DISCRIMINANT ANALYSIS

- A goal of one's research may be to classify a case into one of two or more groups.

- Two methods can be used to perform this task:

1. Logistic regression
    a. Typically used to classify a case into one of two outcome groups.

    b. Often used in medical or epidemiological studies when you want to determine which characters (parameters) are predictive of a response.

    c. Can be used when the multivariate normal assumption is violated or not justified.

2. Discriminant analysis
    a. Suited for classifying a case into one of two or more outcome groups based on a set of specific characteristics or measurements.

    b. Also can be used to determine which characters work best or are best suited for classifying a case or item.

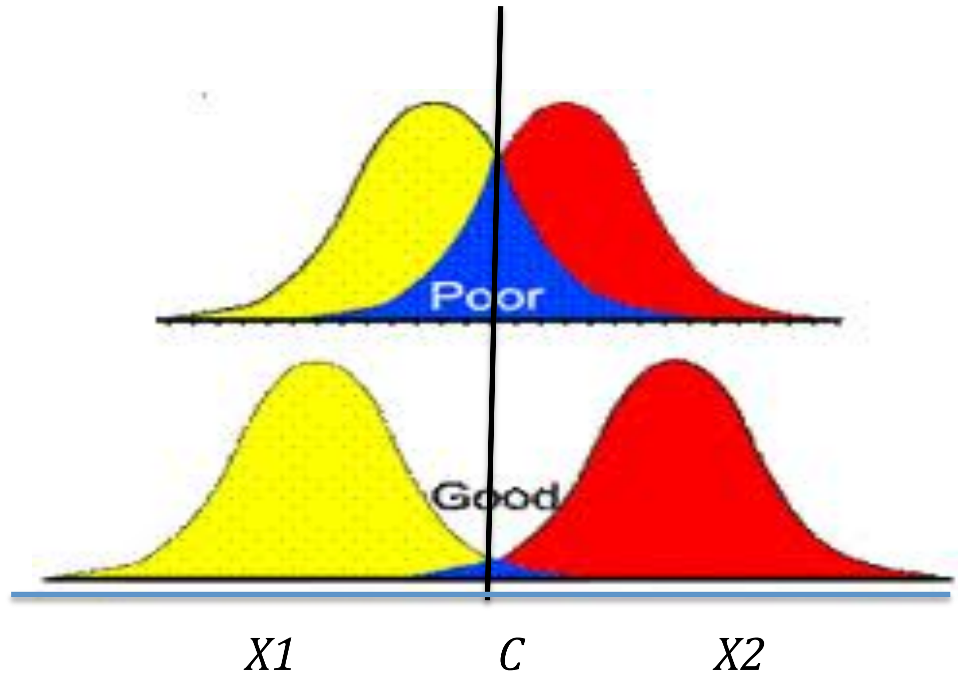
    c. An example would be identifying a new plant that you don't know anything about. Previous research has identified descriptors or variables that can be used to categorize your plant into a group with other similar plants. You collect data on the multiple descriptors including plant morphology (color, types of leaves and flowers, number of anthers, etc.), location where identified, chromosme number, how the plant is propagated, etc. You enter the data, run the analysis, and hopefully you are able to assign the plant a group of like plants.

    d. Training population: Original population on which trait or characteristic data were collected. It is your goal to identify the characteristics or traits that best differentiate the known cases into distinct outcome groups, with as little error or misclassification as possible.

Concepts on Classifying

- A goal in classifying is to use characters that successfully separate the cases into distinct classes with little error.

- Using a single character, let's look at the distribution of two populations to see what would be considered a poor identifier and a favorable identifier.

