Journal of the Korean Statistical Society (2001), **30**: *1, pp 139–151*

Some Diagnostic Results in Discriminant Analysis[†]

Whasoo Bae[1] and Soonyoung Hwang[2]

ABSTRACT

Although lots of works are done in influence diagnostics, results in the multivariate analysis are quite rare. One of recent works done by Fung(1995) is about the single case influence diagnostics in the linear discriminant anaysis. In this paper we extend Fung's results to the multiple cases diagnostics which are necessary in the linear discriminant analysis for two reasons among others ; First, the masking effect cannot be detected by single case diagnostics and secondly two populations are concerned in the discriminant analysis, i.e., influential cases can occur in one or both populations.

Keywords: Influential observations, Masking effect, Swamping phenomenon

1. INTRODUCTION

Identification of influential observations or influential subsets is mainly focused on regression analysis in the last decade. Also, most of influence measures suggested so far are concerned about the influence of observations on the estimates of regression coefficient. Cook's distance(1977) is one of the most widely used influence measure in linear regression, and Kim and Storer(1996) studied reference values for Cook's distance. Cook(1986) suggested the local influence and Kim(1996) suggested the replacement measure. Also, Kim and Hwang(2000) studied the influence diagnostics on the Mallows' C_p. Pregibon(1981) suggested one-step estimator in the logistic regression diagnostics. In Box-Cox transformation model, Cook and Wang(1983), Hinkley and Wang(1988), Tasi and Wu(1990), Kim, Storer and Jeong(1996) studied the influence on the transformation parameter. Also, regression diagnostics in nonparametric regression models are studied by Eubank(1985), Silverman(1985), Thomas(1991), and Kim(1996).

[†] This work was supported by the Inje Scholarship Foundation in 1998.

[1] Department of Data Science, Inje University, Kimhae, 621-749, Korea.

[2] Department of Statistics, Pusan National University, Pusan, 609-735, Korea.

However, influence measures or method of detecting influential observations in multivariate analysis, such as the discriminant analysis, are very few. Among these, Campbell(1978) and Johnson(1987) studied identification of influential observations in discriminant analysis, and Fung(1992, 1995) suggested two basic building blocks and an influence measure on the discriminant score. However, these works are concerned about the influence of single observation from a specific population. As is well known in regression analysis, simultaneous influence of two or more observations is necessary because of the masking effect, and this phenomenon is also important in discriminant analysis.

In this paper we extend two fundamental statistics suggested by Fung (1995) to detect influential subsets on the Fisher's linear discriminant score in discriminant analysis. Notations and some results of Fung(1995) are summarized in Section 2, and the extension of Fung's results are described in Section 3. In Section 4, an example based on real data set is given. Also, concluding remarks are in Section 5.

2. NOTATIONS AND SINGLE CASE DELETION

To extend single case deletion diagnostic by Fung(1995) to multiple cases deletion, we introduce notations for discriminant analysis and Fung's results in this Section.

Let $\mathbf{y}_{1j}$ and $\mathbf{y}_{2j}$, $j = 1, 2, \cdots, n_i$ be p-vector random samples from two normal populations (π_1 and π_2) with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively and a common variance Ω. Also, let $n = n_1 + n_2$ be total observations from both populations. The Fisher's linear discriminant rule is to allocate an observation $\mathbf{y}$ of an unknown population to π_1 if

$$\boldsymbol{\alpha}'\mathbf{y} > \boldsymbol{\alpha}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$$

and to π_2 if otherwise. The discriminant coefficients $\boldsymbol{\alpha}$ can be estimated by $\hat{\boldsymbol{\alpha}} = \mathbf{S}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$, where $\bar{\mathbf{y}}_i$, $i = 1, 2$ are sample means and $\mathbf{S}$ is pooled covariance matrix given by

$$\mathbf{S} = \frac{1}{n-2}\left[\sum_{j=1}^{n_1}(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)' + \sum_{j=1}^{n_2}(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)'\right].$$

