Discriminant Diagnostics

PETER A. LACHENBRUCH
AND YONG-XIAO WANG
UCLA SCHOOL OF PUBLIC HEALTH

Abstract: Diagnostic methods for discriminant analysis are discussed. The equivalence with linear regression is noted and diagnostics used in that situation are considered. It is shown that both the leverage and Cook's distance are functions of the linear discriminant function and the distance of the observation from the group mean. Some examples are given.

Key Words: *Mahalanobis distance, Cook's distance, Leverage*

Introduction: An important component of a regression analysis is the investigation of data points which may distort the analysis because of high influence or unusual values of the dependent or independent variables. This investigation may result in removing the unusual observation from the analysis or in giving it reduced weight. Similar concerns exist in two group Discriminant Analysis: because an observation may be unusual, the discriminant coefficients may be affected and the analysis rendered suspect. We wish to have statistics which detect "important" observations affecting the discriminant coefficients. In regression, the values of the independent variables are (in theory, at least) under the control of the investigator. In discriminant analysis, all observations other than the group identifier are presumed to be realizations of random variables.

It is well known that using group indicators as the dependent variable in a multiple regression leads to a set of coefficients which are proportional to the linear discriminant function. A natural way to develop discriminant diagnostics is to consider the usual regression diagnostics applied to this multiple regression. This has some disadvantages: the diagnostics are not directly interpretable in terms of the discriminant function, and for two major diagnostics, no information is obtained. The plot of residuals against estimated values will plot $y - b'x$ versus $b'x$. In this case y is 1 or 0, so the plot will always be two straight lines with slope -1. This provides no information about possible outlying points. Similarly, studentized residuals have almost the same flaw: very little information is obtained as the plot generally is close to the two straight lines with slope -1.

The leverage and Cook's distance statistics from the regression model can be shown to be easily interpretable as functions of the distance of the candidate point from its group mean and (to a minor extent) its discriminant score. The algebra is almost identical for the two statistics.

Campbell (1978) derived the influence function for general discriminant functions and Mahalanobis distances. He found that the influence was a (messy) function of the discriminant score,and the Mahalanobis distance between groups. Some of his work is similar to the results given below.

We shall assume we are performing a two group discriminant analysis where the groups have a common covariance matrix. The usual LDF arises if we assume multivariate normality of the observations, $\mathbf{x}$, or if we find the linear function which maximizes the ratio of the between groups sum of squares relative to the within group sum of squares (see, e.g. Lachenbruch, 1975). The normality assumption can be used when we consider diagnostic plots. In the following, we will let

$$D_S(\mathbf{X}) = (\mathbf{X} - 1/2(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2))'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

be the sample discriminant function, where $\mathbf{S}$ is the sample pooled covariance matrix, and $\bar{\mathbf{x}}_i$ is the mean in the i^{th} population. Also, $D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ is Mahalanobis distance.

A simple statistic to consider is the LDF itself, standardized within group. $(D_S(\mathbf{X}) - D_S(\mu_i))/\sqrt{\text{var}}(D_S(\mathbf{X}))$ is easily seen to be $(\mathbf{X}-\mu_i)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)$ multiplied by a constant. This quantity will be approximately normally distributed if the observations are. The population mean can be estimated by the appropriate sample mean. Normal plots, and general outlier tests can be applied to this statistic. It also arises in consideration of the leverage and Cook's distance statistics.

There are several useful identities which we use in this article. We shall derive the statistics for observations from the first population. All results obviously carry over to the second. The i^{th} deleted mean is defined as the mean of the observations deleting the i^{th} observation and is denoted as $\bar{\mathbf{x}}_{1(i)}$. The first identity gives $\bar{\mathbf{x}}_1 = \bar{\mathbf{x}}_{1(i)} + \mathbf{d}_{1(i)}/n_1$, where $\mathbf{d}_{1(i)} = \mathbf{x}_{1(i)} - \bar{\mathbf{x}}_{1(i)}$ and the parentheses indicate the mean computed excluding the i^{th} observation. Second, if we can write a matrix as $\mathbf{A} + \mathbf{uv}'$ where $\mathbf{u}$ and $\mathbf{v}$ are column vectors, then $(\mathbf{A} + \mathbf{uv}')^{-1} = \mathbf{A}^{-1} - (\mathbf{A}^{-1}\mathbf{uv}'\mathbf{A}^{-1})/(1+\mathbf{u}'\mathbf{A}^{-1}\mathbf{v}$. Each of these can be verified by some simple algebra. The second identity is useful when computing statistics excluding the i^{th} observation.

