한국통계학회논문집 제3권 2호
1996년 8월 pp. 13-19

Cutoff Values for Cook's Distance[1)]

Choongrak Kim[2)]

Abstract

Cook's distance(Cook, 1977) is one of the most widely used influence measures to assess the influence of single observations or sets of observations in the linear regression model. After computing Cook's distance for a set of observations, one needs cutoff values. Cook (1977) suggested guidelines based on a confidence ellipsoid for the regression parameter β. In this paper, we suggest cutoff values for Cook's distance via Monte Carlo simulation, and compare them with Cook's guidelines. An example based on a real data set is given.

1. Introduction

To assess the influence of observation on the fitted values in linear regression, there are many influence measures (see Chatterjee and Hadi (1986)). Among them, Cook's distance (Cook, 1977) is one of the most widely used influence measures because it is easy to interpret and has a relatively simple form. Also, it is available in most statistical packages such as MINITAB, SAS, and BMDP. Consider a linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{y}$ is an n-vector of response, $\mathbf{X}$ is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is p-vector of unknown coefficients, and $\boldsymbol{\varepsilon}$ is an n-vector of error terms with mean 0 and variance $\sigma^2 I$. Cook's distance for the ith observation is given by

$$C_i = \frac{1}{ps^2} \cdot \frac{e_i^2 h_{ii}}{(1-h_{ii})^2}, \tag{1.2}$$

where s^2 is unbiased estimator of σ^2, $e_i = y_i - \hat{y}_i$ is residual, and h_{ii} is the ith diagonal element of the hat matrix H. We say the ith observation is influential if C_i is large. Almost

1) This paper was supported (in part) by NON DIRECTED RESEARCH FUND, Korea Research Foundation.

2) Associate Professor, Department of Statistics, Pusan National University, Pusan, 609-735, Korea.

two decades have passed since Cook's distance was first proposed, however, there is no widely accepted cutoff value for the Cook's distance. Cook(1977) proposed $F_{.50}(p, n-p)$, the 50 *th* percentile of the F distribution with degree of freedom p and $n-p$ as a cutoff value. His idea stemed from the fact that $(1-\alpha)\times 100\%$ confidence ellipsoid fot β based on $\hat{\beta}$ is given by the set of all β^* such that

$$(\beta^*-\beta)^t X^t X(\beta^*-\beta)/ps^2 \le F_{1-\alpha}(p, n-p) \tag{1.3}$$

and C_i has a similar from to (1.3). His idea is quite reasonable and intuitive, however, our simulation results show that $F_{.50}(p, n-p)$ does not reflect the actual behavior of Cook's distance. In the paper, we suggest a guideline for the cutoff values of Cook's distance based on both the hypothetical design and the Monte Carlo studies, and compare our results with the Cook's suggestion. In Section 2, a cutoff value based on hypothetical design is verified, and in Section 3, results based on the Monte Carlo studies are summarized. Numerical example is given in Section 4, and remarks are given in Section 5.

2. Equal Leverage Case

Since the residuals are correlated with each other, the Cook's distance are not independent random from a certain distribution. If we assume $\varepsilon_i \sim N(0, \sigma^2)$, then we can obtain a joint distribution for $C_1, \cdots, C_n$; however, this does not lead to a tractable solution.

Assume that $\varepsilon_i \sim N(0, \sigma^2)$ and $h_{ii} = p/n$ for all i, average value of leverages. In fact, the assumption $h_{ii} = p/n$ for all i is not realistic except some special cases of 2^m factorial design. But, this assumption can lead to a feasible solution because Cook's distance under this assumption becomes

$$C_i = \frac{r_i^2}{n-p},$$

where $r_i = e_i/s\sqrt{1-h_{ii}}$, and $C_i \sim Beta(1/2, (n-p-1)/2)$ (see Cook and Weisberg (1982, p.19) for details). Therefore, as a cutoff value under this assumption, $Beta_{.95}(1/2, (n-p-1)/2)$, 95*th* percentile of the Beta distribution with degrees of freedom 1/2 and $(n-p-1)/2$, would be useful. We evaluate $Beta_{.95}(1/2, (n-p-1)/2)$ for n = 20(10)50 and p = 2(1)7 and list them in Table 2. As shown in this Table, $Beta_{.95}(1/2, (n-p-1)/2)$ is very close to the 95*th* percentile of our simulation results, though it slightly overestimate. The analytic

expression, $Beta_{.95}(1/2, (n-p-1)/2)$, however, is based on the assumption of equal leverage. It is not guaranteed to be applied to other cases.

