fits $\hat{y}^* - \hat{y} = \hat{e} - \hat{e}^* = \Delta e$, say.

Now consider the geometry of these two fits. Let $\mathbf{P} = (P_1, \cdots, P_n)$ be the orthogonal projection matrix of the error space associated with the least-squares fit of the original model (1.1) so that $\hat{e} = \mathbf{P}y$. Denote the error space for (1.1) by $<\mathbf{P}> = span(P_1, \cdots, P_n)$, the column space of $\mathbf{P}$.

The k parameters introduced in (5.1) affect the size of the error space. All vectors in the span of the orthogonal projection of $\mathbf{X}^*$ onto the error space $<\mathbf{P}>$ of (1.1) are excluded from the error space of (5.1). If we let $<\mathbf{S}>$ denote the column space of $\mathbf{PX}^*$ then the error space of (5.1) is simply the orthogonal complement, $<\mathbf{S}>^{\perp}$ say, of $<\mathbf{S}>$ in $<\mathbf{P}>$.

Now if $\mathbf{P}_{<S>}$ and $\mathbf{P}_{<S>^{\perp}}$ denote the orthogonal projection matrices for $<\mathbf{S}>$ and $<\mathbf{S}>^{\perp}$ respectively, then, $\mathbf{P} = \mathbf{P}_{<S>} + \mathbf{P}_{<S>^{\perp}}$ and we have

$$\Delta e = \mathbf{P}_{<S>}y = \mathbf{P}_{<S>}\hat{e} \ . \tag{5.2}$$

Thus, the difference in least-squares fits for the models (5.1) and (1.1) is just the projection of the original residual vector $\hat{e}$ onto $<\mathbf{S}>$. In particular, the closer $<\mathbf{S}>$ is to $\hat{e}$ the greater is this difference. Summarizing the difference by the angle $\theta \epsilon [0,\frac{\pi}{2}]$ between $<\mathbf{S}>$ and $\hat{e}$ is equivalent to the well known extra-sum-of-squares principle to test the model of (1.1) versus that of (5.1).

Applying these results, which are true for general $\mathbf{X}^*$, to the particular case $\mathbf{X}^* = (\mathbf{0},\mathbf{I}_k)^T$ has led to focus on the angle θ (e.g. see Andrews [1971] and Dempster and Gasko-Green [1981]). Indeed, the test based on the extra-sum-of-squares principle has been advocated by Gentleman [1980] and Draper and John [1981] as a test for outliers.

However, closer examination of (5.2) and the particular vectors $P_{n-k+1}, \cdots, P_n$ given to generate $<\mathbf{S}>$ will show that θ alone misses much of the information on the influence of the k observations. Equation (5.2) may be rewritten as

$$\Delta e = \mathbf{U}(\mathbf{U}^T\hat{e}) \tag{5.3}$$

where $\mathbf{U}$ is the $n\times k$ matrix from the singular value decomposition of $\mathbf{S}$ $(\mathbf{U}^T\mathbf{U}=\mathbf{I}_k)$. The change in fit is thus expressible as the product of two interpretable factors. The first factor $\mathbf{U}$, together with the singular values of $\mathbf{S}$, describes the structure of $\mathbf{S}$ while the second factor, $\mathbf{U}^T\hat{e}$, describes the orientation of $\mathbf{S}$ to $\hat{e}$. These two components, the structure of $\mathbf{S}$ and its orientation to $\hat{e}$, provide simple starting places to generate summaries of the information in Δe.

The structure of $\mathbf{S}$ is the source of "leverage"[†] or *potential* influence that the group of k observations may have on the determination of

[†] This differs slightly from that of Hoaglin and Welsch [1978] where the leverage of a single observation is described in terms of the potential influence it has on its own fit.

