Journal of the Korean Statistical Society (2000), **29**: *3, pp 285–295*

Effect of Outliers on Sample Correlation Coefficient †

Choongrak Kim[1], Byeong U. Park[2], Kook L. Choi[3] and Whasoo Bae[3]

Abstract

In analyzing bivariate data the sample correlation coefficient is often used, and it is quite sensitive to one or few isolated cases. In this article we derive a formula for the effect of k observations on the sample correlation coefficient by the deletion method. To give a reference value for the isolated cases the asymptotic distribution for the formula is derived. Also, we give some interpretations on several types of isolated cases and an example based on a real data set.

Key Words and Phrases : Cancelling Effect, Correlation Coefficient, High Leverage Point, Masking Effect, Outlier.

1. INTRODUCTION

One of the most frequently used statistic when analyzing bivariate data is the sample correlation coefficient. This statistic is very simple and easy to interpret so that many people including nonstatistician use and cite very often. However, like other statistics, the sample correlation coefficient is quite sensitively influenced by one or few observations. Figure 1 shows the scatter diagram for 1985 and 1986 batting averages for 124 American League Players taken from Wardrop(1995). For these Batting Average data, sample correlation coefficient $r = 0.554$. If we delete case 92 then $r = 0.669$. Wardrop(1995) noted that

[†] This research is supported by a grant from Korea Science and Engineering Foundation, Project Number 981-0105-022-1.

[1] Department of Statistics, Pusan National University, Pusan, Korea, 609-735. Member of the Research Institute of Computer, Information and Communication.

[2] Department of Computer Science and Statistics, Seoul National University, Seoul, Koıea, 151-742.

[3] Department of Statistics, Inje University, Kimhae, Korea, 621-749.

"Dropping just one case from 124 - less than 1% of the data - results in a 10% increase in r ". Therefore, this case seems to be very influential on r. As in the regression diagnostics context, we might be interested in the following issues : 1. What kind of isolated cases make r larger or smaller when they are deleted? 2. Can we have some guideline to treat a case influential? 3. Is it necessary to delete more than one case simultaneously to detect the masking effect or the cancelling effect?

