The Korean Communications
in Statistics Vol. 12 No. 3, 2005
pp. 635-642

Influence Measures for a Test Statistic on Independence of Two Random Vectors[1)]

Kang-Mo Jung[2)]

Abstract

In statistical diagnostics a large number of influence measures have been proposed for identifying outliers and influential observations. However it seems to be few accounts of the influence diagnostics on test statistics. We study influence analysis on the likelihood ratio test statistic whether the two sets of variables are uncorrelated with one another or not. The influence of observations is measured using the case-deletion approach, the influence function. We compared the proposed influence measures through two illustrative examples.

Keywords : Covariance matrix, deletion, influence function, likelihood ratio test.

1. Introduction

The detection of outliers or influential observations has a long history. However, many diagnostic measures have been proposed for influence analysis in the context of estimation. A few works that treat detection of influential observations for test statistics in multivariate analysis are found. Among others, Jung (2001) investigated the influence of observations on the likelihood ratio test (LRT) statistics in the canonical correlation analysis using the local influence method introduced by Cook (1986). Influence analysis in testing problems is very important because in extreme situations, few observations can dominate our conclusion about the hypothesis as can be seen in Section 3.

Assume that the random vector $z = (x^T, y^T)^T$ has the covariance matrix Σ, where Σ is partitioned such that

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

And that x and y are p and q dimensional random vectors, respectively.

1) This work was supported by Korea Research Foundation Grant (KRF-2005-202-C00076).

2) Associate Professor, Department of Informatics & Statistics, Kunsan National University, 68 Miryong-Dong, Kunsan, Chollapuk-Do, 573-701, Korea.
Email : kmjung@kunsan.ac.kr

Consider the hypothesis

$$H_0 : \Sigma_{12} = 0 , \tag{1}$$

which means the two sets of variables are uncorrelated with one another. Under the normality, the LRT statistic for testing H_0 is given by

$$T = -(n-(p+q+3)/2)\log\frac{|S|}{|S_{11}||S_{22}|} \tag{2}$$

where S, S_{11}, S_{22} are the maximum likelihood estimators (MLE) of $\Sigma, \Sigma_{11}, \Sigma_{22}$, respectively, and n is the number of observations. Then the test statistic is approximately distributed as a chi-squared distribution with pq degrees of freedom using Bartlett's approximation (Mardia, et al., 1979, pp. 288).

It is well known that the sample covariance matrix is very sensitive to outliers (Critchley, 1985), and so is the test statistic T. To investigate the influence of observations on the test statistic (2), Jung (2001) considered the local influence method using the fact that the test statistic can be written by the squared canonical correlation coefficients. Even though the local influence method is effective in finding outliers and influential observations, the deletion diagnostics are fundamental for confirmatory analysis (Fung, 1993). To accomplish these objectives in this work we considered the deletion approach and the influence function. The proposed diagnostic measures can be expressed in terms of statistics without involving the actual deletion of observations. The deletion approach is widely used in many statistical analysis (Cook and Weisberg, 1982). However, case-deletion diagnostics require amount of computation time. We obtained the case-deletion diagnostic measure which can be expressed in terms of statistics without involving the actual deletion of observations. It is usual to use single case-deletion diagnostic for influence analysis, because double case-deletion diagnostic has somewhat complex form. We obtained double case-deletion diagnostic on the test statistic T. Thus the phenomenon behind influential observations can be explained through double or conditional case-deletion.

In Section 2 we will derive the case-deletion diagnostic and the influence function of T. The former has the results which are single case-deletion, double case-deletion and conditional deletion, while the latter has three sample versions which are empirical influence function, sample influence function and deleted influence function. In Section 3 two numerical examples will be given for illustration.

2. Influence Measures

The random sample $\{z_1, \ldots, z_n\}$ is drawn from $(p+q)$-variate normal distribution $N(\mu, \Sigma)$. Assume that z_u is decomposed as in Section 1, that is, $z_u = (x_u^T, y_u^T)^T$. Then MLE of Σ becomes $S = (1/n)\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T$. Also the MLEs S_{11} and S_{22} of Σ_{11} and Σ_{22} are similarly obtained, respectively.

