

THE DISTRIBUTION OF COOK'S D_I STATISTIC

Donald R. Jensen
Department of Statistics
Virginia Polytechnic Institute
Blacksburg, VA 24061

Donald E. Ramirez
Department of Mathematics
University of Virginia
Charlottesville, VA 22903

Abstract

This paper is concerned with the identification of outliers and influential observations in a linear regression model, $Y = X_0\beta + \varepsilon$, of full rank k . Cook's D_I statistic is the scaled Mahalanobis type squared distance $D_I = (\hat{\beta}_I - \hat{\beta})'(X_0'X_0)(\hat{\beta}_I - \hat{\beta})/(rs_I^2)$ between $\hat{\beta}$ (using all the cases) and $\hat{\beta}_I$ (using all cases except those in the subset I with r cases), where s_I^2 is the unbiased estimator for σ^2 with the cases in I omitted. We show how to compute the exact distribution of D_I and the corresponding p -values.

In the case of deleting a single point, Cook's D_I statistic is shown to be a multiple of an F statistic having $(1, N - 1 - k)$ degrees of freedom. When I contains some $r > 1$ cases, the distribution of D_I is a generalized F distribution.

Key Words and Phrases: Cook's D_I statistic, outliers, influential observations, quadratic forms, generalized F distributions.

1. Introduction

Consider the standard linear model $Y = X_0\beta + \varepsilon$, in which X_0 is an $(N \times k)$ matrix of full rank, Y is a vector consisting of N observable responses, and ε is a random vector for which $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2I$. The vector β consists of k unknown parameters, and ε consists of unobservable Gaussian errors. The problem we address in this note concerns the identification of outliers and influential observations in a linear regression model. For example, if some cases are deleted, then what changes occur in estimates for the parameter vector β ? This is the basic idea in *influence analysis* as introduced by Cook (1977). Cook's D_I statistic is based on a scaled Mahalanobis type squared distance between $\hat{\beta}$ (using all the cases) and $\hat{\beta}_I$ (using all cases except those in the subset I), as given by

$$D_I = (\hat{\beta}_I - \hat{\beta})'(X_0'X_0)(\hat{\beta}_I - \hat{\beta})/(c\hat{\sigma}^2), \quad (1.1)$$

with $\hat{\sigma}^2$ as some unbiased estimate of the variance and c a user defined constant. We use the estimator s_I^2 , the residual mean square with the r cases in I omitted; and we use $c = r$. Equivalently, D_I can be written as $D_I = (\hat{Y}_I - \hat{Y})'(\hat{Y}_I - \hat{Y})/(rs_I^2)$ using $\hat{Y}_I = X_0\hat{\beta}_I$. Thus D_I can be viewed as a multiple of the squared Euclidean distance

between the prediction vector using all cases, and the prediction vector with the cases in I deleted. For later reference, let

$$Q_I(X'_0 X_0) = (\hat{\beta}_I - \hat{\beta})'(X'_0 X_0)(\hat{\beta}_I - \hat{\beta}) \quad (1.2)$$

denote the quadratic form in the numerator of Equation (1.1). To obtain other versions of Cook's D_I we introduce the notation

$$D_I(\hat{\beta}, M, c\hat{\sigma}^2) = (\hat{\beta}_I - \hat{\beta})' M (\hat{\beta}_I - \hat{\beta}) / (c\hat{\sigma}^2) \quad (1.3)$$

To use $D_I(\hat{\beta}, X'_0 X_0, ks^2)$ diagnostically, Cook (1977) and Weisberg (1980, p. 108) suggested using the 50th percentile $F(0.50, k, N - k)$ as a benchmark for identifying influential subsets. Since $D_I(\hat{\beta}, X'_0 X_0, ks^2)$ is not distributed as $F(k, N - k)$, they recommended the 50th percentile as a rule-of-thumb for determining influential observations. As noted by Belsley, Kuh, and Welsch (1980, p.27), such empirical procedures should be guided by statistical theory. In this note, we present the required statistical theory for using D_I as a test for outliers by deriving the *cdf*'s of $Q_I(M)$ and $D_I(\hat{\beta}, M, rs^2)$ for $M = X'_0 X_0$ and $M = X'X$. We are able to compute numerically the *cdf* of Cook's D_I statistics, and, in particular, to compute their *p*-values. This approach supports a statistical procedure for identifying joint outliers. In the case of single deletion with $I = \{i\}$ ($1 \leq i \leq N$), we show that D_I (now denoted by D_i) is distributed as a multiple of an $F(1, N - 1 - k)$ random variable, such that