Fisher's linear discriminant rule considers whether or not $\hat{\boldsymbol{\alpha}}'\mathbf{y} - \hat{\boldsymbol{\alpha}}'(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)/2 > 0$, where the quantity $\hat{\boldsymbol{\alpha}}'\mathbf{y} - \hat{\boldsymbol{\alpha}}'(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)/2$ is called the discriminant score. Let

$\hat{\boldsymbol{\beta}}' = (-\hat{\boldsymbol{\alpha}}'(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)/2, \, \hat{\boldsymbol{\alpha}}')$ and $\mathbf{x}' = (1, \mathbf{y}')$, then we have $\hat{\boldsymbol{\beta}}'\mathbf{x} = \hat{\boldsymbol{\alpha}}'\mathbf{y} - \hat{\boldsymbol{\alpha}}'(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)/2$. We are interested in the effect of the deletion of observation i on the linear discriminant rule. Fung(1995) proposed

$$E(\hat{\boldsymbol{\beta}}'\mathbf{x} - \hat{\boldsymbol{\beta}}'_{(i)}\mathbf{x})^2 \tag{2.1}$$

being the mean squared difference of the discriminant scores for the full sample and the sample without observation i. For simplicity, he considered the deletion of observation i from population π_1. Fung(1995) showed that the measure given in (2.1) can be estimated prarametrically and nonparametrically. The parametric version is given by

$$E2 = tB_1^2 + (1-t)B_2^2 + V \ , \qquad t = \frac{n_1}{n}$$

where

$$B_1 = (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{(i)})'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)/2 - \hat{\boldsymbol{\alpha}}'_{(i)}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_{1(i)})/2,$$
$$B_2 = -(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{(i)})'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)/2 - \hat{\boldsymbol{\alpha}}'_{(i)}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_{1(i)})/2,$$

and

$$V = (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{(i)})'\mathbf{S}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{(i)}).$$

Here, $\hat{\boldsymbol{\alpha}}_{(i)}$ is the estimate of $\boldsymbol{\alpha}$ based on $n-1$ observations after deleting the observation i. On the other hand, the nonparametric version is given by

$$F2 = tB_1^2 + (1-t)B_2^2 + \frac{n-2}{n}V$$

which is very close to E2 for a large size n. Fung(1992) proposed two fundamental statistics, d_i^2 and $\hat{\psi}_i$, on which many influence measures depend. The proposed measures, E2 and F2 can be expressed in terms of $d_i^2 = (\mathbf{y}_{1i} - \bar{\mathbf{y}}_1)'\mathbf{S}^{-1}(\mathbf{y}_{1i} - \bar{\mathbf{y}}_1)$ and $\hat{\psi}_i = \hat{\boldsymbol{\alpha}}'(\mathbf{y}_{1i} - \bar{\mathbf{y}}_1)$. Let $D^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$, then it can be easily shown that d_i^2 and $\hat{\psi}_i/D$ are asymptotically ${\chi^2}_p$ and $N(0,1)$ distributed, respectively. Then, the statistics $DIF = d_i^2 - (\hat{\psi}_i/D)^2$ and $\hat{\psi}_i/D$ are asymptotically independent and distributed as ${\chi^2}_{p-1}$ and $N(0,1)$. The proposed measures are useful for detecting singly influential observation.

3. MULTIPLE CASES DELETION

The proposed measures in Section 2 are based on the deletion of single observation. Identification of multiple cases is necessary because of the masking effect. In this section we extend Fung's(1995) results to multiple cases deletion from both the populations π_1 and π_2.

Let $K = \{i_1, \cdots, i_k\}$ and $L = \{j_1, \cdots, j_l\}$ be index set of size k and l, respectively. We delete k observations in K from π_1 and l observations in L from π_2. For notational convenience, let

$$\mathbf{w}_{ij} = (\mathbf{y}_{ij} - \overline{\mathbf{y}}_i) \ , \qquad i = 1, 2$$