2.1 Deletion Diagnostic

We will derive the deletion diagnostic for the statistic T. Let $T_{(u)}$ be the statistic with the deletion of uth observation z_u. Hereafter we denote by the subscript (u) the estimator or statistic based on the reduced data set without observation z_u. Since $S_{(u)} = 1/(n-1)\sum_{i \neq u}^{n}(z_i - \bar{z})(z_i - \bar{z})^T$ and $\bar{z}_{(u)} = \sum_{i \neq u} z_i/(n-1)$, it follows that

$$S_{(u)} = \frac{n}{n-1}[S - \frac{1}{n-1}(z_u - \bar{z})(z_u - \bar{z})^T].$$

Thus we have

$$|S_{(u)}| = (\frac{n}{n-1})^{p+q}|S|[1 - \frac{1}{n-1}D_{z,uu}],$$

where $D_{z,uu} = (z_u - \bar{z})^T S^{-1}(z_u - \bar{z})$. Similarly $|S_{11(u)}|$ and $|S_{22(u)}|$ can be obtained, where $D_{x,uu}$ and $D_{y,uu}$ are similarly defined as $D_{z,uu}$. Therefore

$$T_{(u)} = (1 + 1/c)T + (c+1)\left\{\log(1 - \frac{D_{z,uu}}{n-1}) - \log(1 - \frac{D_{x,uu}}{n-1}) - \log(1 - \frac{D_{y,uu}}{n-1})\right\}, \quad (3)$$

where $c = -(n - (p+q+3)/2)$.

Further we will derive the deletion diagnostic for the test statistic T when two observations are omitted. Similarly in case single case-deletion the relationship for double case-deletion

$$|S_{(uv)}| = (\frac{n-1}{n-2})^{p+q}|S_{(u)}|(1 - \frac{D_{z,(u)v}}{n-1}),$$

where $D_{z,(u)v} = (z_v - \bar{z}_{(u)})^T S_{(u)}^{-1}(z_v - \bar{z}_{(u)})$ gives

$$|S_{(uv)}| = (\frac{n}{n-2})^{p+q}|S|A_{z,uv},$$

where

$$A_{z,uv} = [1 - \frac{n-1}{n(n-2)}\left\{(D_{z,uu} + D_{z,uv}) + \frac{1}{n-1}(D_{z,uv}^2 + 2D_{z,uv} - D_{z,uu}D_{z,vv})\right\}],$$

which is due to the fact that

$$S_{(uv)}^{-1} = \frac{n-1}{n}[S^{-1} + \frac{1}{n-1-D_{z,uu}}S^{-1}(z_v - \bar{z})(z_v - \bar{z})^T S^{-1}].$$

In the same manner $|S_{11(uv)}|$ and $|S_{22(uv)}|$ can be derived. Thus we obtained $T_{(uv)}$ which can be expressed in terms of statistics without involving the actual deletion of observations. That is,

$$T_{(uv)} = (1 + 2/c)T + (c+2)(\log A_{z,uv} - \log A_{x,uv} - \log A_{y,uv}). \quad (4)$$

This indicates the joint influence of observations u and v on the test statistic. When $T_{(uv)}$ and $T_{(v)}$ are large but $T_{(u)}$ is small, a swamping effect of observation u by observation v

can be inferred (Lawrence, 1995).

The influence effect of observation u on the test statistic T after deleting observation v can be detected by $T_{(u(v))} = T_{(uv)} - T_{(v)}$. This is called a conditional influence measure. Assume that $T_{(v)}$ and $T_{(u(v))}$ are large but $T_{(u)}$ is small. This situation implies that the influence of observation u may be masked by observation v. The proposed conditional influence measure can be useful to detect a masking effect. $T_{(u(v))}$ can be rewritten as

$$T_{(u(v))} = \frac{T}{c} + \log\left(\frac{A_{z,uv}}{A_{x,uv}A_{y,uv}}\right) + (c+1)\left\{\log\left(\frac{A_{z,uv}}{1-D_{z,uv}/(n-1)}\right)\right.$$
$$\left. - \log\left(\frac{A_{x,uv}}{1-D_{x,uv}/(n-1)}\right) - \log\left(\frac{A_{y,uv}}{1-D_{y,uv}/(n-1)}\right)\right\} \tag{5}$$

2.2 Influence Function

In this section we will derive the influence function for the test statistic T and consider three sample versions of the influence function that will be used for investigating the influence of observations on the test statistic for the hypothesis (1).