$$\mathcal{L}(D_i/x'_i(X'_0(i)X_0(i))^{-1}x_i) = F(1, N - 1 - k), \quad (1.4)$$

where x'_i is the i -th row of X_0 and $X_0(i)$ results on deleting the i -th row from X_0 . When I contains more than one index, say, $I = \{i_1, \dots, i_r\}$, then the distribution of Cook's D_I statistic is a generalized F distribution.

2. Basic Results

For basic notation, we denote the Euclidean space of dimension m by R^m and its positive cone by R^m_+ ; and for a random variable Y , we denote the law of the distribution of Y by $\mathcal{L}(Y)$. Probability density and cumulative distribution functions are abbreviated as *pdf* and *cdf*, respectively. $N_m(\mu, \Sigma)$ designates the Gaussian law on R^m having mean μ and dispersion matrix $V(Y) = \Sigma$. Standard distributions on R^1_+ include the chi-squared distribution, $\chi^2(\nu)$, having ν degrees of freedom; the t -distribution, $t(\nu)$, having ν degrees of freedom; and the F -distribution, $F(\nu_1, \nu_2)$, having degrees of freedom (ν_1, ν_2) . For Y_1 and Y_2 two vectors of random variables, we denote the covariance matrix for Y_1 and Y_2 by $cov(Y_1, Y_2)$.

To continue, let I be a subset of $\{1, \dots, N\}$, say $I = \{i_1, \dots, i_r\}$. Let X_0 be partitioned as $X'_0 = [X', Z']$, with Z containing the rows determined by I , and X the remaining rows. We assume that the matrices X_0 , X , and Z are all of full rank, of orders $(N \times k)$, $(n \times k)$, and $(r \times k)$, respectively, such that $k < n$ and $n + r = N$, with $r \leq k$ for notational convenience. Partition $Y' = [Y'_1, Y'_2]$, and set $A = X'X$ and

$B = Z'Z$. Using the positive definite square root, we next diagonalize $A^{-1/2}BA^{-1/2}$ by

$$(X'X)^{-1/2}(Z'Z)(X'X)^{-1/2} = A^{-1/2}BA^{-1/2} = P\Gamma P', \quad (2.1)$$

with P as the matrix of eigenvectors and with Γ as the diagonal matrix of corresponding eigenvalues $\{\gamma_1 \geq \dots \geq \gamma_k \geq 0\}$. The matrix Γ has at most r ($\leq k$) non-zero elements, namely, the eigenvalues of $Z(X'X)^{-1}Z'$. (For the case $r > k$, the rank of $Z(X'X)^{-1}Z'$ is $\min(r, k)$.) In particular, if $I = \{i\}$, then $A^{-1/2}BA^{-1/2}$ has only one positive eigenvalue, namely, $\gamma_1 = x'_i(X'_0(i)X_0(i))^{-1}x_i$.

In Jensen and Ramirez (1993), we used the structure of Equation (2.1) to examine efficiencies pertaining to augmenting or deleting subsets of design points. In a similar manner, we may convert the linear model into an equivalent canonical linear model through a one-to-one reparametrization as in

$$\hat{Y} = \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \end{pmatrix} = \begin{pmatrix} X \\ Z \end{pmatrix} \hat{\beta} = \left[\begin{pmatrix} X \\ Z \end{pmatrix} A^{-1/2} P \right] \left[P' A^{1/2} \hat{\beta} \right] = \begin{pmatrix} X A^{-1/2} P \\ Z A^{-1/2} P \end{pmatrix} \hat{\phi}, \quad (2.2)$$

where $\hat{\phi} = P' A^{1/2} \hat{\beta}$. Observe further that

$$\begin{aligned} \hat{\phi}_I &= [(X A^{-1/2} P)' (X A^{-1/2} P)]^{-1} (X A^{-1/2} P)' Y_1 \\ &= [P' A^{-1/2} X' X A^{-1/2} P]^{-1} P' A^{-1/2} X' Y_1 = P' A^{1/2} \hat{\beta}_I. \end{aligned} \quad (2.3)$$