2. Leverage and Cook's Distance. In the regression model, we assume that the observations are centered about the pooled mean $\bar{\mathbf{x}} = (n_1\bar{\mathbf{x}}_1 - n_2\bar{\mathbf{x}}_2)/(n_1 + n_2)$; that is the observations are of the form $\mathbf{d}_i = \mathbf{x}_i - \bar{\mathbf{x}} = \mathbf{x}_i - \bar{\mathbf{x}}_1 + c_3(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \delta + c_3(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ where $c_3 = n_2/(n_1 + n_2)$. The dependent variable, y, has values 1 in group 1 and 0 in group 2. The regres-

sion coefficients are $(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$. It is a simple matter to show that

$$\begin{aligned}\mathbf{V} &= \mathbf{x}'\mathbf{x} \\ &= (n_1+n_2-2)\mathbf{S} + (\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'n_1n_2/(n_1n_2) \\ &= c_1\mathbf{S} + c_2(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)'\end{aligned}$$

To use the identity from section 1, we let $\mathbf{A} = c_1\mathbf{S}$ and $\mathbf{u} = \mathbf{v} = (\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\sqrt{c_2}$. The leverage for the i^{th} point in group 1 is $\mathbf{d}_{1i}'\mathbf{V}^{-1}\mathbf{d}_{1i}$ which is (dropping the i subscripts)

$$\begin{aligned}h_i &= (\delta + c_3(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2))'\mathbf{V}^{-1}(\delta + c_3(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)) \\ &= 1/c_1\{\delta'\mathbf{S}^{-1}\delta - c_2/c_1(\delta'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1-\mathbf{x}_2))^2 \\ &\quad /1 + c_2/c_1D^2 + 2c_3\delta'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ &\quad - 2c_2c_3/c_1\delta'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)D^2/1+c_2/c_1D^2) \\ &\quad + c_3^2D^2 - (D^2)^2c_3^2c_2/c_1/(1+c_2/c_1D^2)\} \\ &= \{\Phi + 2c_3\gamma + c_3^2D^2 - (c_2/c_1/(1+c_2/c_1D^2)) \\ &\quad (\gamma + c_3D^2)^2\}/c_1\end{aligned}$$

where $\Phi = \delta'\mathbf{S}^{-1}\delta$, and $\gamma = \delta'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) = D_S(\mathbf{x}_{1i}) - D^2/2$. Note that γ is the sample estimate of the numerator of the within-group standardized discriminant statistic mentioned in section 1. Since D^2 is constant for any problem, the only quantities that are a function of the i^{th} observation are Φ and γ. As Φ increases the leverage increases and by taking the derivative of h with respect to γ it is easy to show that while $\gamma < c_3c_1/c_2(= (n_1+n_2-2)/n_1)$ the leverage increases. If γ is negative, the influence will be substantial. Thus, observations from group 1 which are "far" from their group centroid, or like group 2 observations along the discriminant line (i.e. when the discriminant function of the observation is very large the influence will decline – similarly for the second group, when the LDF of the observation is very small its influence will decline). The leverage is an increasing function of γ for most observations inpractically interesting regions of the discriminant line.

For the LDF, Cook's distance is the difference between coefficients with the j^{th} point included and with it excluded, normalized by a suitable metric. Thus, Cook's distance may be chosen to be proportional to $CD = \hat{\beta}_{(j)} - \hat{\beta})'\mathbf{S}(\hat{\beta}_{(j)} - \hat{\beta})$ where $\hat{\beta} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ and the subscript (j) refers to the estimate without the j^{th} observation. Using the identities given in section 1, we have

$$\begin{aligned}\hat{\beta}_{1(j)} &= (k_1\mathbf{S}^{-1} - k_1^2k_2\mathbf{S}^{-1}\mathbf{d}\mathbf{d}'\mathbf{S}^{-1}/(1+k_1k_2\mathbf{d}'\mathbf{S}^{-1}\mathbf{d})) \\ &\quad ((\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) - \mathbf{d}/n_1 - 1) \\ &= \beta(1 - 1/(n_1+n_2-3) - \mathbf{S}^{-1}\mathbf{d}\{1/(n_1-1) \\ &\quad + k_1^2k_2(\mathbf{d}'\hat{\beta} - \Phi)/(1+k_1k_2\Phi)\}\end{aligned}$$