Then, the resulting estimator of the linear discriminant score $\boldsymbol{\alpha}$ after deleting $(k+l)$ observations is given by

$$\hat{\boldsymbol{\alpha}}_{(K \cup L)} = \mathbf{S}^{-1}_{(K \cup L)}(\overline{\mathbf{y}}_{1(K)} - \overline{\mathbf{y}}_{2(L)})$$

where

$$\overline{\mathbf{y}}_{1(K)} = \overline{\mathbf{y}}_1 - \frac{\mathbf{w}_K}{n_1 - k} \ , \qquad \overline{\mathbf{y}}_{2(L)} = \overline{\mathbf{y}}_2 - \frac{\mathbf{w}_L}{n_2 - l} \ ,$$
$$\mathbf{w}_K = \sum_{j \in K} \mathbf{w}_{1j} \ , \qquad \mathbf{w}_L = \sum_{j \in L} \mathbf{w}_{2j}$$

and the pooled covariance matrix based on $(n-k-l)$ observations can be shown (see Appendix for proof) as

$$\begin{aligned} \mathbf{S}_{(K \cup L)} &= \frac{1}{n-k-l-2}[(n-2)\mathbf{S} - \frac{1}{n_1 - k}\mathbf{w}_K \mathbf{w}_K{}' \\ &\quad - \frac{1}{n_2 - l}\mathbf{w}_L \mathbf{w}_L{}' - \sum_{j \in K} \mathbf{w}_{1j}\mathbf{w}_{1j}{}' - \sum_{j \in L} \mathbf{w}_{2j}\mathbf{w}_{2j}{}']. \end{aligned} \tag{3.1}$$

Similar to Eq.(2.1), the mean squared difference of the discriminant scores for the full sample and the sample without $(k+l)$ observations is given by

$$E(\hat{\boldsymbol{\beta}}'\mathbf{x} - \hat{\boldsymbol{\beta}}'_{(K \cup L)}\mathbf{x})^2 .$$

Then, the parametric version becomes

$$E2_{(K \cup L)} = tB^2_{1(K \cup L)} + (1-t)B^2_{2(K \cup L)} + V_{(K \cup L)} \ , \qquad t = \frac{n_1}{n}$$

where

$$\begin{aligned} B_{1(K \cup L)} &= (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{(K \cup L)})'(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_2)/2 - \hat{\boldsymbol{\alpha}}'_{(K \cup L)}(\overline{\mathbf{y}}_1 - \overline{\mathbf{y}}_{1(K)})/2 \\ &\quad - \hat{\boldsymbol{\alpha}}'_{(K \cup L)}(\overline{\mathbf{y}}_2 - \overline{\mathbf{y}}_{2(L)})/2, \end{aligned}$$

$$B_{2(K\cup L)} = -(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{(K\cup L)})'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)/2 - \hat{\boldsymbol{\alpha}}'_{(K\cup L)}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_{1(K)})/2$$
$$-\hat{\boldsymbol{\alpha}}'_{(K\cup L)}(\bar{\mathbf{y}}_2 - \bar{\mathbf{y}}_{2(L)})/2,$$

and

$$V_{(K\cup L)} = (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{(K\cup L)})'\mathbf{S}(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{(K\cup L)}).$$

Similarly, the nonparametric version becomes

$$F2_{(K\cup L)} = tB^2_{1(K\cup L)} + (1-t)B^2_{2(K\cup L)} + \frac{n-2}{n}V_{(K\cup L)}.$$

Set of observations with index $\{K \cup L\}$ can be regarded as influential if $F2_{(K\cup L)}$ is relatively large. We shall mainly discuss $F2_{(K\cup L)}$ because $F2_{(K\cup L)}$ is very close to $E2_{(K\cup L)}$. Also, the two fundamental influence statistics d_i^2 and $\hat{\psi}_i$ can be extended to $d^2_{K\cup L} = \sum_{i\in K\cup L} d_i^2$ and $\hat{\psi}_{K\cup L} = \sum_{i\in K\cup L}\hat{\psi}_i$.