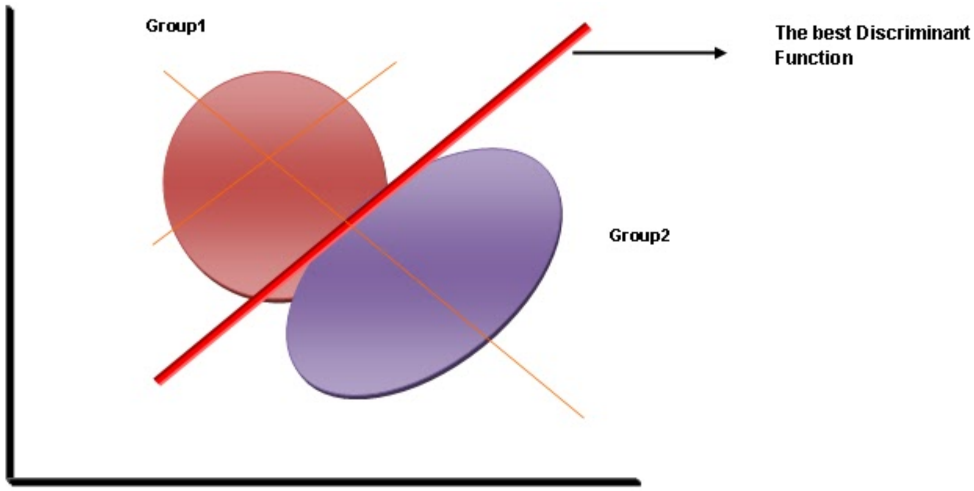


http://t3.gstatic.com/images?q=tbn:ANd9GcTcHzERVxW-nffzCVPOpzkdrbfbV5rnhgrwZcu2c-I8AyY3wGja

- Call the left distribution that for *X1* and the right distribution for *X2*.
- In both populations, a value lower than a certain value, *C,* would be classified in *X1* and if the value is >*C*, then the case would be classified into *X2*.
- In the situation portrayed in the top picture, the chance of misclassifying case is higher than would occur in the lower picture.
- If the variances for the two populations are equal (which is rarely the case), the value for *C* is the average of the means for the two populations.

$$C = \frac{\overline{X_1} + \overline{X_2}}{2}$$

- Generally, use of more than two or more classifying variable will reduce the errors in classifying unknown cases.
- The following figure shows an example of the resultant ellipsoids when classifying cases into two groups based on two variables.
- The intersection region in the previous figure and in the following figure shows the individuals that would be misclassified.



http://1.bp.blogspot.com/_rCLFLMl7aI0/TOlaJqGG1xI/AAAAAAAAAPc/eVZr1BDb3kM/s1600/untitled1.bmp

- The line in both figures showing the division between the two groups was defined by Fisher with the equation $Z = C$.

- Z is referred to as *Fisher's discriminant function* and has the formula:

$$Z = a_1X_1 + a_2X_2 + \ldots + a_pX_p$$

- A separate value of Z can be calculated for each individual in the group and a mean value of $\overline{Z_P}$ can be calculated for each group.

- A pooled sample variance of Z (S_Z^2) can be calculated similar to that done for the t-test.

- The *Mahalanobis distance* (D^2) or the squared distance between the means of the standardized values of Z*'s*.

- The greater the value of D^2 for a variable, the better it is able to differentiate between the groups or classes.

- The formulas for computing the coefficients a_1 and a_2 were derived by Fisher to maximize the D^2 or "distance" between the groups or classes.

<u>Hypothesis Testing in Discriminant Analysis</u>

- Assuming the classifying variables have a multivariate normal distribution, hypothesis testing is available in discriminant analysis.

- *Multivariate normal distribution*: A random vector is said to be *p*-variate normally distributed if every linear combination of its *p* components has a univariate normal distribution.

- **Warning**: The hypothesis tests don't tell you if you were correct in using discriminant analysis to address the question of interest.

- An *F*-test associated with D^2 can be performed to test the hypothesis that the classifying variables are able to differentiate unknown cases into groups better than by random chance (H_o: $D^2 = 0$).

  - $$F = \frac{N_I+N_{II}-P-1}{P(N_I+N_{II}-2)} X \frac{N_I N_{II}}{N_I+N_{II}} X D^2$$

- Another useful *F*-test is one to test the hypothesis that adding an additional variable improves discrimination or your ability to more accurately assign an individual to a group.

  - The hypothesis tested if an additional variable X_{P+1} will significantly increase D^2.