the least-squares fit of the remaining $n-k$ observations. To see this, first note that the structure of $\mathbf{S}$ may be summarized by its principal direction vectors (the column vectors of $\mathbf{U}$) and its dimension indices $\eta_1, \cdots, \eta_k$ of $\mathbf{S}$ with respect to $\mathbf{P}$ (since $\mathbf{X}^* = (0,\mathbf{I}_k)^T$, $\eta_1, \cdots, \eta_k$ are well-defined). Further, since $\lambda_{\max}(\mathbf{P}) = 1$, $\eta_1, \cdots, \eta_k$ are simply the singular values of $\mathbf{S}$. Now, letting $\pi = diag(\eta_1, \cdots, \eta_k)$ and partitioning $\mathbf{U}$ as $\mathbf{U}^T = (\mathbf{U}_1^T, \mathbf{U}_2^T)$ where $\mathbf{U}_2$ has k rows, it is easily shown that

$$\mathbf{U}_2^T\mathbf{U}_2 = \P^2$$

and

$$\mathbf{U}_1^T\mathbf{U}_1 = \mathbf{I}_k - \P^2. \tag{5.4}$$

Together (5.3) and (5.4) indicate that the fit will be perturbed only at the k suspect observations if, and only if, $\eta_1 = \cdots = \eta_k = 1$ (regardless of $\hat{e}$). If some of these dimension indices are small, then the fit of the other observations may be perturbed. The extent of the perturbation will also depend upon $\hat{e}$. Therefore examination of just $\mathbf{U}$ and the dimension indices can determine only the *potential* influence of the group, hence the word "leverage"[†] Note that, as was the case with collinearity, to determine the presence of high leverage it is best to examine the entire set of dimension indices rather than any single uni-dimensional summary like $d_1(\mathbf{S};\mathbf{P})$.

[†] When $k = 1$, $\eta_1^2 = ||P_n||^2 = 1-h_n$ where h_n is the nth diagonal element of the "hat matrix" $\mathbf{H} = \mathbf{I}-\mathbf{P}$. Thus for the one-at-a-time case this notion of leverage corresponds directly to the self-influence one given by Hoaglin and Welsch [1978].

The second factor of (5.3), describing the orientation of $\mathbf{S}$ to $\hat{e}$, is the source of measures of group outlyingness based on θ, the acute angle $<\mathbf{S}>$ makes with $\hat{e}$ (N.B. $||\mathbf{U}^T\hat{e}|| = ||\hat{e}||\cos\theta$). However, more detailed information on this orientation can be found in $\mathbf{U}^T\hat{e}$. In particular, if θ_i denotes the acute angle between the *ith* principal direction of $\mathbf{S}$ and $\hat{e}$, we have $\mathbf{U}^T\hat{e} = ||\hat{e}||(\cos\theta_1, \cdots, \cos\theta_k)^T$. Thus, from (5.3) and (5.4) it can be seen that the angles $\theta_1, \cdots, \theta_k$ also contain important information about the change in fit.

For example, suppose that $\eta_k < 1$ and all other η_i's equal 1. By (5.4), only the last column of $\mathbf{U}_1$ is non-zero and the fit of the first $n-k$ observations will be perturbed, if, and only if, $\hat{e}$ is not orthogonal to the k^{th} column vector of $\mathbf{U}$. The amount of the perturbation will increase as θ_k decreases, regardless of the value taken by θ. The value of θ is only a lower bound for θ_k.

It is possible, therefore, that a group of k observations can be both an outlying group (small θ) and a high leverage group (small η_k) but still have little or no impact on the fit of the remaining $n-k$ points. Such a situation will occur when $\hat{e}$ is orthogonal to all principal directions of $\mathbf{S}$ whose corresponding dimension indices, η_i, are much less than 1. More detailed information on the orientation of $\mathbf{S}$ to $\hat{e}$, involving the k angles $\theta_1, \ldots, \theta_k$, say, between $\hat{e}$ and the principal directions of $\mathbf{S}$, is needed to determine the *actual* influence the k observations have on the remaining ones.

Separate examination of the structure of $\mathbf{S}$ in $\mathbf{P}$ and the orientation of $\mathbf{S}$, to $\hat{e}$, therefore suggests that the n-dimensional information contained in Δe might be reasonably summarized by the two k-dimensional statistics $(\eta_1, \ldots, \eta_k)$ and $(\theta_1, \ldots, \theta_k)$. These statistics correspond roughly to the "leverage" and "outlyingness" components of the influence of a group of k observations, respectively. Together, they can give some indication of the actual influence. However, if many groups of k observations are to be examined and k is large enough, then practicality will require still lower dimensional summaries.