For the multiple cases deletion, asymptotic distributions for the fundamental statistics such as $d^2_{K\cup L}$, $(\hat{\psi}_{K\cup L}/D)^2$, $DIF_{K\cup L}$ and the influence measure $F2_{(K\cup L)}$ are hard to derive. As an alternative approach, we might use Monte Carlo simulation study for the reference values for $d^2_{K\cup L}$, $(\hat{\psi}_{K\cup L}/D)^2$, $DIF_{K\cup L}$ and $F2_{(K\cup L)}$. Of course, the simulation results will depend on the sample size, the number of variables, and the number of cases deleted. There are several methods of simulation study in computing the reference values. Among them, Atkinson(1981) suggested a Monte Carlo testing method and Kim and Storer(1996) suggested a Monte Carlo distribution for the maximum values under the assumption that no influential observation exists. The method suggested by Atkinson(1981) is useful when one is concerned about the reference value for the single case deletion, but, it is almost computationally infeasible for the multiple cases deletion. Here we take the method by Kim and Storer(1996) since it is especially convenient to get a reference value for multiple cases deletion. We will explain this method and apply to the real data in Section 4.

4. EXAMPLE

As an illustrative example, we use an annual financial data(Johnson and Wichern, 1987, p. 526) collected for firms approximately 2 years prior to bankruptcy and for financially sound firms at about the same point in time. Variables considered are X_1 = (cash flow)/(total debt), X_2 = (net income)/(total assets), X_3 = (current assets)/(current liabilities), X_4 = (current assets)/(net sales). Observations from bankrupt firms are labeled from 1 to 21 and those from nonbankrupt firms are labeled from 22 to 46. Therefore, $p = 4$, $n_1 = 21$, and $n_2 = 25$.