where $k_1 = (n_1+n_2-2)/(n_1+n_2-3)$ and $k_2 = (n_1+1)/(n_1(n_1+n_2-3))$. With a bit of algebra we can show that the Cook's distance is

$$CD = \hat{\beta}'\mathbf{S}^{-1}\hat{\beta}/n_1+n_2-3)^2 + F^2\Phi + 2*F\gamma/(n_1+n_2-3)$$

where F is the multiplier of $\mathbf{S}^{-1}\mathbf{d}$ and Φ and γ have the same meaning as before. Thus, Cook's distance is a monotonically increasing function of Φ and γ.

Since $\mathbf{x} - \bar{\mathbf{x}}_1$ is $MVN(0, \Sigma(n_1-1)/n_1)$ it is immediate that $n_1/(n_1-1)\Phi$ is a T^2 variable. For large values of n_1+n_2 this will be approximately X^2 with degrees of freedom equal to the number of variables in the LDF. This suggests that a X^2 plot will be helpful in examining this diagnostic. Large values of Φ will be at the positive extreme of the X^2 distribution. For 2 variables the 99th percentile is 9.21, and for a 4 variable problem, it is 13.28. Similarly, γ is a shifted discriminant function and will be approximately normal with mean 0 and variance about D^2 (more exact statements could be made, but are probably not too useful for graphical procedures). Thus, if $|\gamma|$ is larger than 2.33D we should consider examining the point further. Alternatively, one could standardize γ by dividing it by D.

3. Examples. First, let us examine the behavior of Φ and γ for some values of an observation, X, from group 1. We consider the observations at $\mathbf{x}_1$, $\mathbf{x}_2$ and their mean. For these points, we have

	VALUE OF X		
Statistic	$\bar{\mathbf{x}}_1$	$\bar{\mathbf{x}}_2$	$(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2)/2$
Φ	0	D^2	$D^2/4$
γ	0	$-D^2$	$-D^2/2$

Thus, if the point is near the mean of the alternative group, both Φ and γ will be relatively large. A point near the group mean will have smaller Φ and γ. A value of $D^2/4$ must be evaluated in light of the number of variables in the discriminant function. Thus, if $D^2 = 25$ and there are 2 variables, both Φ and γ would be quite suspicious for a point at the mean. If there were 10 variables, neither one would be cause for concern. A point orthogonal to the plane $\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ will be more sensitive to Φ than to γ. That is, Φ will detect generally bad points, while γ will be sensitive to outliers on the discriminant line.

We use a data set from Afifi and Azen (1972) on shock. We use the variables Systolic Pressure (SP), Cardiac Index (CI), and Appearance Time (APTIME). These variables show substantial differences in the two groups on Systolic Pressure but not on the Cardiac Index or Appearance Time. The two groups are those who Survived and those who Died. The authors give initial and final measurements; we use only the initial measurement. For these data, the means and within group covariance matrix are given in Table 1.

TABLE 1
STATISTICS FOR SHOCK DATA

Means	SP	CI	APTIME
Survived (70)	114.6	27.1	9.1
Died (43)	92.3	23.4	9.6

COVARIANCE MATRIX

SP	830.6		
CI	39.8	215.6	
APTIME	-6.7	-27.7	31.3

Figure 1 gives the scatter plot matrix of these data with 95% ellipsoids of concentration superimposed. One sees that there are several suspicious points which seem to lie outside the ellipsoids for the survivors. These outliers seem to have large values of CI or a small value of SP. Figure 2a, 2b and 2c are Normal plots of these three variables with the appropriate line indicated in the plot. On all three, the slopes appear to be almost identical, suggesting the validity of the common variance assumption (this says nothing about equal correlations however). On both CI and APTIME plots there is a suspicious dropping off at the low end in both groups. However, a Kolmogorov)Smirnov test with the Lilliefors adjustment does not indicate a significant difference from normality. The tests are close to signficance at the 0.05 level.