  - $$H_o: D_{P+1}^2 = D^2$$

  - $$F = \frac{(N_I+N_{II}-P-2)(N_I N_2)(D_{P+1}^2-D_P^2}{(N_I+N_{II})(N_I+N_{II}-2)+(N_I N_2)D_P^2}$$

- The probability of assigning an individual into the wrong class can be calculated and it is call the *Posterior Probability*.

  - The probability of assigning an individual to group 1 is: $\frac{1}{1+e^{(-Z+C)}}$

  - The probability of assigning an individual to group 2 is: 1-probability of being assigned to group 1.

<u>Determining if Your Discriminant Analysis Was Successful in Classifying Cases Into Groups</u>

- A *measure of goodness* to determine if your discriminant analysis was "successful" in classifying is to calculate the probabilities of misclassification, probability (II given I; classifying case as group II when it is actually belongs in group I) and probability (I given II; classifying as group I when the case belongs in group II).

- Two methods are available for determining unbiased estimated of the probabilities.

1. *Cross-validation*: An unbiased method where the original population is divided into two sub-populations. One sub-population is used as the training set and the other sub-population is used for validation. A possible problem occurs if the original population is small.

2. *Jackknife procedure*: An unbiased systematic method where one individual is excluded from the first group in the population, the discriminant function is then estimated, and that function is used to classify the excluded observation. This will allow you to estimate the probability of (II given I). You can use a similar procedure on the second group to estimate the probability of (I given II).

Adjusting the Dividing Point (*C*) Between the Groups

- The default in discriminant analysis is to have the dividing point set so there is an equal chance of misclassifying group I individuals into group II, and *vice versa*.
  - $C = \frac{\bar{Z}_I + \bar{Z}_{II}}{2} \; if \; q_I = q_{II}$
  - Where q_I is the prior probability of a case being assigned to group I and q_{II} is the prior probability of a case being assigned to group II.
- It is possible to establish a value of *C* where any desired ratio of the probabilities of the errors is established.
- An understanding of how to adjust the dividing point requires knowledge of the prior probability of having a case assigned to a specific group.
  - For example, going into research to establish causes of mental depression, was there *a priori* percentage of cases that were going to be assigned to either the depress or non-depressed group?
  - The assumptions going into the research was that 80% of the people would be labeled as non-depressed and 20% would be labeled as depressed.
  - Therefore, the prior probability of being non-depressed was 80%, which is labeled as $q_I = 0.8$. Likewise, the prior probability of being labeled depressed was 20%.
- The goal is to choose a dividing point of *C* so the total probability of misclassification is minimized.
  - This total probability is defined as: q_I[Prob(II given I)] + q_{II}[Prob(I given II)].
  - The formula for *C* that works for any values of q_1 and q_2 is:

$$C = \frac{\bar{Z}_I + \bar{Z}_{II}}{2} + ln\frac{q_{II}}{q_I}$$

Incorporating the Costs of Misclassification Into the Choice of *C*

- The cost of misclassification an individual into the wrong group can be figured into the discriminant analysis.

- For example, suppose it is four times more serious to misclassify a Group II case (e.g. depressed as non-depressed) into Group I than to misclassify a Group I case into Group II (e.g. non-depressed as depressed). These costs can be denoted as:

  - Cost(II given I)=1

  - Cost(I given II)=4

- The dividing point *C* can then be adjusted to minimize the cost of misclassification.

  - {q_I[Prob(II given I)] [cost(II given I)]}+ {q_{II}[Prob(I given II)][cost(I given II)]}

- The formula for *C* then becomes:

  - $= \frac{\bar{Z}_I+\bar{Z}_{II}}{2} + K$, where $K = ln \frac{q_{II}[cost(I\ given\ II)]}{q_I[cost(II\ given\ I)]}$

  - Using the depression example, $K = ln \frac{0.2(4)}{0.8(1)} = \ln(1) = 0$

Using SAS for Performing Discriminant Analysis

- SAS commands for Discriminant Analysis using a single classifying variable

```
proc discrim crosslisterr mahalanobis;
class cases;
var beddays;
title 'Discriminant analysis using only beddays';
run;
```

  - The crosslisterr option of proc discrim list those entries that are misclassified. Other options available are crosslist and crossvalidate.