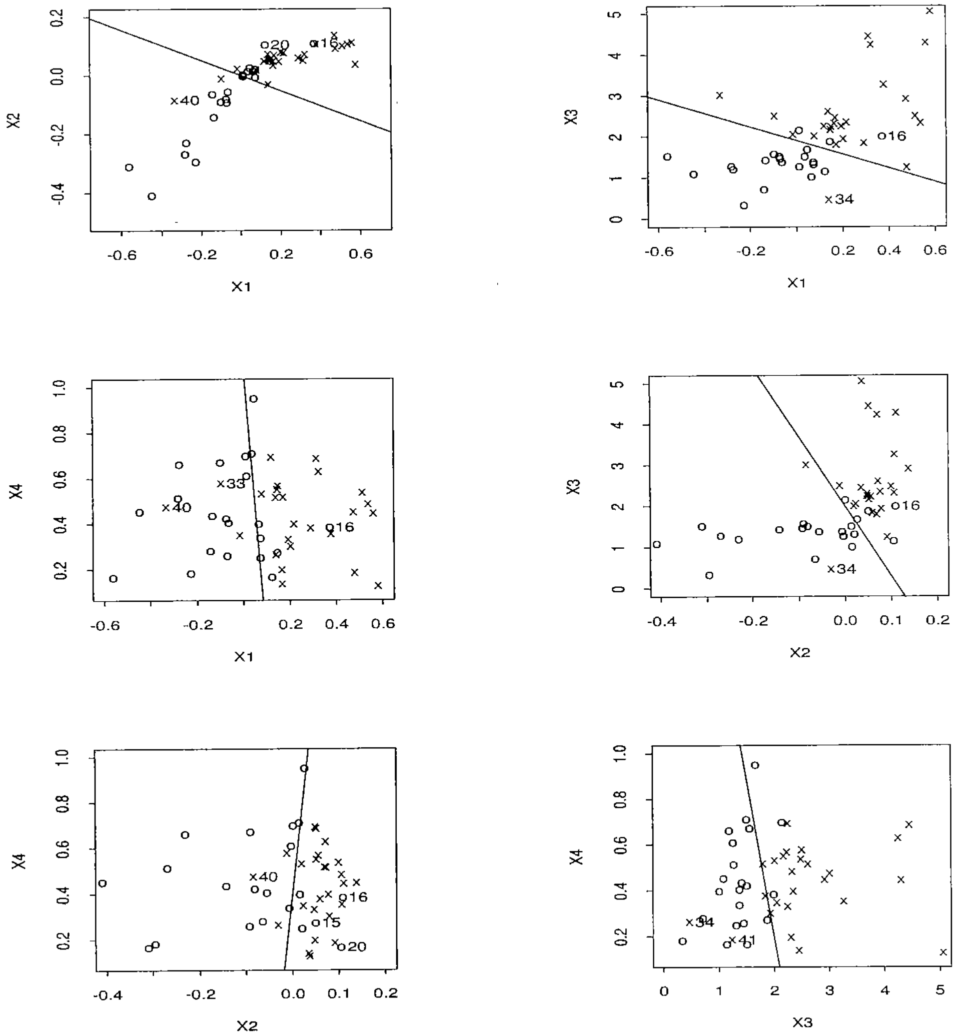

Figure 4.1. The plots for the pairs of variables (X_1, X_2), (X_1, X_3), (X_1, X_4), (X_2, X_3), (X_2, X_4) and (X_3, X_4). In each plot, a straight line denotes a Fisher's linear discriminant rule. (o : population 1, x : population 2)

The plots for the pairs of variables (X_1, X_2), (X_1, X_3), (X_1, X_4), (X_2, X_3), (X_2, X_4), and (X_3, X_4) are in Figure 4.1. As shown in Figure 4.1, cases 15, 16, 20, 33, 34, 40, 41 seem to be influential.

However, to assess the exact influence of each observation, we have to evaluate basic building blocks $d^2_{K\cup L}$, $(\hat{\psi}_{K\cup L}/D)^2$, $DIF_{K\cup L}$ and influence measure $F2_{(K\cup L)}$. The five largest observations with $d^2_{K\cup L}$, $(\hat{\psi}_{K\cup L}/D)^2$, and $DIF_{K\cup L}$, for $k+l = 1,2,3,4$ are given in Table 4.1 and those with $F2_{(K\cup L)}$ are given in Table 4.2. As shown in Table 4.1, observations with large values of $F2_{(K\cup L)}$ tend to have large values at least one of $d^2_{K\cup L}$, $(\hat{\psi}_{K\cup L}/D)^2$, $DIF_{K\cup L}$. Therefore, they behave quite well as basic building blocks. When one case $(k+l=1)$ is deleted observations 46 and 34 are very influential. If we delete two cases, the most influential set is (34, 46), and this set is influential due to the swamping phenomenon. However, if we delete three cases, the most influential set is (31, 36, 44). This set cannot be detected as influential if single case deletion diagnostic is used. This set is a good illustration of the masking effect which can only be revealed by the multiple cases deletion. Conclusively, observations 31, 34, 36, 44, 46 are quite influential. Note that these observations are quite different from those detected by "eye"(15, 16, 20, 34, 40) in Figure 4.1.

To get Monte Carlo reference values for the influence measures $d^2_{K\cup L}$, $(\hat{\psi}_{K\cup L}/D)^2$, $DIF_{K\cup L}$ and $F2_{(K\cup L)}$, we take the method of Kim and Storer(1996). First, generate $n_1 = 21$ random vectors $(p=4)$ from a multivariate normal distribution with mean $\boldsymbol{\mu}_1 = \mathbf{0}$ and covariance matrix $\boldsymbol{\Omega} = \mathbf{S}$. Also, generate $n_2 = 25$ random vectors from $N_4(\mathbf{0}, \mathbf{S})$. Note that $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ can be set to be different, however, the statistic $F2_{(K\cup L)}$ is location-invariant, and therefore it is not unrealistic to set $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$. For the generated random numbers, compute $d^2_{K\cup L}$, $(\hat{\psi}_{K\cup L}/D)^2$, $DIF_{K\cup L}$ and $F2_{(K\cup L)}$ for $\begin{pmatrix} n \\ k+l \end{pmatrix}$ cases, and find the maximum values of $d^2_{K\cup L}$, $(\hat{\psi}_{K\cup L}/D)^2$, $DIF_{K\cup L}$ and $F2_{(K\cup L)}$. Repeat this process 100 times and get a 95-*th* percentile of 100 maximum values. The 95-*th* percentiles of our Monte Carlo study for $k+l = 1,2,3$ are listed in Tables 4.1 and 4.2. The result for $k+l = 4$ took too much computation and it is omitted. Note that these reference values are only a guideline to those observations which might be considered with special attention, and they are not a strict cutoff value to determine some observations are influential or not.