The discriminant function on the full data set is

$$D^S(\mathbf{X}) = 0.02629SP + 0.01207CI + 0.00120\,\text{APTIME} - 3.03576.$$

With this data, we allocate 46 of the 70 survivors correctly and 27 of the 43 deaths correctly. The discriminant scores range from -2.3 to 2.7 in the survivor group and from -2.6 to 1.9 in the death group, with a considerable overlap. The D^2 is 0.6297 (computed from the $F = 5.49$ with 3 and 109 d.f.). We used the 99th percentile of X^2 as a criterion for noting suspect observations. For this data the X^2 (3,0.99) = 11.3. There are three individuals in the survivor group with large values of Φ: case 33, $\Phi = 12.2$, case 47, Φ, and case 86, $\Phi = 18.3$; there are no individuals in the dead group with large values of Φ. The maximum of Φ in that group is 9.1. These suggest some potential extreme values in the data. In group 1, since $\gamma = D_S(x) - D^2/2$ the values of that will concern us are those greater than 2.18 or less than -1.49 (using the 99th percentile) while the negatives of these values may be used in group 2 (This uses the plug-in estimate of the standard deviation of γ, D = 0.79). There were no values of γ which indicated problems in either group. Note also that we can find $D_S(\mathbf{X})$ by subtracting the values of Φ in each group and dividing by 2.

The values of the variables of the cases were:

Case 33: SP = 153, CI = 75.8, APTIME = 4.7, $\Phi_1 = 12.2$, $\Phi_2 = 16.0$
Case 47: SP = 63, CI = 9.2, APTIME = 26.1, $\Phi_1 = 12.0$, $\Phi_2 = 9.5$
Case 86: SP = 82, CI = 76.3, APTIME = 14.5, $\Phi_1 = 18.3$, $\Phi_2 = 18.4$.

For these cases, it is clear that they are quite unlike the Survivors (indeed they were also quite unlike the Deaths). The difference from the group 1 mean values is large for the CI variable for case 33 and case 86, and large for the APTIME variable for case 47. For SP each case is at least one standard deviation away. These cases are so atypical that they might best be deleted from the data set. Figure 3 gives the Chi squared plots for 3 d.f. for groups 1 and 2. In group 1, there are three points at the upper tail which appear to he too large. These, of course, correspond to the large values of Φ.

In the plot for group 2, one point appears visually to be too large. Note, however, the scales of these diagrams are different.

We modified the data by removing the offending cases. This changed the means and covariance matrix to

TABLE 2
MEANS AND COVARIANCES
WITH MODIFIED DATA SET

Means	SP	CI	APTIME
Survivors	115.3	25.9	9.6
Died (43)	92.3	23.4	9.6

COVARIANCE MATRIX

SP	805.2		
CI	30.5	173.3	
APTIME	4.66	-26.4	29.0

and the D^2 value was 0.6862, about a 9% increase. For these modified data, there were no extreme values of Φ. The discriminant function became

$$D_S(\mathbf{X}) = 0.02850SP + 0.00509CI - 0.02549\,\text{APTIME} - 2.85112.$$

The major changes in the function are the change in sign and mag-nitude of the coefficient of APTIME, and the change in the constant. Since CI and APTIME are similar in the two groups, the change in the value of the APTIME coefficient is of little importance (one probably should not even use them in the dis-criminant function).

4. Discussion. The question of the appropriate "alarm" level for these statistics naturally arises. Since their distributions are approximately known, rough critical values may be obtained from the X^2 distribution for Φ, and for the normal distribution for γ. Since mul-

tiple tests are likely to be done since we don't know which observations are suspect a priori, using conservative critical values is appropriate. In our example, we used the 0.99 point, but would not argue with a stricter criterion for rejection. We deleted the three large values of Φ in our data. This is quite arbitrary and one could also argue that one should shrink the data along the ray from the group mean. This was suggested by Broffitt, Clarke, and Lachenbruch (1980). In that paper, good results were obtained in a scale contamination problem by forming the pseudo-observation along that ray with length $\min(\Phi, K)$ where K was a suitably chosen percentile of D^2.

The Φ statistic seems to be useful in selecting generally wild observations, hardly a surprising result. The γ statistic seems to be highly correlated with Φ (in the SHOCK data, the correlation was about 0.5 in each of the two groups). Because γ will be normally distributed if $\mathbf{X}$ is, it will be useful in detecting deviations from normality. In some studies (made by replacing a single observation in the data set by bad data) it was clear that both Φ and γ were sensitive to such single outliers. Lack of normality seems to be easily detected using γ.