3. Simulation Results

Intuitively, the magnitude of C_i is highly dependent on n, p and s^2. In our simulation, we examined n = 20(10)50, p = 2(1)7, and assumed $\sigma^2 = 1$. Under consideration is the multiple linear regression model

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon ,$$

where $X_i \sim U(1,10)$ and $\varepsilon \sim N(0,1)$. For each n and p, we performed 1000 replications. In each replication, the independent variable X was fixed, all $\beta_i's$ were equal to 1, and y was obtained by generating ε using the IMSL subroutines GGUBS and GGNML. The computation of $\hat{\beta}_{(i)}$ was done by LINPACK subroutines DQRDC and DQRSL. We obtained $C_{(1)}, C_{(2)}, \cdots, C_{(n)}$ the order statistics of $C_1, C_2, \cdots, C_n$ from each replication by using the IMSL subroutine VSRTA, and then obtained $\overline{C}_{(1)}, \overline{C}_{(2)}, \cdots, \overline{C}_{(n)}$ means of $C_{(1)}, C_{(2)}, \cdots, C_{(n)}$. To see whether particular choices of distributions for the independent variables affected the magnitude of the $C_i's$, we performed another simulation under the same conditions for $n = 20$ and $p = 5$, using a different of $\mathbf{X}$. As we see in Table 1 (listed are values of $\overline{C}_{(1)}, \overline{C}_{(2)}, \cdots, \overline{C}_{(n)}$), the magnitude of $C_i's$ does not appear to depend strongly on $\mathbf{X}$.

Table 1. Simulation results for the ordered Cook's distance when $n = 20, p = 5$

(a) $X_1, X_2, X_3, X_4 \sim U(1,10)$

.0003	.0011	.0025	.0047	.0073	.0102	.0138	.0179	.0227	.0278
.0342	.0438	.0540	.0671	.0838	.1029	.1263	.1674	.2473	.4068

(b) $X_1 \sim U(1,10), X_2 \sim N(10,4), X_3 \sim B(1,5), X_4 \sim P(1)$

.0004	.0011	.0024	.0043	.0064	.0094	.0126	.0164	.0216	.0218
.0346	.0435	.0554	.0659	.0852	.1042	.1295	.1774	.2411	.4474

Plot of empirical density estimation using S graphics for each n and p are given in Figure 1. As shown in these plots, we find the following facts :

1) All the densities are skewed to the right;

2) For $20 \leq n \leq 50, 2 \leq p \leq 7$, most of the $C_i's$ are less than 0.5;