As with collinearity, it is tempting here to interpret the d_1-effective dimension as a measure of the effective number of observations determining the fit. When k observations are overly influential the error vector $\hat{e}$ will lie close to a small k dimensional space $<\mathbf{S}>$. By considering $\mathbf{S}$, the evidence given by the dimension indices may indicate that few of the orthogonal directions of $\mathbf{S}$ are well-determined and so even the dimensionality of k is suspect. Further, having $\hat{e}$ lie close to a principal direction of $\mathbf{S}$ associated with a small dimension index causes the influence of these points to be greater still. Influential points are associated with having $\hat{e}$ in close proximity to a subspace, $<\mathbf{S}>$, whose matrix of generating vectors has small effective dimension.

Considering weighted least-squares reinforces this interpretation. In (3.1), b would be estimated by $(\mathbf{X}^T\mathbf{D}_w\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}_w y$ where $\mathbf{D}_w$ is a diagonal matrix of weights in $[0,1]$. The effect of downweighting the influential observations can be seen by taking $\mathbf{P}$ and $\hat{e}$ to be $\mathbf{I}-\mathbf{D}_w^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{D}_w\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}_w^{1/2}$ and $\mathbf{D}_w^{1/2}\mathbf{P}y$. Downweighting k observations will have the following effects on the corresponding $\mathbf{S}$ and $\hat{e}$. The vectors of $\mathbf{S}$ will be elongated and spread apart, and the angle between $\hat{e}$ and $\mathbf{S}$ will be increased, thus increasing $d_1(\mathbf{S};\mathbf{P})$ as the weights decrease. These effects are most dramatic when $\mathbf{S}$ contains all k influential points. The geometric effect of influential points in least-squares and of downweighting these points coincides with the intuition that a few influential points may effectively determine the fit, and that downweighting these points will increase the effective number of observations contributing to the fit.

6. Two popular k-at-a-time diagnostics

A number of uni-dimensional statistics have been suggested in the literature. Two of them, Cook's [1977] statistic and the Andrews-Pregibon [1978] statistic are now described geometrically. These have also been discussed by Dempster and Gasko-Green [1981], and by Draper and John [1981]. As will be shown, each captures the structure of $\mathbf{S}$ (leverage) and its orientation to $\hat{e}$ (outlyingness) in different ways.

Let Q be the extra sum of squares, $\Delta e^T\Delta e$, due to fitting the model (5.1). The angle, θ, between $\hat{e}$ and the span of the vectors $P_{n-k+1}, \ldots, P_n$ may be measured by

$$\cot^2\theta = \frac{Q}{\hat{e}^T\hat{e}-Q}$$

or equivalently

$$\sin^2\theta = \frac{\hat{e}^T\hat{e}-Q}{\hat{e}^T\hat{e}} .$$

One-at-a-time statistics combine this outlier information with the leverage information, $||P_i||^2$, in simple ways (see Dempster and Gasko-Green [1981]).

For k-at-a-time diagnostics, the Andrews-Pregibon (AP) and Cook (C) statistics may be expressed, up to a multiplicative constant, as follows,

$$AP = \sin^2\theta \ . \ \det(\mathbf{P}_{22}) = \sin^2\theta \prod_{i=1}^{k} \eta_i^2$$

and

$$C = \cos^2\theta \left(\frac{\hat{e}^T\mathbf{P}_2\mathbf{P}_{22}^{-2}\mathbf{P}_2^T\hat{e}}{Q} - 1 \right). \qquad (6.1)$$

where $\mathbf{P}_2 = (\mathbf{P}_{21}, \ \mathbf{P}_{22})^T$ is the $n\times k$ matrix given by the last k columns of the projection matrix $\mathbf{P}$ of Section 5. Small values of AP and large C identify "influential" groups.