Unequal covariance matrices may not be detected with these statistics. There are general multivariate statistics for this purpose (see e.g. Anderson 1984); unfortunately they tend to be affected by violations of the multivariate normality assumption.

In 1963, Wilks considered tests for multivariate outliers based on the likelihood ratio statistic. The Mahalanobis distance statistic is equivalent to this.

The computation of these statistics is a straightforward procedure in several statistical packages. BMDP (1988) gives the (squared) Mahalanobis distance, Φ, of each observation from both group means as part of the standard output. Thus, it is easy to examine the data for outliers. The SYSTAT (1988) program provides a method of saving output in the MGLH module which gives the square root of Φ from each group as part of its output. This data can be examined as is, or can be squared for plotting purposes. SAS (1985) does not display the Mahalanobis distances as part of standard output. However, it is an easy matter to write a SAS IML program which prints this data. We append a copy of such a program.

REFERENCES:

Afifi, A. and Azen, S. (1972) "Statistical Analysis: A Computer Oriented Approach." New York: Academic Press (Dataset is on pages 17-22)

Anderson, T. W. (1984) "An Introduction to Multivariate Statistical Analysis, 2nd Edition." New York: John Wiley and Sons. (Chapter 10)

Broffitt, B., Clarke, W. R., and Lachenbruch, P. A. (1980) "The Effect of Huberizing and Trimming on the Quadratic Discriminant Function." *Commun. Statist. Theor. Meth*, **A9**(1):13-25.

Campbell, N. (1978) "The Influence Function as an Aid in Outlier Detection in Discriminant Analysis." *Applied Statistics*, **27**(3):251-258.

Dixon, W. J., Chief Editor (1988) "BMDP Statistical Software Manual." Berkeley: University of California Press (pages 337-356).

SAS Institute, Inc. (1985) "SAS User's Guide: Statistics, Version 5 Edition."

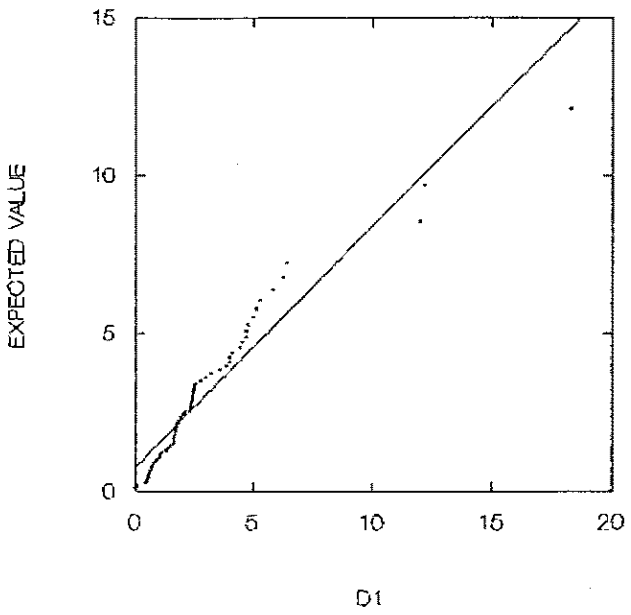
Wilkinson, L. (1988) "SYSTAT: The System for Statistics." Evanston, IL: SYSTAT, Inc. (MGLH program).

Wilks, S. S. (1963) "Multivariate Statistical Outliers." *Sankhya: The Indian Journal of Statistics: Series A.*, **25**:407-426.