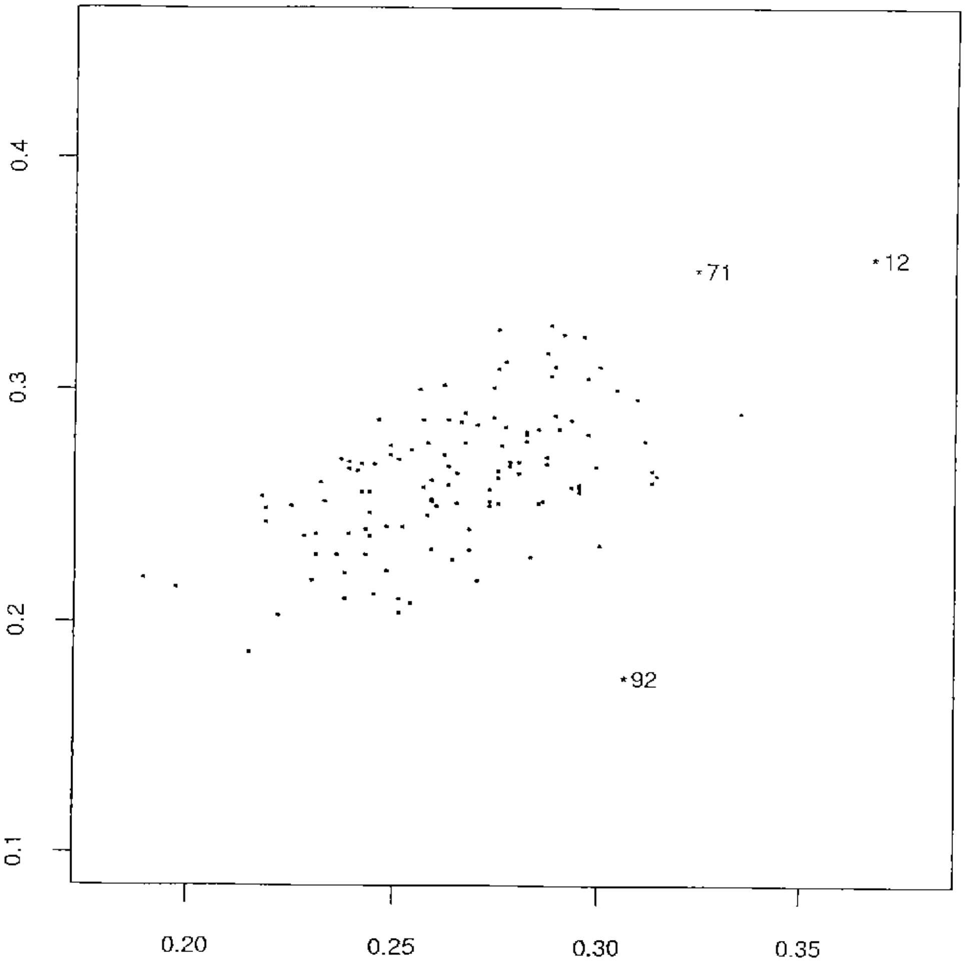


Figure 1. Scatterplot of 1986 versus 1985 American League batting averages

In this article we derive the influence of one or few observations on the sample correlation coefficient. To do this the most intuitive way is computing the difference of two sample correlation coefficients based on the full sample and the reduced sample, respectively. A formula for this difference is obtained and the asymptotic distribution of the difference is derived in Section 2. The asymptotic distribution can serve as a guide to decide whether some observations are influential or not. In Section 3, we give some interpretations on several types of isolated cases and an illustrative example based on the Batting Average data.

2. INFLUENCE ON THE CORRELATION COEFFICIENT

Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be independent and identically distributed n pairs of random variables. Then we often use the sample correlation coefficient

$$r = S_{XY}/\sqrt{S_{XX}S_{YY}}$$

as an estimator of the correlation between X_1 and Y_1, where $S_{XX} = \sum(X_i - \bar{X})^2$, $S_{YY} = \sum(Y_i - \bar{Y})^2$, and $S_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y})$. Let $K = \{i_1, \cdots, i_k\}$ be the index set of a set of k observations. To see the effect of the k observations on r, it is natural to see the difference

$$\Delta_{(K)} = r_{(K)} - r$$

where $r_{(K)}$ denotes the sample correlation coefficient based on $n-k$ observations after deleting the k observations with index in K. After some tedious algebra, it can be shown (see Appendix) that

$$\Delta_{(K)} = (\frac{1}{\sqrt{bc}} - 1)r - \frac{a}{\sqrt{bcS_{XX}S_{YY}}} \tag{1}$$

where

$$a = \frac{1}{n-k}\sum_{j\in K} p_j \sum_{j\in K} q_j + \sum_{j\in K} p_j q_j \tag{2}$$

$$b = 1 - \frac{(\sum_{j\in K} p_j)^2}{(n-k)S_{XX}} - \frac{\sum_{j\in K} p_j^2}{S_{XX}} \tag{3}$$

$$c = 1 - \frac{(\sum_{j\in K} q_j)^2}{(n-k)S_{YY}} - \frac{\sum_{j\in K} q_j^2}{S_{YY}} \tag{4}$$

and

$$p_j = X_j - \bar{X}\ , \qquad q_j = Y_j - \bar{Y}. \tag{5}$$

Now, it would be very useful if we suggest a guideline to flag potentially influential observations on r. To do this assumes that X_1 and Y_1 have finite fourth moments, and let $\mu_X = E(X_1)$, $\mu_Y = E(Y_1)$, $\sigma_X^2 = Var(X_1)$, $\sigma_Y^2 = Var(Y_1)$, $\sigma_{XY} = Cov(X_1, Y_1)$, and $\rho = Corr(X_1, Y_1)$. Also, let $Z_i = (X_i - \mu_X)/\sigma_X$ and $W_i = (Y_i - \mu_Y)/\sigma_Y$.