Let F be a distribution function defined on the p-dimensional Euclidean space and $\theta = \theta(F)$ be a parameter of interest which is a functional of F. The mean vector and covariance matrix for the distribution F are written as $\mu = \mu(F)$ and $\Sigma = \Sigma(F)$, respectively. For $0 \leqq \epsilon \leqq 1$, the perturbation of F at z is defined by $F_\epsilon = (1-\epsilon)F + \epsilon\delta_z$, where δ_z denotes the distribution having unit mass at z. The perturbation of θ at z is $\theta(F_\epsilon)$. The influence function for θ at z (Hampel, 1974) is defined by

$$\lim_{\epsilon \to 0} \frac{\theta(F_\epsilon) - \theta(F)}{\epsilon}. \tag{6}$$

The influence function for a parameter at z measures the effect of an infinitesimal contamination at z on the estimator of the parameter. Hence the influence function can serve as a diagnostic method of detecting influential observations in performing a test of hypothesis.

We have known that

$$\Sigma(F_\epsilon) = \Sigma(F) + \{(z-\mu(F))(z-\mu(F))^T - \Sigma(F)\}\epsilon + O(\epsilon^2).$$

The equation (3.1) of Jung (2002) yields

$$|\Sigma(F_\epsilon)| = |\Sigma(F)| + |\Sigma(F)|\,\mathrm{tr}\,\{\Sigma(F)^{-1}(z-\mu(F))(z-\mu(F))^T - I\}\epsilon + O(\epsilon^2).$$

It follows immediately that

$$IF(T,F,z) = c\left\{(z-\mu(F))^T\Sigma^{-1}(F)(z-\mu(F)) - (x-\mu_1(F))^T\Sigma_{11}^{-1}(F)(x-\mu_1(F))\right.$$
$$\left. - (y-\mu_{22}(F))^T\Sigma_{22}^{-1}(F)(y-\mu_{22}(F))\right\} \tag{7}$$

We will consider three sample versions as in Critchley (1985) : the empirical influence function (EIF), the deleted empirical influence function (DIF) and the sample influence function (SIF). A large absolute value of each sample version indicates that the corresponding

observation is influential.

The EIF is obtained by substituting the empirical distribution function $\hat{F}$ and observation z_u for F and z in (7), respectively. The EIF of T at z_u becomes

$$EIF(T, z_u) = c\,(D_{z,uu} - D_{x,uu} - D_{y,uu}) \tag{8}$$

Equation (8) can be rewritten as

$$EIF(T, z_u) = c\,(z_u - \bar{z}\,)^T(S^{-1} - S_0^{-1})(z_u - \bar{z}\,),$$

where S_0 is the MLE of covariance matrix under H_0 in (1).

The SIF can be obtained by setting $F = \hat{F}$ and taking $\epsilon = -1/(n-1)$ in the definition of the influence function (7) instead of taking a limit. Then the SIF for a parameter θ at z_u can be rewritten as $(n-1)\{\theta(\hat{F}) - \theta(\hat{F}_u)\}$, where $\hat{F}_{(u)} = (1 + (n-1)^{-1})\hat{F} - (n-1)^{-1}\delta_{z_u}$ is the deleted version of $\hat{F}$ with the uth observation z_u deleted. Thus from (3) it follows that the SIF of T as

$$SIF(T, z_u) = (n-1)\,(T - T_{(u)}) \tag{9}$$

The DIF is obtained by replacing F with $\hat{F}_{(u)}$ in (7) and it measures the effect of deleting the uth observation on the estimator. The mean vector for $\hat{F}_{(u)}$ is given by $\mu(\hat{F}_{(u)}) = \bar{z}_{(u)}$ and the covariance matrix for $\hat{F}_{(u)}$ is $\Sigma(\hat{F}_{(u)}) = S_{(u)}$ computed in the previous subsection. Thus we got