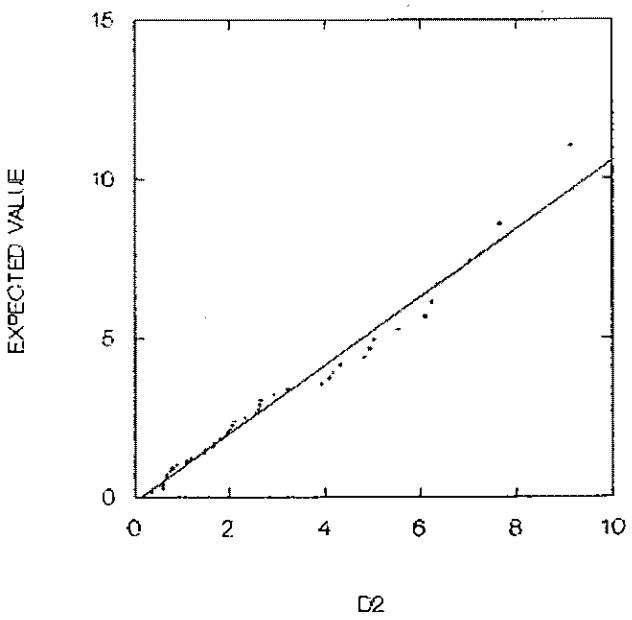
APPENDIX
SAS PROC MATRIX CODE FOR DISCRIMINANT DIAGNOSTICS

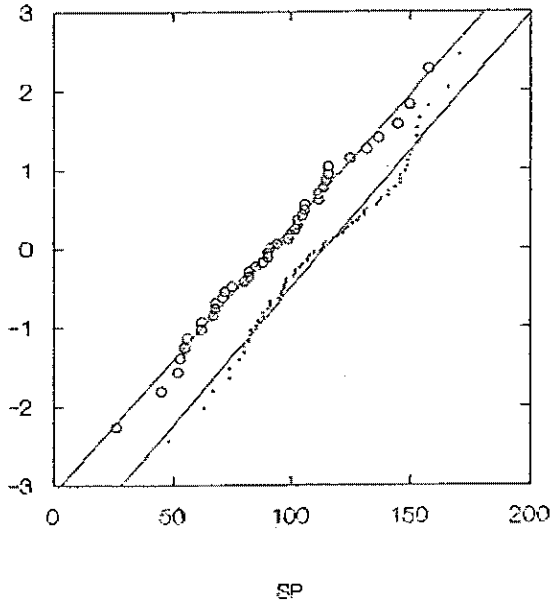
```
*     DATA SET B1 CONTAINS THE DATA FROM
THE SURVIVORS
*     DATA SET B2 CONTAINS THE DATA FROM
      THE DEATHS
*     DATA SET C1 CONTAINS THE COVARIANCE
      MATRIX AND MEANS FOR THE SURVIVORS
*     DATA SET C2 CONTAINS THE COVARIANCE
      MATRIX AND MEANS FOR THE DEATHS
PROC MATRIX;
      FETCH B1 DATA B1;
      FETCH B2 DATA B2;
      FETCH C1 DATA C1;
      FETCH C2 DATA C2;
      BB=B1//B2;      * FOR THE OUTPUT ;
      N1=C1(6,1);     * NUMBER OF
OBSERVATIONS FROM THE SURVIVORS;
      N2=C2(6,1);     * NUMBER OF
OBSERVATIONS FROM THE DEATHS;
*     COMPUTE THE POOLED COVARIANCE MATRIX
      ;
      S=((N1-1)#C1(1 2 3,) + (N2-1)#C2(1 2
3,))#/(N1+N2-2);
      INS=INV(S);     * INVERSE OF S;
      X1=B1(,2 3 4); * OBSERVATIONS FROM
GROUP 1;
      X2=B2(,2 3 4); * OBSERVATIONS FROM
GROUP 2;
*     COMPUTE DIFFERENCE OF OBSERVATIONS
      FROM FIRST GROUP MEAN;
      DX1=X1-J(N1,1)*C1(4,);
*     COMPUTE DIFFERENCE OF OBSERVATIONS
      FROM SECOND GROUP MEAN;
      DX2=X2-J(N2,1)*C2(4,);
      DX=DX1//DX2;    * CONCATENATE THESE
DIFFERENCES;
      CR=J(N1+N2,1)*(C1(4,)-C2(4,));      *
      VECTOR OF DIFFERENCE OF MEANS;
      NN=DX*INS*CR;
      FFI=DX*INS*DX;
      D1=CR*INS*CR;
      GAMMA=DIAG(NN)*J(N1+N2,1);     *
      DISCRIMINANT COMPONENT;
*     COMPUTE MAHALANOBIS DISTANCE
      SQUARED;
      DD=(DIAG(D1))##0.5*J(N1+N2,1);
      RR=GAMMA#/DD;   *  STANDARDIZED
      GAMMA;
*     GET MAHALANOBIS DISTANCE COMPONENT
      OF h;
      FI=DIAG(FFI)*J(N1+N2,1);
      INFOR=FI||GAMMA||RR;      *  FIRST
      COLUMN IS MAHALANOBIS DISTANCE,
      SECOND COLUMN IS ALIGNED
      DISCRIMINANT, THIRD COLUMN IS
      STANDARDIZED ALIGNED DISCRIMINANT;
      C='PHI' 'ADF'  'ADF/D2';
      PRINT INFOR COLNAME=C;
```