Since a uni-dimensional statistic is used in either case, care must be taken when combining the two sources of information. For example, having the "outlier" part, θ, enter each statistic multiplicatively through $\sin^2\theta$ and $\cos^2\theta$ has certain drawbacks. In particular, if the fit including the i^{th} observation is exactly the same as that excluding it, then

$\cos^2\theta = 0$ and AP may be small if $||P_i||^2 = (1-h_i)$ is small. Draper and John [1981] recommended C over AP for this reason alone. However, if $\cos^2\theta = 1$, then removal of the i^{th} observation gives a perfect fit to the remaining observations and C does not detect this. This has been pointed out by Dempster and Gasko-Green [1981]. Using the cotangent in place of the cosine, or the tangent in place of the sine, removes these difficulties, but may, in the above uni-dimensional statistics, emphasize the "outlier" part of each statistic too much.

While the AP statistic factors simply into an "outlier" part, $\sin^2\theta$, representing the orientation of $<\mathbf{S}>$ to $\hat{e}$, and a "leverage" part, $\Pi\eta_i^2$, summarizing the structure of $\mathbf{S}$, this is not the case for the C statistic. Its second factor in (6.1) is not a simple "leverage" component. Rather, the second factor summarizes both the structure of $\mathbf{S}$ and the orientation of $\mathbf{S}$ to $\hat{e}$. This may be seen by reexpressing this factor $(+1)$ as

$$\frac{1}{\cos^2\theta} \cdot \sum_{i=1}^{k} \left(\frac{\cos\theta_i}{\eta_i} \right)^2 \tag{6.2}$$

where θ_i is the angle between $\hat{e}$ and the i^{th} principal direction of $\mathbf{S}$. Clearly, both the structure of $\mathbf{S}$ and its orientation to $\hat{e}$ are captured by (6.2). Indeed (6.1) can now be rewritten as

$$C = \sum_{i=1}^{k} \left(\frac{\cos\theta_i}{\eta_i} \right)^2 - \cos^2\theta. \tag{6.3}$$

The closter $\hat{e}$ lies to a principal direction of **S** which has a small dimension index the larger is C. This is a fundamental difference between C and AP. The Andrews-Pregibon statistic uses the orientation of $<\mathbf{S}>$ to $\hat{e}$ whereas the Cook statistic uses the orientation of **S** to $\hat{e}$ with that of $<\mathbf{S}>$ to $\hat{e}$ appearing more as a correction factor in (6.3). For this reason, we recommend the Cook statistic over the Andrews-Pregibon one.

Of course, two or more dimensional summaries are preferred over either statistic. Oldford [1983] compares a number of two-dimensional statistics, derived from the above statistics, on the Gessell adaptive score data-set found in Mickey, Dunn, and Clark [1967]. There, plots of $C/\cos^2\theta$ versus $|\cot\theta|$ are seen to work well.

7. Summary and Concluding Remarks

The n-dimensional geometry often used to describe linear models (e.g. Seber [1965], Guttman [1983]) can be extended to accommodate modern regression diagnostics. This has been done by concentrating on certain sets of vectors rather than on the entire vector space they generate. The geometry of other diagnostics not considered here, like variance inflation factors and variance decomposition proportions (e.g. Belsley, Kuh, and Welsch [1980]), could be just as easily explored.

As in the usual linear model theory, the n-dimensional geometry here sheds light on important problems in regression analysis. One immediately sees that the singular values of the corresponding matrix (by way of the dimension indices η_i) and its principal directions in $\mathbb{R}^n$ play important roles in *both* collinearity *and* influential data diagnostics. More specifically, the geometry has led to an understanding of Thisted's [1980, 1982] indices *mci* and *pmci* and their inappropriateness as collinearity diagnostics. On the other hand, it has reinforced Belsley, Kuh, and Welsch's [1980] condition indices (η_i^{-1}). Similarly, the Cook [1977] statistic, C, seems preferable to the Andrews and Pregibon [1978] statistic, AP — mainly on the grounds that the former pays attention to the orientation of individual principal directions of the $\mathbf{S}$ matrix to $\hat{e}$ while the latter does not.

Further, having adopted the effective dimension throughout this geometric examination, some commonality in the various diagnostics emerged. The d_1-effective dimension and its components were observed to consistently mimic the behaviour one would intuitively expect of the "effective number" of observations (or parameters) in the situations considered. While other measures have been suggested to coincide with this intuitive behaviour, none have been applicable in more than one setting (e.g., Huber [1981, p. 160], Mosteller and Tukey [1977, p. 348], Thisted [1982]), nor have they had the advantage of a clear geometric and mathematical interpretation. This common and intuitive theme has much pedagogic value and invites its application in areas other than least-squares linear regression.

