Discriminant Analysis

The classification problem consists in classifying an unknown individual w into one of k populations $W_1, W_2, W_3 \ldots W_k$ on the basis of measurements $x_1, \ldots x_p$ on p characteristics. The following two examples illustrate the nature of this problem.

e.g. (1) The admissions committee of a college wishes to classify an applicant into W_1 = the population of students who successfully complete college, or W_2 = the population of students who do not complete college. Its decision is based on the applicant's scores on p entrance examinations.

e.g. (2) A physician wishes to diagnose a patient as having one of k diseases on the basis of the existence or absence of p symptoms.

Discussion of computer printouts.

9.2

Discriminant Analysis for Polit's book Page 260.

```
-> get file='d:\stat601.14\polit\p260.sav'.

-> discriminant groups=finish(0,1)/
->         variables=motiv,gpa_g/
->         analysis=motiv,gpa_g/
->         method=direct/statistics=mean,raw,table/plot cases.
```

- - - - - - - - D I S C R I M I N A N T A N A L Y S I S - - - - - - - -

On groups defined by FINISH

20 (Unweighted) cases were processed.
0 of these were excluded from the analysis.
20 (Unweighted) cases will be used in the analysis.

Number of cases by group

	Number of cases		
FINISH	Unweighted	Weighted	Label
0	8	8.0	no
1	12	12.0	yes
Total	20	20.0	

Group means

FINISH	MOTIV	GPA_G
0	61.25000	2.82500
1	75.41667	3.46667
Total	69.75000	3.21000

- - - - - - - - D I S C R I M I N A N T A N A L Y S I S - - - - - - - -

On groups defined by FINISH

Analysis number 1

Direct method: all variables passing the tolerance test are entered.

Minimum tolerance level..... .00100

9.3

Canonical Discriminant Functions

Maximum number of functions..... 1
Minimum cumulative percent of variance... 100.00
Maximum significance of Wilks' Lambda.... 1.0000

Prior probability for each group is .50000

Canonical Discriminant Functions

Fcn	Eigenvalue	Pct of Variance	Cum Pct	Canonical Corr	After Fcn	Wilks' Lambda	Chi-square	df	Sig
					: 0	.410063	15.155	2	.0005
1*	1.4386	100.00	100.00	.7681	:				

* Marks the 1 canonical discriminant functions remaining in the analysis.

Standardized canonical discriminant function coefficients

	Func 1
MOTIV	.54014
GPA_G	.60439

$DF = 0.540\,Z_{MOTIV} + 0.604\,Z_{GPA\_G}$

	μ_1	GPA_G
0	61.25	2.85
1	75.41	3.64

Structure matrix:

Pooled within-groups correlations between discriminating variables and canonical discriminant functions
(Variables ordered by size of correlation within function)

	Func 1
GPA_G	.88812
MOTIV	.85761

Unstandardized canonical discriminant function coefficients

	Func 1
MOTIV	.0759485
GPA_G	1.9430075
(Constant)	-11.5344617

$DF = -11.534 + 0.0759\,MOTIV + 1.943\,GPA\_G$

$-11.534 + 0.0759 \times 61.25 + 1.943 \times 2.825 = -1.39362$ for GP=0

9.4

3

Canonical discriminant functions evaluated at group means (group centroids)

Group	Func 1
0	-1.39362
1	.92908

For D=1, MOTIV = 75, GPA-G = 3.6

D Score = $-11.534 + 0.0759 \times 75 + 1.943 \times 3.1 = 1.1533$

$-11.534 + 0.0759 \times 75.42 + 1.943 \times 3.4667 = 0.92$

Case Number	Mis Val	Sel	Actual Group	Highest Group	P(D/G)	P(G/D)	2nd Highest Group	P(G/D)	Discrim Scores
1			1	1	.8201	.9618	0	.0382	1.1565
2		**	1	0	.3151	.5900	1	.4100	-.3890
3			1	1	.1165	.9982	0	.0018	2.4989
4			0	0	.8859	.9539	1	.0461	-1.5371
5			1	1	.1185	.9982	0	.0018	2.4900
6			0	0	.5946	.9808	1	.0192	-1.9257
7			1	1	.8720	.9108	0	.0892	.7679
8			0	0	.4730	.9874	1	.0126	-2.1112
9		**	0	1	.7222	.8667	0	.1333	.5736
10			1	1	.4676	.7331	0	.2669	.2027
11			1	1	.5437	.9838	0	.0162	1.5362
12		**	1	0	.3194	.5950	1	.4050	-.3979
13			0	0	.8064	.8936	1	.1064	-1.1485
14			1	1	.4676	.7331	0	.2669	.2027
15			1	1	.3194	.9934	0	.0066	1.9248
16			1	1	.5825	.8054	0	.1946	.3793
17			0	0	.3572	.9921	1	.0079	-2.3143
18			0	0	.4730	.9874	1	.0126	-2.1112
19			0	0	.4127	.6889	1	.3111	-.5745
20			1	1	.8789	.9124	0	.0876	.7768