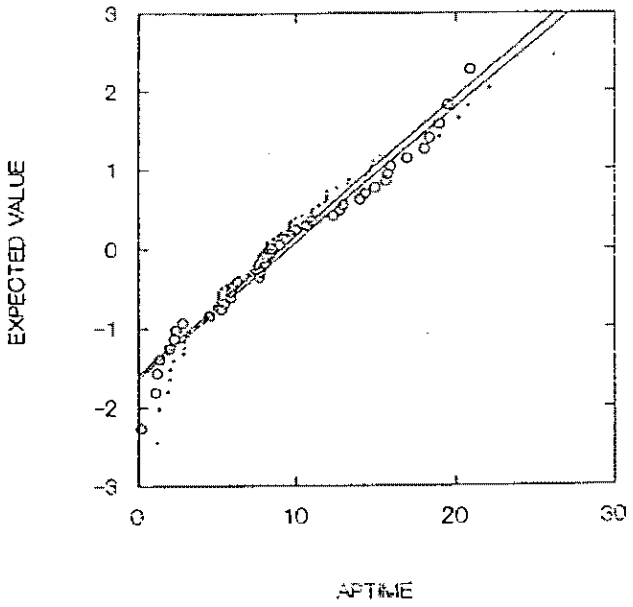
CHISQ(3) PLOT - GROUP=1

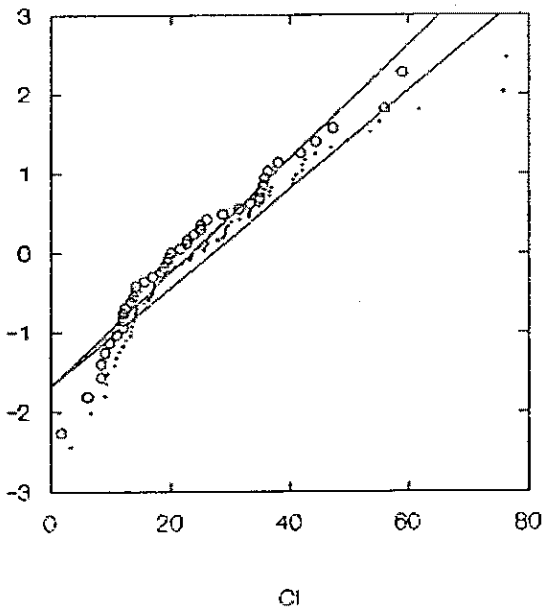


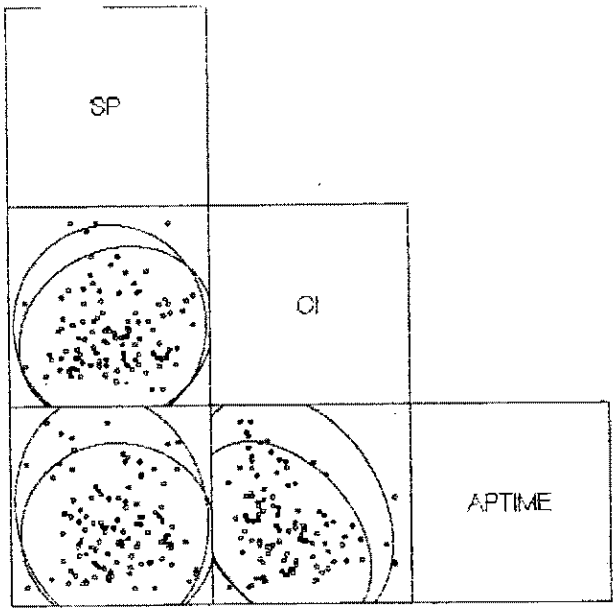
CHISQ(3) PLOT - GROUP=2




3
2
1
0
-1
-2
-3
0
50
100
150
200
SP


EXPECTED VALUE
3
2
1
0
-1
-2
-3
0
10
20
30
APTIME


3
2
1
0
-1
-2
-3
0
20
40
60
80
CI


SP
CI
APTIME