To compare the result of Atkinson's(1981) suggestion, we obtain the *envelop*(see Atkinson(1981) for details) for $k+l = 1$, which is shown in Figure 4.2. We see that cases 46 and 34 are significantly influential and this result coincides with the above Monte Carlo study.

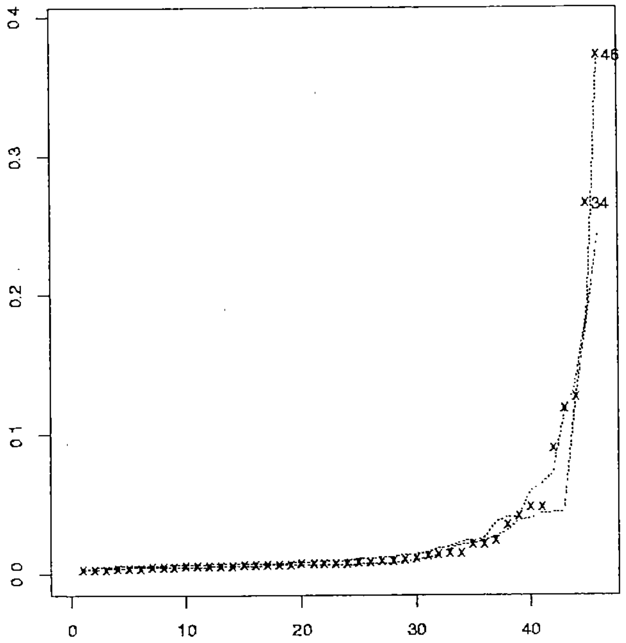

Figure 4.2. $F2_{(K\cup L)}$ and its corresponding envelop for $k+l=1$. Cases 46 and 34 are significantly influential.

5. CONCLUDING REMARKS

Detection of influential observations are very important not only in regression analysis but also in multivariate analysis. Fung(1995) suggested single case deletion diagnostic in the discriminant analysis. In this paper we extend Fung's results to multiple cases deletion diagnostic which is necessary to detect influential observations with masking effect. Through a real data set we have shown that the masking effect really exists, and those observations with the masking effect can never be detected by the single case deletion diagnostic.

Also, we consider some reference values for $d^2_{K\cup L}$, $(\hat{\psi}_{K\cup L}/D)^2$, $DIF_{K\cup L}$ and $F2_{(K\cup L)}$ which can flag observations requiring special attention. To do this, Monte Carlo distributions and their 95-*th* percentiles were derived.

Table 4.1. The five largest observations with the basic building blocks $d^2_{K\cup L}$, $(\hat{\psi}_{K\cup L}/D)^2$, and $DIF_{K\cup L}$ for $k + l = 1, 2, 3, 4$. Values in parenthesis denote Monte Carlo reference values.

$k+l$	*set*	$d^2_{K\cup L}$	*set*	$(\frac{\hat{\psi}_{K\cup L}}{D})^2$	*set*	$DIF_{K\cup L}$
	46	17.30	46	7.25	40	10.81
	1	11.72	34	4.25	46	10.05
1	40	11.45	42	3.91	2	8.73
	2	9.69	1	3.74	18	8.30
	11	8.63	16	3.22	1	7.99
		(11.56)		(3.84)		(9.53)
	1, 46	29.02	42, 46	21.81	1, 46	28.45
	40, 46	28.76	16, 46	20.12	40, 46	25.19
2	2, 46	26.99	27, 46	16.52	34, 46	25.14
	11, 46	25.67	15, 46	16.36	11, 46	24.77
	18, 46	25.25	26, 46	16.31	2, 46	24.07
		(25.53)		(16.27)		(24.52)
	1, 40, 46	40.48	16, 42, 46	41.78	1, 40, 46	40.48
	1, 2, 46	38.72	27, 42, 46	36.51	1, 2, 46	38.67
3	2, 40, 46	38.45	15, 42, 46	36.28	2, 40, 46	37.62
	1, 11, 46	37.66	26, 42, 46	36.20	11, 40, 46	37.31
	1, 18, 46	37.39	16, 27, 46	34.32	34, 40, 46	36.96
		(37.56)		(34.29)		(37.66)
	1, 2, 40, 46	50.17	16, 27, 42, 46	61.40	1, 2, 40, 46	49.12
	1, 11, 40, 46	49.11	15, 16, 42, 46	61.10	1, 18, 40, 46	48.80
4	1, 18, 40, 46	48.85	16, 26, 42, 46	61.00	1, 40, 41, 46	47.71
	1, 34, 40, 46	48.71	16, 20, 42, 46	56.34	1, 2, 18, 46	47.08
	1, 40, 41, 46	47.99	16, 24, 42, 46	55.81	2, 11, 40, 46	46.58