$= \frac{-1.39+0.93}{2} = -0.23$

> 3, GP=1

< 3, GP=0

Classification results -

Actual Group		No. of Cases	Predicted Group Membership 0	1
Group no	0	8	7 87.5%	1 12.5%
Group yes	1	12	2 16.7%	10 83.3%

Percent of "grouped" cases correctly classified: 85.00% $= \frac{7+10}{20} = 85\%$

Classification processing summary

20 (Unweighted) cases were processed.
0 cases were excluded for missing or out-of-range group codes.
0 cases had at least one missing discriminating variable.
20 (Unweighted) cases were used for printed output.

Factor Analysis

Factor analysis is a statistical technique used to identify a relatively small number of factors that can be used to represent relationships among sets of many interrelated variables. For example, variables such as scores on a battery of aptitude tests may be expressed as a linear combination of factors that represent verbal skills, mathematical aptitude, and perceptual speed. Variables such as consumer ratings of products in a survey can be expressed as a function of factors such as product quality and utility. Factor analysis helps identify these underlying, not directly observable, constructs.

See computer printout for discussions.

Reliability Analysis:

From the moment we're born the world begins to "score us. One minute after birth, we're rated on the 10-point Apgar scale, followed closely by the five-minute Apgar score, and then on to countless other scales that will track our intelligence, creditworthiness & so on.

When we want to measure characteristics such as driving ability, mastery of course materials, or the ability to function independently, we must construct some type of measurement device. Usually we develop a scale or test that is composed of a variety of related items. The responses to each of the items can be graded and summed, resulting in a score for each case. A question that frequently arises is, How good is our scale?

One of the most commonly used reliability coefficients is Cronbach's alpha. Alpha (or α) is based on the "internal consistency" of a test. Like correlation coefficients, reliability coefficients can range between -1.00 and +1.00, but they are almost always positive. The higher the coefficient, the greater the reliability of the instrument. Reliability coefficients generally should be at least 0.70 and in some contexts may need to be even higher to be considered acceptable.

See computer printouts for discussion.

```
COXREG
  time  /STATUS=status(1)
  /PATTERN BY ln_yesno
  /CONTRAST (ln_yesno)=Indicator(1)
  /METHOD=ENTER ln_yesno
  /PLOT SURVIVAL
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .
```

9.7

Cox Regression-survival analysis for Breast Cancer data

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	72	6.0%
	Censored	1135	94.0%
	Total	1207	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with non-positive time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		1207	100.0%

a. Dependent Variable: TIME Time (months)

Categorical Variable Codings[a,b]

		Freqency	(1)
LN_YESNO	0=No	929	.000
	1=Yes	278	1.000

a. Indicator Parameter Coding

b. Category variable: LN_YESNO (Lymph Nodes?)

Block 0: Beginning Block

Omnibus Tests of Model Coefficients

-2 Log Likelihood
917.091

Block 1: Method = Enter

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step		
	Chi-square	df	Sig.	Chi-square	df	Sig.
903.359	15.986	1	.000	13.732	1	.000

9.8

Omnibus Tests of Model Coefficients[a,b]