We convert the quadratic form $Q_I(X'_0 X_0)$ into canonical form through several steps as in

$$\begin{aligned} Q_I(X'_0 X_0) &= (\hat{\beta}_I - \hat{\beta})' (X'_0 X_0) (\hat{\beta}_I - \hat{\beta}) \\ &= (\hat{\beta}_I - \hat{\beta})' (A^{1/2} P) P' (I + A^{-1/2} B A^{-1/2}) P (P' A^{1/2}) (\hat{\beta}_I - \hat{\beta}) \\ &= (\hat{\phi}_I - \hat{\phi})' (I + \Gamma) (\hat{\phi}_I - \hat{\phi}), \end{aligned} \quad (2.4)$$

with $I + \Gamma$ a diagonal matrix.

By construction, $\hat{\phi}_I$ and $\hat{\phi}$ are jointly Gaussian variables, since $\hat{\beta}_I$ and $\hat{\beta}$ have multivariate normal distributions. We next compute the expectations and dispersion matrices for $\hat{\phi}_I$, $\hat{\phi}$, and $\hat{\phi}_I - \hat{\phi}$. First, we note that

$$E(\hat{\phi}_I) = E(P' A^{1/2} \hat{\beta}_I) = P' A^{1/2} \beta, \quad (2.5)$$

and

$$E(\hat{\phi}) = E(P' A^{1/2} \hat{\beta}) = P' A^{1/2} \beta, \quad (2.6)$$

so that

$$E(\hat{\phi}_I - \hat{\phi}) = 0. \quad (2.7)$$

The dispersion matrices for $\hat{\phi}_I$ and $\hat{\phi}$ are given by

$$V(\hat{\phi}_I) = P' A^{1/2} V(\hat{\beta}_I) A^{1/2} P = \sigma^2 P' A^{1/2} (A)^{-1} A^{1/2} P = \sigma^2 I, \quad (2.8)$$

and

$$\begin{aligned} V(\hat{\phi}) &= P' A^{1/2} V(\hat{\beta}) A^{1/2} P = \sigma^2 P' A^{1/2} (A + B)^{-1} A^{1/2} P \\ &= \sigma^2 P' (A^{-1/2} (A + B) A^{-1/2})^{-1} P = \sigma^2 (I + \Gamma)^{-1}. \end{aligned} \quad (2.9)$$

The covariance between $\hat{\phi}_I$ and $\hat{\phi}$ is given by

$$\begin{aligned}
\text{cov}(\hat{\phi}_I, \hat{\phi}) &= \text{cov}(P' A^{1/2} \hat{\beta}_I, P' A^{1/2} \hat{\beta}) & (2.10) \\
&= P' A^{1/2} \text{cov}(A^{-1} X' Y_1, (A+B)^{-1} (X' Y_1 + Z' Y_2)) A^{1/2} P \\
&= P' A^{1/2} A^{-1} X' \text{cov}(Y_1, Y_1) X (A+B)^{-1} A^{1/2} P + 0 \\
&= \sigma^2 P' A^{1/2} (A+B)^{-1} A^{1/2} P = \sigma^2 (I + \Gamma)^{-1}.
\end{aligned}$$

The corresponding result in terms of $\hat{\beta}_I$ and $\hat{\beta}$ has

$$\text{cov}(\hat{\beta}_I, \hat{\beta}) = \sigma^2 (X_0' X_0)^{-1}. \quad (2.11)$$

To find the dispersion matrix for $\hat{\phi}_I - \hat{\phi}$, use Equations (2.8), (2.9), and (2.10) to get

$$V(\hat{\phi}_I - \hat{\phi}) = \sigma^2 (I - (I + \Gamma)^{-1}). \quad (2.12)$$

The corresponding result in terms of $\hat{\beta}_I$ and $\hat{\beta}$ is

$$V(\hat{\beta}_I - \hat{\beta}) = \sigma^2 ((X' X)^{-1} - (X_0' X_0)^{-1}). \quad (2.13)$$

We now give the structure theorem for $Q_I(X_0' X_0)$, the numerator of Cook's D_I statistic.