Theorem 1. Suppose $k \to \infty$ in such a way that $k/n \to \lambda(0 < \lambda < 1)$ as $n \to \infty$. Then $n^{1/2}(r_{(K)} - r)$ converges in distribution to $N(0, \tau^2)$ where $\tau^2 = \{\lambda/(1-\lambda)\}\ Var(Z_1W_1 - \rho Z_1^2/2 - \rho W_1^2/2)$.

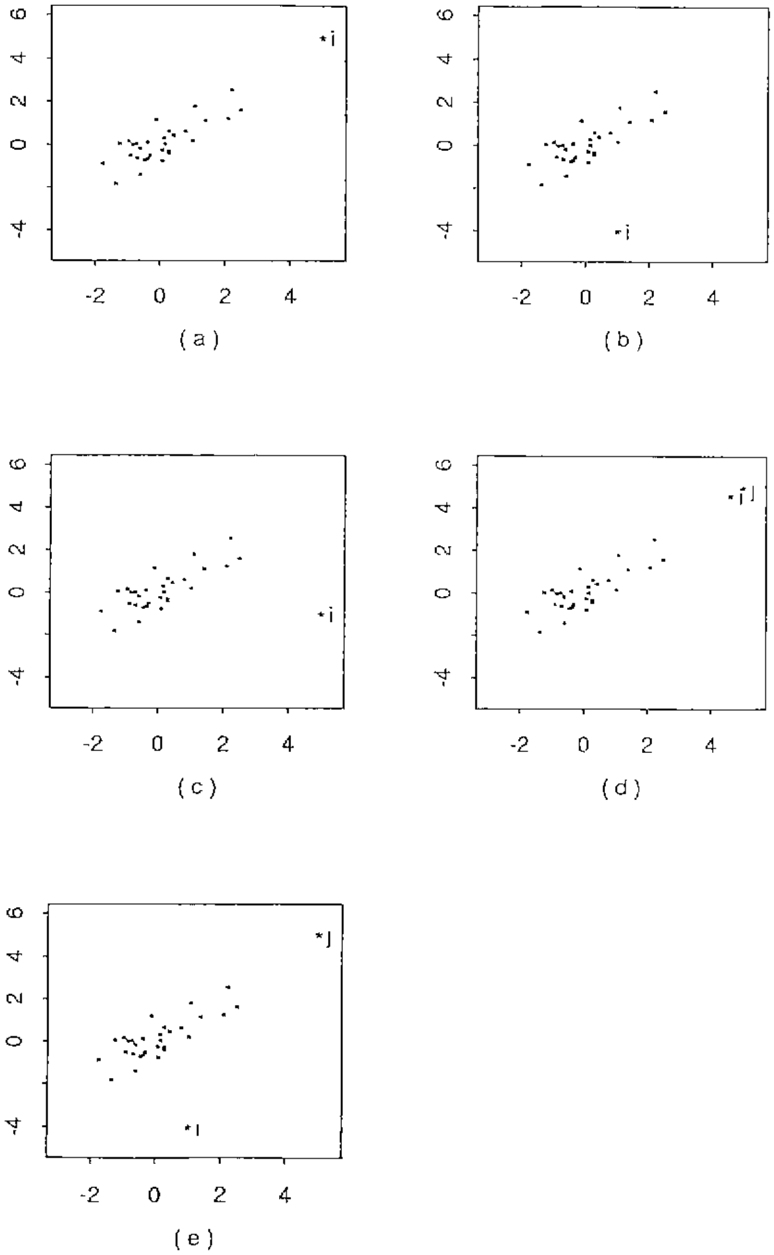
See the Appendix for the proof of Theorem 1. If (X_1, Y_1) follows standard bivariate normal distribution, i.e.

$$(X_1, Y_1) \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

then we can easily show that $\tau^2 = \{\lambda/(1-\lambda)\}(1-\rho^2)^2$. For practical use, we can replace τ^2 by $\hat{\tau}^2 = \{\hat{\lambda}/(1-\hat{\lambda})\}(1-r^2)^2$, where $\hat{\lambda} = k/n$.