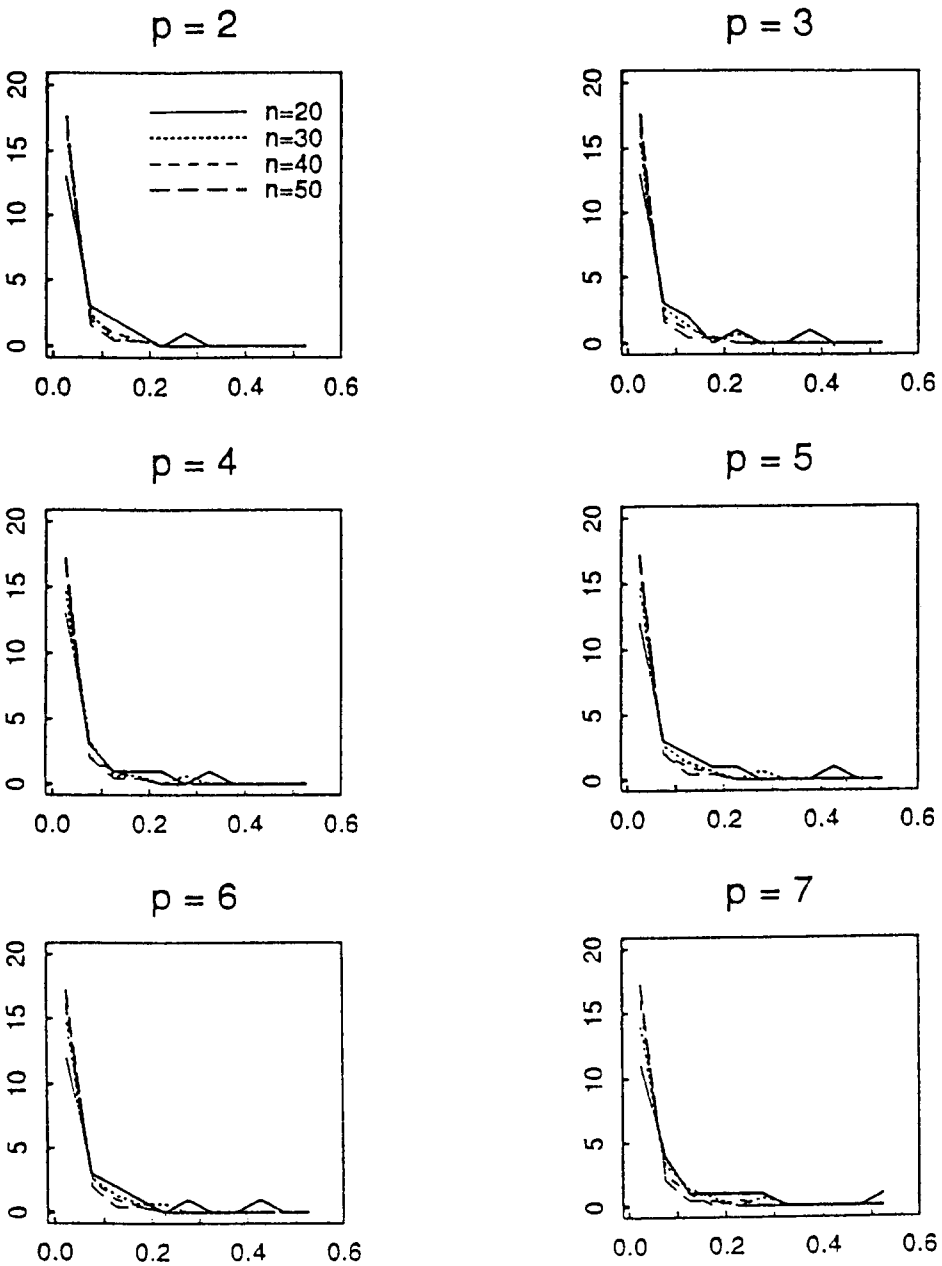

Figure 1 : Empirical density of Cook's distance

To see the properties of our empirical results in detail, we list summary statistics for $\overline{C}_{(1)}$, $\overline{C}_{(2)}, \cdots, \overline{C}_{(n)}$ in Table 2 from which we find :

1) For fixed p, mean values of $C_i's$ are decreasing as n increases;
2) For fixed n, mean values of $C_i's$ are increasing as p increases;
3) Mean values of the $C_i's$ are well approximated by $1/(n-p)$. This seems quite reasonables, since if we let $h_{ii}=p/n$ and $e_i^2=(n-p)s^2/n$, then C_i becomes $1/(n-p)$;
4) The average $95th$ percentile, which we suggest can be used as a cutoff value for C_i is approximated by a function of n and p : $f(n,p)=3.67/(n-p)$ which is proportional to the mean value of the $C_i's$.
5) $F_{.50}(p, n-p)$, suggested by Cook (1977) as a cutoff value for C_i, is too large and rarely reveal the effect of n. Also, the Kolmogorov - Smirnov test for goodness-of-fit based on the empirical distribution function shows that the F distribution is far away from that of the $C_i's$.

Table 2. Summary Statistics for Cook's Distance C_i

n	p	mean	$1/(n-p)$	$95th\%$	$3.67/(n-p)$	$Beta_{.95}$	$F_{.50}$
20	2	.056	.056	.188	.204	.208	.721
	3	.062	.059	.217	.216	.219	.821
	4	.066	.063	.225	.230	.233	.876
	5	.072	.067	.247	.245	.247	.911
	6	.076	.071	.255	.262	.264	.936
	7	.088	.077	.294	.282	.284	.955
30	2	.036	.036	.128	.131	.135	.711
	3	.038	.037	.131	.136	.140	.809
	4	.039	.039	.136	.141	.145	.862
	5	.041	.040	.145	.147	.151	.917
	6	.042	.044	.158	.160	.164	.934
40	2	.0026	.026	.096	.097	.100	.706
	3	.027	.027	.101	.099	.103	.803
	4	.028	.028	.102	.102	.105	.855
	5	.029	.029	.108	.105	.108	.887
	6	.030	.029	.105	.108	.112	.909
	7	.032	.030	.111	.112	.115	.925
50	2	.021	.021	.078	.076	.079	.703
	3	.021	.021	.080	.078	.081	.800
	4	.022	.022	.081	.080	.083	.851
	5	.023	.022	.085	.082	.085	.884
	6	.023	.023	.084	.083	.086	.905
	7	.024	.023	.085	.086	.088	.921

Conclusively, the $95th$ percentile of our simulation results can be approximated by $3.67/(n-p)$, and $Beta_{.95}(1/2,(n-p-1)/2)$ is very close to them.

4. Example

Belsley, Kuh, and Welsch(1980) considered influential observations for the model that how savings ratio is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variable $(n=50, p=5)$. For more details and complete list of data, see Belsley, Kuh, and Welsch(1980,p.39-42). We computed Cook's distance C_i (see Figure 2), and the cutoff value in this case is $3.67/(50-5)=.082$. As shown in this Figure, 3 points (49, 23, 46) have larger cook's distance than .082, and they are clearly away from others. On the other hand, no point has larger Cook's distance than $F_{.50}(5,45)=.884$.