Theorem 1. *With $\mathcal{L}(Y) = N_N(X_0 \beta, \sigma^2 I_N)$ and the notation above, the quadratic form $Q_I(X_0' X_0)$ is distributed as a weighted sum of independent chi-squared random variables, namely,*

$$\mathcal{L}(Q_I(X_0' X_0)) = \mathcal{L}(\gamma_1 \sigma^2 Z_1^2 + \cdots + \gamma_r \sigma^2 Z_r^2), \quad (2.14)$$

where $\mathcal{L}(Z) = N(0, 1)$ and the weights $\{\gamma_1 \geq \cdots \geq \gamma_r > 0\}$ are the non-zero eigenvalues of $Z(X' X)^{-1} Z'$, or equivalently, the non-zero eigenvalues of Γ .

Proof. Standard results for quadratic forms (for example, Mathai and Provost (1992, p. 90)) show that $Q_I(X_0' X_0)$ may be represented in distribution as a weighted sum of independent chi-squared random variables with weights as the non-zero eigenvalues $\{\gamma_1, \dots, \gamma_r\}$ of $(I - (I + \Gamma)^{-1})^{1/2} (I + \Gamma) (I - (I + \Gamma)^{-1})^{1/2} = \Gamma$. ■

We next consider the distribution of Cook's D_I statistic. Its numerator has been treated in Theorem 1. The denominator is rs_I^2 , where we have chosen s_I^2 because it is independent of $\hat{\beta}_I - \hat{\beta}$. This follows from

$$\begin{aligned}
\hat{\beta}_I - \hat{\beta} &= (X' X)^{-1} X' Y_1 - (X_0' X_0)^{-1} (X' Y_1 + Z' Y_2) & (2.15) \\
&= ((X' X)^{-1} - (X_0' X_0)^{-1}) X' Y_1 - (X_0' X_0)^{-1} Z' Y_2.
\end{aligned}$$

The estimator s_I^2 is clearly independent of the second term on the right of Equation (2.15). It is independent of the first term as well since $\text{cov}(X' Y_1, Y_1 - \hat{Y}_1) = 0$, as X' and $I - X(X' X)^{-1} X'$ are perpendicular. Since $\mathcal{L}((n-k)s_I^2/\sigma^2) = \chi^2(n-k)$, we have the following theorem.

Theorem 2. With $\mathcal{L}(Y) = N_N(X_0\beta, \sigma^2 I_N)$ and the notation above,

$$\mathcal{L}\left(\frac{r D_I}{n-k}\right) = \mathcal{L}\left(\frac{Q_I(X'_0 X_0)}{V_I}\right) = \mathcal{L}\left(\frac{\gamma_1 \sigma^2 Z_1^2 + \cdots + \gamma_r \sigma^2 Z_r^2}{V_I}\right), \quad (2.16)$$

with $Q_I(X'_0 X_0) = (\hat{\beta}_I - \hat{\beta})'(X'_0 X_0)(\hat{\beta}_I - \hat{\beta})$ and $V_I = (n-k)s_I^2$, and with $Q_I(X'_0 X_0)$ and V_I as independent random variables. When $I = \{i\}$,

$$\mathcal{L}(r D_i / x'_i (X'_0(i) X_0(i))^{-1} x_i) = F(1, N-1-k). \quad (2.17)$$

We now define the *generalized F distribution* based on Equation (2.16). Suppose that the elements of $\mathbf{U} = [U_1, \dots, U_r]'$ are independent $\{N_1(0, 1); 1 \leq i \leq r\}$ random variables; let $\{\alpha_1, \dots, \alpha_r\}$ be non-increasing positive weights; and identify

$$Q = \alpha_1 U_1^2 + \cdots + \alpha_r U_r^2. \quad (2.18)$$

If $\mathcal{L}(V) = \chi^2(v)$ independently of \mathbf{U} , then the *cdf* of

$$W = \frac{Q/r}{V/v} \quad (2.19)$$

is denoted by $F_r(w; \alpha_1, \dots, \alpha_r; v)$. If $\{\mathcal{L}(U_i) = N_1(\omega_i, 1); 1 \leq i \leq r\}$, then the *cdf* of W is denoted by $F_r(w; \alpha_1, \dots, \alpha_r; \omega_1, \dots, \omega_r; v)$. Thus we have shown