Acknowledgements

I would like to thank D.F. Andrews, D.A. Belsley, R.A. Thisted, and an anonymous referee for their helpful comments on this work.

Bibliography

Andrews, D.F. [1971], "Significance Tests Based on Residuals", *Biometrika*, 58, pp. 139-148.

Andrews, D.F. and D. Pregibon [1978], "Finding the Outliers that Matter", *J.R.S.S.* B, 40, pp. 85-93.

Belsley, D.A. [1984], "Demeaning Conditioning Diagnostics Through Centering (with discussion)", *The American Statistician*, 38, pp. 73-94.

Belsley, D.A., E. Kuh, and R.E. Welsch [1980], *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* John Wiley and Sons, Inc., New York.

Belsley, D.A. and R.W. Oldford [1986], "The general problem of ill conditioning and its role in statistical analysis", *Computational Statistics and Data Analysis*, 4, pp. 103-120.

Chambers, John M. [1977], *Computational Methods for Data Analysis*, John Wiley and Sons, Inc., New York.

Cook, R.D. [1977], "Detection of Influential Observations in Linear regression", *Technometrics*, 19, pp. 15-18.

Dempster, A.P. and M. Gasko-Green [1981], "New Tools for Residual Analysis", *Annals of Statistics*, 9, pp. 945-959.

Draper, N.R. and J.A. John [1981], "Influential Observations and Outliers in Regression", *Technometrics*, 23, pp. 21-26.

Gentleman, J.F. [1980], "Finding the K Most Likely Outliers in Two-way Tables", *Technometrics*, 22, pp. 591-600.

Gunst, R. [1984], "Comment: Toward a Balanced Assessment of Collinearity Diagnostics", *The American Statistician*, 38, pp. 79-82.

Guttman, I. [1983], *Linear Models: An Introduction*, John Wiley & Sons, Inc., New York.

Hoaglin, D.C. and R.E. Welsch [1978], "The Hat Matrix in Regression and ANOVA", *American Statistician*, 32, pp. 17-22.

Huber, P.J. [1981], *Robust Statistics*, John Wiley & Sons, Inc., New York.

Mickey, M.R., O.J. Dunn, and V. Clark [1967], "Note on the Use of Stepwise Regression in Detecting Outliers", *Computers and Biomedical Research*, 1, pp. 105-111.

Mosteller, F. and J.W. Tukey [1977], *Data Analysis and Regression*, Addison-Wesley Publishing Co., Reading, Massachusetts.

Oldford, R.W. [1983], "Collinearity and Influential Observations as Dimensionality Problems", *Proceedings of the American Statistical Association: Statistical Computing Section*, pp. 153-158.

Oldford, R.W. [1985], "New geometric theory for the linear model", Technical Report No. 48, Massachusetts Institute of Technology, Center for Computational Research in Economics and Management Science, Cambridge, MA.

Seber, G.A.F. [1965], *The Linear Hypothesis: A General Theory*, Griffin's Statistical Monographs No. 19. Griffin, London.

Smith, G. and F. Campbell [1980], "A Critique of Some Ridge Regression Methods (Invited Paper)", *JASA*, 75, pp. 74-104.

Thisted, R.A. [1980], Discussion of "A Critique of Some Ridge Regression Methods", by G. Smith and F. Campbell, *JASA*, 75, pp. 81-86.

Thisted, R.A. [1982], "Decision-Theoretic Regression Diagnostics", *Statistical Decision Theory and Related Topics III*, Vol. 2, pp. 363-382, Academic Press, Inc., New York.

Thisted, R.A. and C.N. Morris [1980], "Theoretical Results for Adaptive Ordinary Ridge Regression Estimators", Technical Report No. 94 (revised), University of Chicago, Department of Statistics,

Wilkinson, J.H. [1965], *The Algebraic Eigenvalue Problem*, Oxford University Press.

Received by Editor March, 1986; Revised February, 1987.

Recommended by George P. H. Styan, McGill University, Montreal, Canada.

Refereed by H. Caussinus, Universite Paul Sabatier, Toulouse, France.