

Computing the *cdf* of Cook's D_I Statistic

D. R. Jensen¹ and D. E. Ramirez²

¹ Virginia Polytechnic Institute
Department of Statistics, Blacksburg, VA 24061 USA

² University of Virginia
Department of Mathematics, Charlottesville, VA 22903 USA

1 Summary

Consider a linear model $Y = X_0\beta + \varepsilon$, with $(Y, \varepsilon) \in R^N$, X_0 of order $(N \times k)$, and with $\beta \in R^k$ unknown. Partition $Y = [Y_1', Y_2']'$ and $X_0 = [X', Z']'$ conformably, with $Y_1 \in R^n$, $Y_2 \in R^r$, X of order $(n \times k)$ of rank $k < n$, and Z of order $(r \times k)$, where $N = n + r$ and $r \leq k$. Let $\hat{\beta}$ be the least-squares estimator for β from the full data (X_0, Y) , and $\hat{\beta}_I$ from the reduced data (X, Y_1) . Cook's D_I statistics take the form $D_I(\hat{\beta}, M, rs_I^2) = (\hat{\beta}_I - \hat{\beta})' M (\hat{\beta}_I - \hat{\beta}) / (rs_I^2)$ where $M(k \times k)$ is non-negative definite and s_I^2 is the residual mean square from the reduced data. Commonly used choices for M are $X_0'X_0$ and $X'X$. Suppose that 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 $T = \alpha_1 U_1^2 + \dots + \alpha_r U_r^2$. If $\mathcal{L}(V) = \chi^2(v)$ independently of \mathbf{U} , then the *cdf* of $W = (T/r)/(V/v)$ is denoted by $F_r(t; \alpha_1, \dots, \alpha_r; v)$. The *pdf* of $T/V = rW/v$ has the representation $\sum_{i=0}^{\infty} (c_i/\delta) ((n-k)/(r+2i)) f_1(((n-k)/(r+2i))(t/\delta); r+2i, n-k)$ with f_1 as the *pdf* of $F(v_1, v_2)$, and whose coefficients $\{c_i\}$ are defined recursively such that $0 < \delta < \alpha_r$. We have developed an algorithm for computing the distribution and *p*-values of T/V . See Jensen and Ramirez (1996a).

If $\mathcal{L}(Y) = N_N(X_0\beta, \sigma^2 I_N)$, then (1) the *cdf* of $D_I(\hat{\beta}, X_0'X_0, rs_I^2)$ is $F_r(t; \gamma_1^2, \dots, \gamma_r^2; v)$ where $\{\gamma_1^2 \geq \dots \geq \gamma_r^2\}$ are the ordered eigenvalues of $Z(X'X)^{-1}Z'$; (2) the *cdf* of $D_I(\hat{\beta}, X'X, rs_I^2)$ is $F_r(t; \lambda_1, \dots, \lambda_r; v)$ where $\{\lambda_1 \geq \dots \geq \lambda_r\}$ are the ordered eigenvalues of $Z(X_0'X_0)^{-1}Z'$ and $\lambda_i = \gamma_i^2/(1 + \gamma_i^2) = h_{ii}, 1 \leq i \leq r$; and (3) the *cdf* of $D_I(\hat{\beta}, \Sigma^-, rs_I^2)$ (a modified D_I statistic) is $F_r(t; 1, \dots, 1; v)$ where Σ^- is the Moore-Penrose generalized inverse of $\text{cov}(\hat{\beta}_I - \hat{\beta})$ (Jensen and Ramirez (1996b)). We prefer (2) to (1) computationally, since the number of terms required in the series expansion for the *cdf* of T/V is smaller. This number increases with the condition numbers, which satisfy $\gamma_1^2/\gamma_r^2 \geq \lambda_1/\lambda_r$.

When $r = 1$, the three Cook's D_I statistics all have scaled $F(1, N-1-k)$ distributions, and thus all have the same p -values. These, in turn, are equal to the p -value from the Studentized deleted residuals test $(y_i - \hat{y}_{(i)}) / (s_i \sqrt{1 + x_i(X'X)^{-1}x_i'})$, or equivalently, from the R-Student test $(y_i - \hat{y}_i) / (s_i \sqrt{1 - h_{ii}})$. That the three p -values for the Cook's D_I statistics are equal follows from the characterizations of their distributions as given in Theorems 3, 4, and 5 of Jensen and Ramirez (1996b).

When $r > 1$, the cdf 's for the two Cook's D_I statistics satisfy stochastic bounds based on the largest eigenvalue (for the lower bound) and on their geometric mean (for the upper bound). Generally, the bounds are tighter using $D_I(\hat{\beta}, X'X, rs_I^2)$. To estimate the p -values we have found that $F_r(t; \bar{\alpha}, \dots, \bar{\alpha}; v)$ is a good approximation for $F_r(t; \alpha_1, \dots, \alpha_r; v)$. The p -values using the modified D_I statistic (3) are the same as those from the F -test for testing the mean shift model with additional parameters added for each row in I .

We use the Drill Data Set from Cook and Weisberg (1982, p. 149) with $N = 31$ and $k = 10$. With $r = 1$, we find the influential rows (with $p < .025$) to be row 9 ($p = .003$), row 28 ($p = .011$), and row 31 ($p = .019$). The p -values are (necessarily) the same using $D_I(\hat{\beta}, X'X, rs_I^2)$, $D_I(\hat{\beta}, X_0'X_0, rs_I^2)$, $D_I(\hat{\beta}, \Sigma^-, rs_I^2)$, or RSTUDENT. With $r = 2$, we find 35 influential pairs (with $p < .01$) using $\bar{\lambda}$ from $Z(X_0'X_0)^{-1}Z'$, 39 pairs using $\bar{\gamma}^2 = \text{avg}(\gamma_i^2)$ from $Z(X'X)^{-1}Z'$, and 16 pairs using $D_I(\hat{\beta}, \Sigma^-, rs_I^2)$. All contain at least one row from $\{9, 28, 31\}$ except for the pair (5, 26). The p -values for this pair, using $D_I(\hat{\beta}, X'X, rs_I^2)$, $D_I(\hat{\beta}, X_0'X_0, rs_I^2)$, and $D_I(\hat{\beta}, \Sigma^-, rs_I^2)$, are .000115, .000129, and .000161, respectively. The number of terms used in the partial sums for the cdf 's of $D_I(\hat{\beta}, X'X, rs_I^2)$ and $D_I(\hat{\beta}, X_0'X_0, rs_I^2)$ are 20 and 50, respectively. The mean ranges (upper bound - lower bound) are .075 and .103, respectively. The mean condition numbers λ_1/λ_2 and γ_1^2/γ_2^2 are 2.733 and 4.336, respectively.

References

- Cook, R. D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall.
- Jensen, D. R. & Ramirez, D. E. (1996a). The distribution of Cook's D_I statistic. In review.
- Jensen, D. R. & Ramirez, D. E. (1996b). Some exact properties of Cook's D_I . In review.