Change From Previous Block		
Chi-square	df	Sig.
13.732	1	.000

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 917.091

b. Beginning Block Number 1. Method = Enter

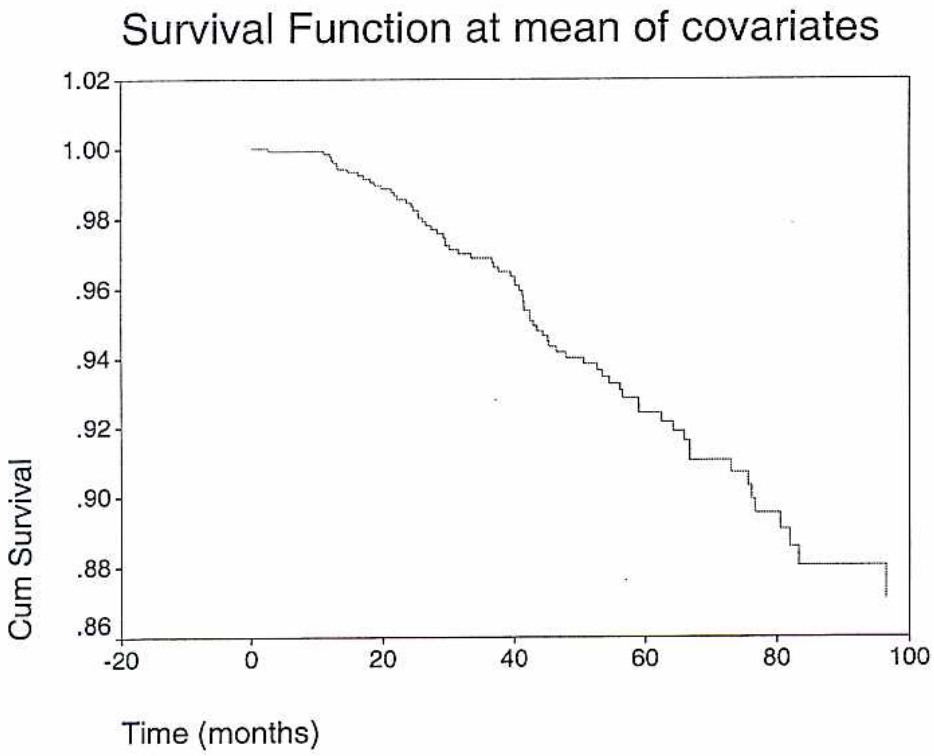
Variables in the Equation

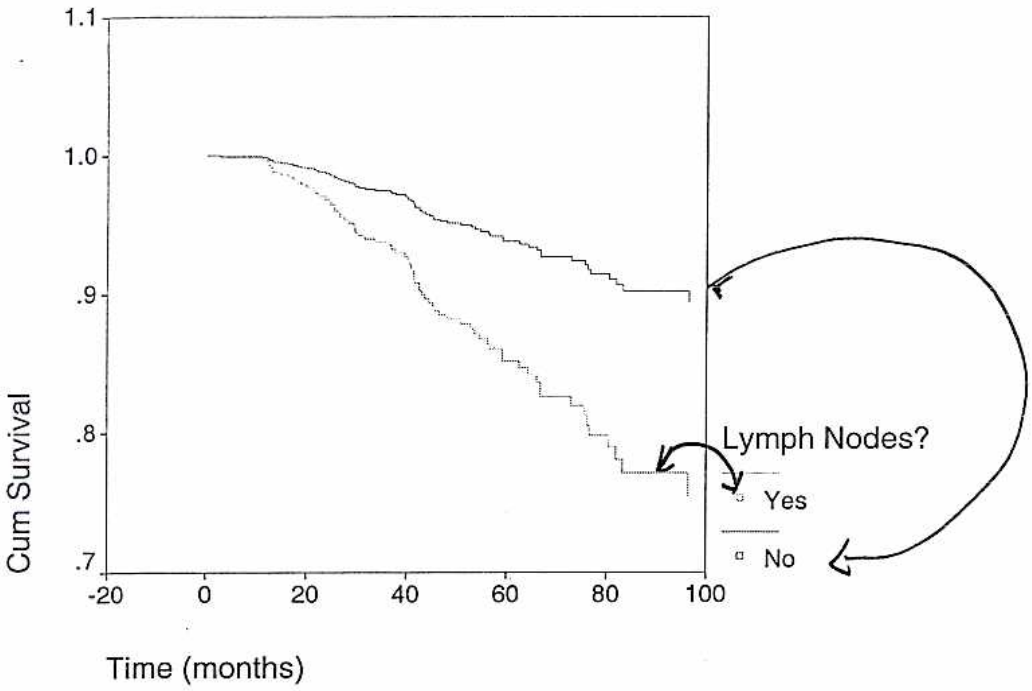
	B	SE	Wald	df	Sig.	Exp(B)
LN_YESNO	.923	.239	14.899	1	.000	2.517

← Hazard function

Covariate Means and Pattern Values

	Mean	Pattern	
		1	2
LN_YESNO	.230	.000	1.000




Survival Function for patterns 1 - 2
1.1
1.0
.9
.8
.7
Cum Survival
-20
0
20
40
60
80
100
Time (months)
Lymph Nodes?
Yes
No