Table 4.2. The five largest observations with the influence measure $F2_{(K \cup L)}$ and the reference values from the Monte Carlo study for $k + l = 1, 2, 3, 4$

$k + l$	*set*	$F2_{K \cup L}$	*Monte Carlo reference value*
1	46	0.37	0.21
	34	0.27	
	16	0.13	
	40	0.12	
	20	0.09	
2	34, 46	1.25	0.83
	42, 46	1.18	
	16, 46	0.95	
	34, 41	0.72	
	15, 46	0.68	
3	31, 36, 44	2.95	2.36
	16, 42, 46	2.53	
	26, 31, 40	2.21	
	16, 34, 46	2.20	
	26, 27, 46	1.97	
4	27, 34, 42, 46	6.02	
	26, 27, 42, 46	5.18	
	16, 34, 42, 46	5.17	
	26, 34, 42, 46	5.08	
	16, 27, 42, 46	4.24	

APPENDIX : Proof of Eq. (3.1)

Let $K = \{i_1, \cdots, i_k\}$ and $L = \{j_1, \cdots, j_l\}$, then $\mathbf{w}_K = \sum_{j \in K}(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1) = \sum_{j \in K} \mathbf{w}_{1j}$ and $\mathbf{w}_L = \sum_{j \in L}(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2) = \sum_{j \in L} \mathbf{w}_{2j}$.
Now, we have

$$\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_{1(K)} = \bar{\mathbf{y}}_1 - \frac{\sum_{j=1}^{n_1} \mathbf{y}_{1j} - \sum_{j \in K} \mathbf{y}_{1j}}{n_1 - k} = \frac{\sum_{j \in K}(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)}{n_1 - k} = \frac{\mathbf{w}_K}{n_1 - k}$$

and

$$\bar{\mathbf{y}}_2 - \bar{\mathbf{y}}_{2(L)} = \bar{\mathbf{y}}_2 - \frac{\sum_{j=1}^{n_2} \mathbf{y}_{2j} - \sum_{j \in L} \mathbf{y}_{2j}}{n_2 - l} = \frac{\sum_{j \in L}(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)}{n_2 - l} = \frac{\mathbf{w}_L}{n_2 - l}.$$