$$DIF(T, z_u) = c\,(D_{z,(u)u} - D_{x,(u)u} - D_{y,(u)u}),$$

where $D_{z,(u)u}$ is previously defined in Section 2.1. It follows from $D_{z,(u)u} = n(n-1)D_{uu}$ that

$$DIF(T, z_u) = c\,(D_{z,uu} - D_{x,uu} - D_{y,uu})$$

which is equivalent to $EIF(T, z_u)$.

3. Numerical Examples

3.1. Head-length data

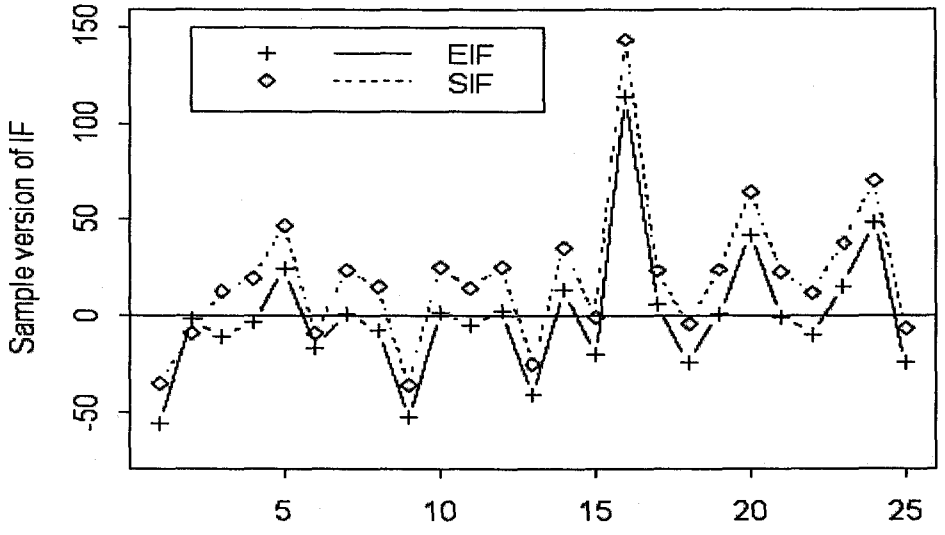
Diagnostic measures described in Section 2 was applied to the head-length data (Mardia, et al., 1979, p. 121, Table 5.1.1) previously analyzed by Jung (2001) based on the local influence method. For this data set, the number of observations is $n = 25$ and the dimensions are $p = 2, q = 2$. The LRT statistic based on the full data set is 20.96, and therefore we conclude that the null hypothesis is strongly rejected from the p-value 0.003.

We obtained information about influential observations for the test statistic T using the deletion diagnostic and the influence function.

<Table 1> Case-deletion results for the head-length data

Single case-deletion		Double case-deletion		Conditional case-deletion		
J	$T_{(J)} - T$	J	$T_{(J)} - T$	I	J	$T_{(J(I))}$
16	-5.96	16, 20	-9.35	16	20	-6.65
24	-2.92	16, 24	-8.99	16	6	-6.60
20	-2.69	16, 5	-8.15	16	23	-6.43
5	-1.96	16, 23	-8.02	16	13	-6.38
23	-1.59	16, 14	-7.29	16	22	-6.36

We carried out the single, double and conditional case-deletions, and the results are summarized in Table 1. Numbers in Table 1 are arranged in decreasing order of corresponding to the deletion measure, where $T_{(J)}$ denotes the statistic after deletion of the corresponding index set. The case deletion results show that observations 16, 20 and 24 are individually influential and observations 20 and 24 with observation 16 are jointly influential. And also the results of the conditional case-deletion indicates that there are no masking effects in the data set. Furthermore, the test statistic without observations 16, 20, 24 becomes 8.71. This gives the conclusion that the null hypothesis is not rejected. It implies that opposite conclusions are made by removing observations 16, 20, 24 or not. We conclude that observations 16, 20 and 24 are influential observations on the LRT statistic, and observation 16 is most influential.