3. INTERPRETATION OF ISOLATED CASES AND AN EXAMPLE

Figure 2 shows several types of isolated cases. We first discuss Figure 2(a) and 2(b). Deletion of the i-th case in Figure 2(a) makes r smaller ($r = 0.878$, $r_{(i)} = 0.781$), and deletion of the i-th case in Figure 2(b) makes r larger ($r = 0.504$, $r_{(i)} = 0.781$). In the context of linear regression, these cases can be regarded as "high leverage point" and "outlier", respectively. Therefore, a high leverage point makes r larger and an outlier makes r smaller. The case i in Figure 2(c) is both a high leverage point and an outlier, and deletion of the case i makes r larger(r= 0.444, $r_{(i)} = 0.781$). Hence, if a high leverage point is an outlier, the "high leverage" effect is hidden by the "outlier" effect. For Figure 2(d), two cases i and j are high leverage points, and $r = 0.910$, $r_{(i)} = 0.878$, $r_{(j)} = 0.869$, and $r_{(i,j)} = 0.781$, i.e., deletion of case i or case j does not alter r very much, but deletion of both cases i and j does make r smaller. Therefore, cases i and j are individually not influential, but are simultaneously influential. This phenomenon is so called the "masking effect" in the regression diagnostics context. Finally,



(a) (b) (c) (d) (e)

Figure 2. Various types of isolated cases

case i is a high leverage point and case j is an outlier in Figure 2(e). Since $r = 0.697$, $r_{(i)} = 0.504$, $r_{(j)} = 0.878$, and $r_{(i,j)} = 0.781$, they are individually influential, but are not simultaneously influential. We call this phenomenon as "cancelling effect", an apposite concept of the masking effect.

As an illustrative example we use the Batting Average data introduced in Section 1, and the data are given in Table 1. As shown in Figure 1 we see three cases appear to be isolated : case 12 (Wade Boggs(.368, .357)), case 71 (Don Mattingly (.324, .352)) and case 92 (Floyd Rayford (.306, .176)). Cases 12 and 71 are high leverage points and case 92 is an outlier. Table 2 lists ten most influential cases based on $\Delta_K = r_K - r$ for k=1 and 2. Reference values based on Theorem 1 are .011 and .016 for k=1 and 2, respectively. (α=.01 is used). Based on these values, 7 cases turned out to be influential for k=1, and 478 pairs of cases out of $\begin{pmatrix} 124 \\ 2 \end{pmatrix}$ pairs are influential for k=2. However, it is not realistic in the sense that almost all cases are influential when k=2. As argued by Kim and Storer(1996), "relative influence" must be considered. For example, Δ_K of case 92 is .05541 and this value is much larger than others, but Δ_K of cases (92, 97) is .0719 and it is not relatively larger than others. Note that ten most influential pairs of cases for k=2 contain case 92 except for the second pair (12, 71). Hence, they are influential due to the "swamping phenomenon". Also, cases 92 and 12 are individually influential, and not simultaneously influential (r=.554 and Δ_K=.018 for K=(92, 12)) because of the cancelling effect. Conclusively, we might say that cases 92, 12, and 71 are individually influential.

Table 1. The Batting Average data in Wardrop(1995).