Using these expressions, we have

$$\begin{aligned}
\mathbf{S}_{(K \cup L)} &= \frac{1}{n-k-l-2}[\sum_{j \notin K}(\mathbf{y}_{1j} - \bar{\mathbf{y}}_{1(K)})(\mathbf{y}_{1j} - \bar{\mathbf{y}}_{1(K)})' \\
&\quad + \sum_{j \notin L}(\mathbf{y}_{2j} - \bar{\mathbf{y}}_{2(L)})(\mathbf{y}_{2j} - \bar{\mathbf{y}}_{2(L)})'] \\
&= \frac{1}{n-k-l-2}[\sum_{j=1}^{n_1}(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1 + \frac{\mathbf{w}_K}{n_1 - k})(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1 + \frac{\mathbf{w}_K}{n_1 - k})' \\
&\quad + \sum_{j=1}^{n_2}(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2 + \frac{\mathbf{w}_L}{n_2 - l})(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2 + \frac{\mathbf{w}_L}{n_2 - l})' \\
&\quad - \sum_{j \in K}(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1 + \frac{\mathbf{w}_K}{n_1 - k})(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1 + \frac{\mathbf{w}_K}{n_1 - k})' \\
&\quad - \sum_{j \in L}(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2 + \frac{\mathbf{w}_L}{n_2 - l})(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2 + \frac{\mathbf{w}_L}{n_2 - l})'] \\
&= \frac{1}{n-k-l-2}[\sum_{j=1}^{n_1}(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)' + \sum_{j=1}^{n_2}(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)' \\
&\quad + \frac{n_1}{(n_1 - k)^2}\mathbf{w}_K\mathbf{w}_K{}' + \frac{n_2}{(n_2 - l)^2}\mathbf{w}_L\mathbf{w}_L{}' \\
&\quad - \sum_{j \in K}(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)(\mathbf{y}_{1j} - \bar{\mathbf{y}}_1)' - \frac{2n_1 - k}{(n_1 - k)^2}\mathbf{w}_K\mathbf{w}_K{}' \\
&\quad - \sum_{j \in L}(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)(\mathbf{y}_{2j} - \bar{\mathbf{y}}_2)' - \frac{2n_2 - l}{(n_2 - l)^2}\mathbf{w}_L\mathbf{w}_L{}'] \\
&= \frac{1}{n-k-l-2}[(n-2)\mathbf{S} - \frac{1}{n_1 - k}\mathbf{w}_K\mathbf{w}_K{}' - \frac{1}{n_2 - l}\mathbf{w}_L\mathbf{w}_L{}' \\
&\quad - \sum_{j \in K}\mathbf{w}_{1j}\mathbf{w}_{1j}{}' - \sum_{j \in L}\mathbf{w}_{2j}\mathbf{w}_{2j}{}'].
\end{aligned}$$

REFERENCES

Atkinson, A. C. (1981) Two graphical displays for outlying and influential observations in regression, *Biometrika*, **68**, 13-20.

Campbell, N. A.(1978) The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics*, **27**, 251-258.

Cook, R. D.(1977) Detection of influential observation in linear regression, *Technometrics*, **19**, 15-18.

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

Cook, R. D., and Wang, P. C.(1983) Transformations and influential cases in regression, *Technometrics*, **25**, 337-343.

Eubank, R. L.(1985) Diagnostics for smoothing splines, *Journal of the Royal Statistical Society*, Ser. B, **47**, 332-341.

Fung, W. K.(1992) Some diagnostic measures in discriminant analysis, *Statistics and Probability Letters*, **13**, 279-285.

Fung, W. K.(1995) Diagnostics in linear discriminant analysis, *Journal of the American Statistical Association*, **90**, 952-956.

Hinkley, D. V., and Wang, S.(1988) More about transformations and influential cases in regression, *Technometrics*, **30**, 435-440.

Johnson, W.(1987) The detection of influential observations for allocation, separation, and the determination of probabilities in a Bayesian framework, *Journal of Business and Economic Statistics*, **5**, 369-381.

Johnson, R. A. and Wichern, D. W.(1987) *Applied Multivariate Statistical Analysis*, Prentice Hall.

Kim, C. (1996) Cook's distance in spline smoothing. *Statistics and Probability Letters*, **31** 139-144.

Kim, C.(1996) Local influence and replacement measure, *Communications in Statistics - Theory and Methods*, **25**, 49-61.

Kim, C., and Storer, B. E.(1996) Reference values for Cook's distance, *Communications in Statistics - Simulation and Computation*, **25**, 691-709.

Kim, C., Storer, B. E., and Jeong, M.(1996) A note on Box-Cox transformation diagnostics, *Technometrics*, **38**,178-180.

Kim, C., and Hwang, S. Y.(2000) Influential subsets on the variable selection, *Communications in Statistics - Theory and Methods*, To appear.

Pregibon, D.(1981), Logistic regression diagnostics, *The Annals of Statistics*, **9**, 705-724.

Silverman, B. W.(1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion), *Journal of the Royal Statistical Society*, Ser. B, **47**, 1-52.

Tsai, C. L., and Wu, X.(1990) Diagnostics in transformation and weighted regression, *Technometrics*, **32**, 3155-322.

Thomas, W.(1991) Influence diagnostics for the cross-validated smoothing parameter in spline smoothing, *Journal of the American Statistical Association*, **86**, 693-698.