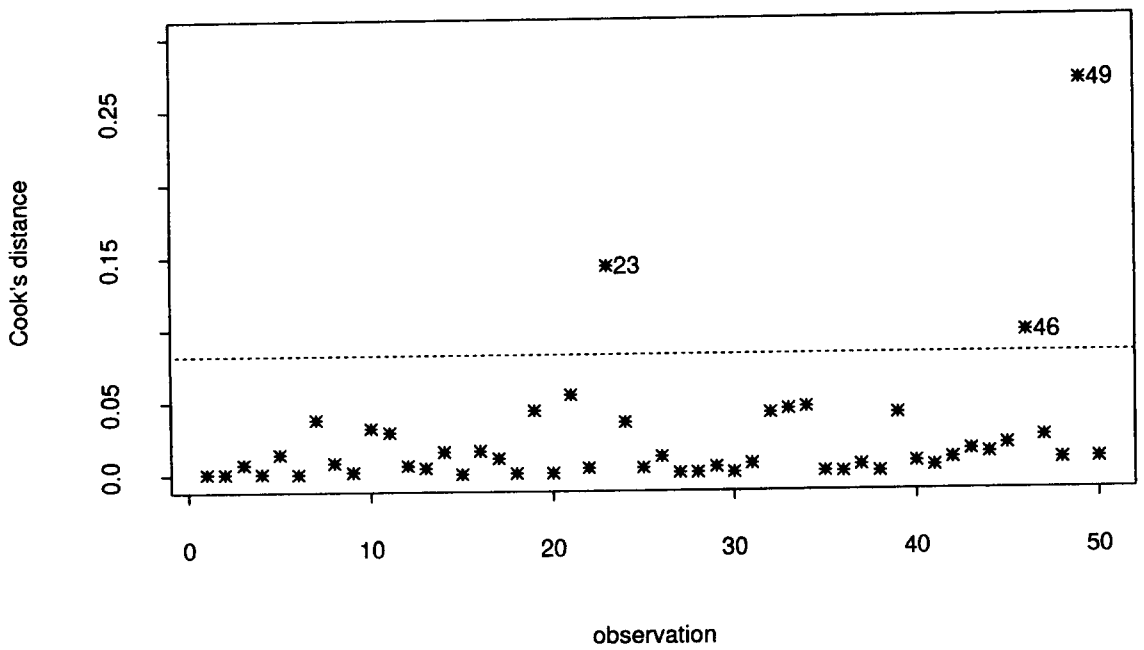

Figure 2 : Cook's distance in savings data

5. Remarks and Future Research

After computing Cook's distance, one needs some guidelines indicating influential observation deserving special attention. To get this cutoff value, we used two approaches ; one is deriving analytic distribution of Cook's distance by assuming equal leverage in hat matrix, and the other is approximating the 95*th* percetile of empirical distribution from the Monte Carlo studies. They turned out to be $Beta_{.95}(1/2,(n-p-1)/2)$ and $3.67/(n-p)$, respectively, and they are quite close for moderate n and p. Our cutoff values do not match well with those suggested by Cook(1977). In fact, $F_{.50}(p, n-p)$ does not reflect the actual behavior of Cook's distance, and is too large. Also, it does not reflect the dependence on n very well. However, $F_{.50}(p, n-p)$ has its own geometric interpretation. For example, if $C_i = F_{.50}(p, n-p)$ then the deletion of the ith case would move $\hat{\beta}_{(i)}$ to the edge of a 50% confidence ellipsoid relative to $\hat{\beta}$.

One thing to note is that these cutoff values are not strict or unique rule distinguishing "influential" from "not-influential" . They should be used as a guideline for identification, and further accomodations are up to the analysts.

In this paper, we suggested cutoff values of Cook's distance for single obervations only. Due to the masking effect or swamping phenomenon, we definitely need sone cutoff values of Cook's distance for sets of observations. Of course, it will be requried a lot of computation times and a careful set-up for random number generating mechanism. We are working on this project and will obtain some results in the near future. Alse, it should be noted that the more variations in $\mathbf{X}$ the higher chance of getting large h_{ii}. Therefore, we have to be cautious in generating $\mathbf{X}$ in Monte Carlo studies.

References

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

[2] Chatterjee, S. and Hadi, A.S. (1986). Influential observations, high leverage points, and outliers in linear regression(with discussion), *Statistical Science,* Vol. 1, 379-416

[3] Cook, R. D. (1977). Detection of influential obervations in linear regression, *Technometrcs,* Vol. 19, 15-18.

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