No	1985	1986	No	1985	1986	No	1985	1986	No	1985	1986
1	0.265	0.264	32	0.248	0.241	63	0.230	0.218	94	0.286	0.252
2	0.309	0.296	33	0.263	0.259	64	0.293	0.287	95	0.282	0.282
3	0.268	0.240	34	0.289	0.310	65	0.277	0.284	96	0.251	0.210
4	0.243	0.229	35	0.238	0.221	66	0.259	0.261	97	0.300	0.233
5	0.289	0.289	36	0.256	0.300	67	0.265	0.251	98	0.219	0.249
6	0.266	0.286	37	0.288	0.306	68	0.263	0.287	99	0.262	0.272
7	0.231	0.238	38	0.246	0.287	69	0.264	0.227	100	0.283	0.228
8	0.275	0.309	39	0.225	0.250	70	0.218	0.254	101	0.228	0.237
9	0.304	0.300	40	0.254	0.274	71	0.324	0.352	102	0.280	0.264
10	0.274	0.301	41	0.282	0.281	72	0.239	0.266	103	0.258	0.246
11	0.295	0.256	42	0.295	0.258	73	0.259	0.252	104	0.257	0.287
12	0.368	0.357	43	0.287	0.268	73	0.297	0.281	105	0.245	0.212
13	0.248	0.222	44	0.242	0.268	75	0.232	0.260	106	0.275	0.326
14	0.300	0.310	45	0.270	0.285	76	0.259	0.253	107	0.251	0.204
15	0.335	0.290	46	0.273	0.250	77	0.222	0.203	108	0.215	0.187
16	0.237	0.270	47	0.270	0.218	78	0.295	0.259	109	0.236	0.229
17	0.242	0.256	48	0.296	0.323	79	0.233	0.252	110	0.313	0.265
18	0.299	0.267	49	0.282	0.278	80	0.297	0.305	111	0.258	0.277
19	0.219	0.243	50	0.241	0.265	81	0.267	0.290	112	0.275	0.251
20	0.239	0.269	51	0.314	0.263	82	0.290	0.283	113	0.258	0.277
21	0.311	0.278	52	0.244	0.247	83	0.267	0.277	114	0.287	0.316
22	0.262	0.302	53	0.285	0.283	84	0.259	0.231	115	0.280	0.269
23	0.251	0.270	54	0.278	0.267	85	0.239	0.238	116	0.249	0.272
24	0.293	0.258	55	0.268	0.231	86	0.273	0.257	117	0.245	0.268
25	0.197	0.215	56	0.313	0.260	87	0.249	0.276	118	0.285	0.251
26	0.287	0.271	57	0.252	0.241	88	0.257	0.258	119	0.189	0.219
27	0.287	0.268	58	0.274	0.288	89	0.275	0.265	120	0.244	0.237
28	0.244	0.256	59	0.260	0.250	90	0.288	0.328	121	0.278	0.269
29	0.254	0.208	60	0.231	0.229	91	0.276	0.276	122	0.275	0.262
30	0.263	0.267	61	0.243	0.240	92	0.306	0.176	123	0.273	0.252
31	0.262	0.302	62	0.238	0.210	93	0.291	0.324	124	0.277	0.312

Table 2. Ten largest cases based on $\Delta_K = r_{(K)} - r$ for k=1 and 2 in the Batting Average data.

cases	Δ_K	*cases*		Δ_K
92.	0.05541	92.	97.	0.07190
12.	-0.04101	12.	71.	-0.06678
71.	-0.02053	92.	100.	0.06564
108.	-0.01591	56.	92.	0.06463
97.	0.01483	51.	92.	0.06356
25.	-0.01197	47.	92.	0.06281
77.	-0.01102	36.	92.	0.06253
119.	-0.01001	38.	92.	0.06242
100.	0.00903	92.	110.	0.06233
56.	0.00795	22.	92.	0.06067

APPENDIX

(Proof of Eq.(1))

The equation (1) is a direct consequence of the following three identities :

$$S_{XX(K)} = S_{XX} - \frac{1}{n-k}\{\sum_{j\in K} p_j\}^2 - \sum_{j\in K} p_j^2 \tag{A.1}$$

$$S_{YY(K)} = S_{YY} - \frac{1}{n-k}\{\sum_{j\in K} q_j\}^2 - \sum_{j\in K} q_j^2 \tag{A.2}$$

$$S_{XY(K)} = S_{XY} - \frac{1}{n-k}\sum_{j\in K} p_j \sum_{j\in K} q_j - \sum_{j\in K} p_j q_j \tag{A.3}$$

where $S_{XX(K)}$, $S_{YY(K)}$ and $S_{XY(K)}$ are the corresponding versions of S_{XX}, S_{YY} and S_{XY} with the k observations deleted. We prove (A.3) only. The other two equations follow by a parallel argument.