<Fig. 1> The index plots for the LRT statistic T using EIF and SIF.

The results based on the influence function are presented in Fig. 1. The EIF and SIF have similar patterns, and they have the same influence information with single case-deletion.

From the results of deletion method and influence function we may conclude that observation 16 is most influential from all influence measures. Observations 20 and 24 are candidate for influential observations. The influence of observations 16, 20 and 24 are confirmed by p −value 0.068 of the LRT statistic with the remaining data set discarding those observations.

Single case-deletion shows the individual influence of an observation on the test statistic, while double case-deletion present the influence information about joint influence. And the conditional deletion measure provides the influence information about the masking effect.

3.2. Diabetics data

We considered another data set to show the effectiveness of the proposed method. The data set (Rencher, 1995, p. 74, Table 3.6) was surveyed for comparing normal patients and diabetics. Five variables (the first two variables are minor of interest and the last three variables are major) are measured for 46 patients. The LRT statistic based on the full data set is 13.73, and therefore we conclude that the null hypothesis about the independence between two groups is rejected with 5% significance level.

The results of single and double case-deletions are summarized in Table 2. Even though observations 6 and 37 are most influential from the individual influence point of view, the omission of observations 26 and 27 changes the rejection about the null hypothesis based on the full data into the acceptance. Thus the latter observations are more influential than the former. The data set without observations in the results of double case-deletion has the p -values larger than 0.05. That is, the corresponding data sets reject the null hypothesis. In the double case-deletion observations 26 and 27 are still influential. To investigate the swamping effect of observations 26 and 27 we conducted conditional case-deletions. Table 2 shows that there is no swamping effect between observations 26 and 27. Thus we may conclude that observations 26 and 27 are very influential. This numerical examples illustrated that conditional case-deletion confirmed the joint influence of individually influential observations.

4. Concluding Remarks

Case-deletion diagnostics are fundamental for investigating the influence of observations on the statistic or estimator. However, the computing load is unmanageable for large data set. We derived the single and double case-deletion diagnostics for the testing statistic about the independence of random vectors, which are expressed in terms of statistics without involving the actual deletion of observations. From these diagnostics we may observe

<Table 2> Case-deletion results for the diabetics data

Single case-deletion		Double case-deletion		Conditional case-deletion		
J	$p-value$	J	$p-value$	I	J	$p-value$
6	.002	26, 27	.310	27	6	.002
37	.008	15, 26	.177	27	26	.111
27	.105	15, 27	.176	27	37	.009
26	.101	25, 26	.175	27	8	.017
8	.015	25, 37	.172	27	15	.058

the single influence, the joint influence and the conditional influence. And we showed the algebraic relationship between case-deletion diagnostics and the influence function diagnostics.

References

[1] Cook, R.D. (1986). Assessment of local influence (with discussions), *Journal of the Royal Statistical Society B*, 48, 133-169.

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

[3] Critchley, F. (1985). Influence in principal component analysis, *Biometrika*, 72, 627-636.

[4] Fung, W.-K. (1993). Unmasking outliers and leverage points : A confirmation, *Journal of the American Statistical Association*, 88, 545-519.

[5] Hampel, F. R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, 69, 383-393.

[6] Jung, K.-M. (2001). Influence analysis on a test statistic in canonical correlation coefficients, *The Korean Communications in Statistics*, 8, 347-355.

[7] Jung, K.-M. (2002). Influence function of the likelihood ratio test statistic for multivariate normal sample, *Communications in Statistics - Theory and Methods*, 31, 1273-1281.

[8] Lawrence, A. J. (1995). Deletion influence and masking in regression, *Journal of the Royal Statistical Society B*, 57, 181-189.

[9] Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, New York.

[10] Rencher, A. C. (1995). *Methods of Multivariate Analysis*, John Wiley & Sons, New York.

[Received April 2005, Accepted September 2005]