Theorem 3. With $\mathcal{L}(Y) = N_N(X_0\beta, \sigma^2 I_N)$ and the notation above, the distribution of $D_I(\hat{\beta}, X'_0 X_0, r s_I^2)$ is given by

$$\mathcal{L}(D_I(\hat{\beta}, X'_0 X_0, r s_I^2)) = F_r(w; \gamma_1, \dots, \gamma_r; n-k). \quad (2.20)$$

Another natural candidate for M is $X'X$. We now derive the distribution of $D_I(\hat{\beta}, X'X, r s_I^2)$. First let the ordered eigenvalues of $Z(X'_0 X_0)^{-1} Z'$ be denoted by $\{\lambda_1 \geq \cdots \geq \lambda_r > 0\}$ with $\{\lambda_i = \gamma_i / (1 + \gamma_i); 1 \leq i \leq r\}$ the canonical leverages (also denoted $h_{ii}, 1 \leq i \leq r$). As in Equation (2.4), we can convert the quadratic form $Q_I(X'X)$ into canonical form by

$$\begin{aligned} Q_I(X'X) &= (\hat{\beta}_I - \hat{\beta})'(X'X)(\hat{\beta}_I - \hat{\beta}) \\ &= (\hat{\beta}_I - \hat{\beta})'(A^{1/2} P)P'(I)P(P' A^{1/2})(\hat{\beta}_I - \hat{\beta}) \\ &= (\hat{\phi}_I - \hat{\phi})'(I)(\hat{\phi}_I - \hat{\phi}), \end{aligned} \quad (2.21)$$

Standard results for quadratic forms (for example, Mathai and Provost (1992, p. 90)) show that $Q_I(X'X)$ may be represented in distribution as a weighted sum of independent chi-squared random variables with weights as the non-zero eigenvalues $\{\lambda_1, \dots, \lambda_r\}$ of $(I - (I + \Gamma)^{-1})^{1/2} (I) (I - (I + \Gamma)^{-1})^{1/2} = I - (I + \Gamma)^{-1}$. Thus it follows that

Theorem 4. With $\mathcal{L}(Y) = N_N(X_0\beta, \sigma^2 I_N)$ and the notation above, the distributions of $Q_I(X'X)$ and $D_I(\hat{\beta}, X'X, r s_I^2)$ are given by

$$\mathcal{L}(Q_I(X'X)) = \mathcal{L}(\lambda_1 \sigma^2 Z_1^2 + \cdots + \lambda_r \sigma^2 Z_r^2), \quad (2.22)$$

and

$$\mathcal{L}(D_I(\hat{\beta}, X'X, r s_I^2)) = F_r(w; \lambda_1, \dots, \lambda_r; n-k). \quad (2.23)$$

Further extension of these results are given in Jensen and Ramirez (1996) where we adapted the theory of singular decompositions to transform a linear model into canonical form.

References

- [1] Belsley, D., Kuh, E., and Welsch, R. (1980). *Regression Diagnostics*. John Wiley & Sons, Inc., New York.
- [2] Cook, R. (1977). Detection of influential observations in linear regression, *Technometrics*, 19, 15-18.
- [3] Jensen, D. and Ramirez, D. (1993). Efficiency comparisons in linear inference, *J. Statistical Planning and Inference*, 37, 51-68.
- [4] Jensen, D. and Ramirez, D. (1996). Some exact properties of Cook's D_I , under review.
- [5] Mathai, A. and Provost, S. (1992). *Quadratic Forms in Random Variables: Theory and Applications*. Marcel Dekker, Inc., New York.
- [6] Weisberg, S. (1980). *Applied Linear Regression*. John Wiley & Sons, Inc., New York.