We start by noting that $\bar{X}-\bar{X}_{(K)} = \sum_{j\in K} p_j/(n-k)$ and $\bar{Y}-\bar{Y}_{(K)} = \sum_{j\in K} q_j/(n-k)$. Now

$$\begin{aligned} S_{XY(K)} &= \sum_{j\notin K}(X_j - \bar{X} + \bar{X} - \bar{X}_{(K)})(Y_j - \bar{Y} + \bar{Y} - \bar{Y}_{(K)}) \\ &= \sum_{j\notin K} p_j q_j + (\bar{X} - \bar{X}_{(K)}) \sum_{j\notin K} q_j + (\bar{Y} - \bar{Y}_{(K)}) \sum_{j\notin K} p_j \\ &\qquad +(n-k)(\bar{X} - \bar{X}_{(K)})(\bar{Y} - \bar{Y}_{(K)}) \\ &= \sum_{j\notin K} p_j q_j + \frac{1}{n-k} \sum_{j\in K} p_j (-\sum_{j\in K} q_j) + \frac{1}{n-k} \sum_{j\in K} q_j (-\sum_{j\in K} p_j) \\ &\quad +\frac{1}{n-k} \sum_{j\in K} p_j \sum_{j\in K} q_j . \end{aligned}$$

The desired result follows directly from this.

(Proof of Theorem 1)

Without lose of generality, we may assume $\mu_X = \mu_Y = 0$. We first observe that, for $l = 0, 1, 2$,

$$n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^l \, (Y_i - \bar{Y})^{2-l} = n^{-1} \sum_{i=1}^{n} X_i^l \, Y_i^{2-l} + o_p(n^{-1/2}) \, .$$

This enables use to linearize r in the form of

$$\begin{aligned} r &= T_{XY}/\sigma_X\sigma_Y - \rho\,(T_{XX} - \sigma_X^2)/2\sigma_X^2 - \rho\,(T_{YY} - \sigma_Y^2)/2\sigma_Y^2 + o_p(n^{-1/2}) \\ &= \rho + T_{XY}/\sigma_X\sigma_Y - \rho\,T_{XX}/2\sigma_X^2 - \rho\,T_{YY}/2\sigma_Y^2 + o_p(n^{-1/2}) \qquad \text{(A.4)} \end{aligned}$$

where $T_{XX} = \sum_{i=1}^{n} X_i^2/n$, $T_{YY} = \sum_{i=1}^{n} Y_i^2/n$, and $T_{XY} = \sum_{i=1}^{n} X_i Y_i/n$. The asymptotic expansion (A.4) is valid for $r_{(K)}$ too if we replace the sample moments T_{XX}, T_{YY}, T_{XY} by the corresponding partial averages of X_i^2, Y_i^2, X_iY_i for $i \notin K$. Let

$$h_i = \begin{cases} -1/n^{1/2} & if \quad i \in K \\ k/\{n^{1/2}(n-k)\} & if \quad i \notin K \end{cases}$$

then from (A.4) and its equivalent for $r_{(K)}$ we can write

$$n^{1/2}(r_{(K)} - r) = \sum_{i=1}^{n} h_i (Z_i W_i - \rho\, Z_i^2/2 - \rho\, W_i^2/2) + o_p(1) \qquad \text{(A.5)}$$

Note that $\sum_{i=1}^{n} h_i = 0$ and $\sum_{i=1}^{n} h_i^2 = k/(n-k)$, and therefore the mean and the variance of the sum in (A.5) are zero and $\{k/(n-k)\} Var(Z_1W_1 - \rho Z_1^2/2 - \rho W_1^2/2)$, respectively. The asymptotic normality can be easily verified by checking the Lindeberg condition.

References

Kim. C and Storer, B. E. (1996) Reference values for Cook's distance, *Communications in Statistics - Simulation and Computations*, 25; 691-708.

Wardrop, R. L. (1995) *Statistics : Learning in the Presence of Variation*, Wm. C. Brown Publishers : Dubuque, Iowa.