  - The mahalanobis option of proc discrim displays the D^2 values, the *F*-value, and the probabilities of a greater D^2 between the group means.

Discriminant analysis using only beddays

The DISCRIM Procedure

Total Sample Size	294	**DF Total**	293
Variables	1	**DF Within Classes**	292
Classes	2	**DF Between Classes**	1

Number of Observations Read	294
Number of Observations Used	294

Class Level Information					
cases	**Variable Name**	**Frequency**	**Weight**	**Proportion**	**Prior Probability**
0	_0	244	244.0000	0.829932	0.500000
1	_1	50	50.0000	0.170068	0.500000

Pooled Covariance Matrix Information	
Covariance Matrix Rank	**Natural Log of the Determinant of the Covariance Matrix**
1	-1.82766

Discriminant analysis using only beddays

The DISCRIM Procedure

Squared Distance to cases		
From cases	0	1
0	0	0.38211
1	0.38211	0

D^2 values

F Statistics, NDF=1, DDF=292 for Squared Distance to cases		
From cases	0	1
0	0	15.85620
1	15.85620	0

F-statistic to test the null hypothesis (H_o: $D^2 = 0$).

Prob > Mahalanobis Distance for Squared Distance to cases		
From cases	0	1
0	1.0000	<.0001
1	<.0001	1.0000

Probabilities of >*F* for the test of the hypothesis H_o: $D^2 = 0$.

Discriminant analysis using only beddays

The DISCRIM Procedure

Generalized Squared Distance to cases		
From cases	0	1
0	0	0.38211
1	0.38211	0

Linear Discriminant Function for cases		
Variable	0	1
Constant	-0.09214	-0.54854
beddays	1.07054	2.61211

Variables	Classification function Group I=0 (non-depressed)	 Group II=1 (depressed)	Discriminant function
Constant	-0.09214	-0.54845	*C= -0.4564 = (-0.54854 – -0.09214)†*
Bed days	1.07054	2.61211	a_1= -1.54157 = (1.07054-2.61211)

†Note: Calculation of C, the **dividing point**, is done in reverse order, right value – left value.

- Discriminant function is *Z=-1.54157(bed days)*

Discriminant analysis using only beddays

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.DEPRESS
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into cases			
From cases	**0**	**1**	**Total**
0	202 82.79	42 17.21	244 100.00
1	29 58.00	21 42.00	50 100.00
Total	231 78.57	63 21.43	294 100.00
Priors	0.5	0.5	

Number and percent of misclassified cases.

Error Count Estimates for cases			
	0	**1**	**Total**
Rate	0.1721	0.5800	0.3761
Priors	0.5000	0.5000	

Discriminant analysis using only beddays

The DISCRIM Procedure
Classification Results for Calibration Data: WORK.DEPRESS
Cross-validation Results using Linear Discriminant Function

Posterior Probability of Membership in cases					
Obs	**From cases**	**Classified into cases**		**0**	**1**
5	**0**	**1**	*	0.2469	0.7531
10	**0**	**1**	*	0.2469	0.7531
14	**0**	**1**	*	0.2469	0.7531
15	**0**	**1**	*	0.2469	0.7531
27	**0**	**1**	*	0.2469	0.7531
28	**0**	**1**	*	0.2469	0.7531
37	**0**	**1**	*	0.2469	0.7531
43	**0**	**1**	*	0.2469	0.7531
54	**0**	**1**	*	0.2469	0.7531
58	**0**	**1**	*	0.2469	0.7531
62	**0**	**1**	*	0.2469	0.7531
65	**0**	**1**	*	0.2469	0.7531
71	**0**	**1**	*	0.2469	0.7531
72	**0**	**1**	*	0.2469	0.7531
81	**0**	**1**	*	0.2469	0.7531
87	**0**	**1**	*	0.2469	0.7531
88	**0**	**1**	*	0.2469	0.7531
89	**0**	**1**	*	0.2469	0.7531
91	**0**	**1**	*	0.2469	0.7531
92	**0**	**1**	*	0.2469	0.7531
94	**0**	**1**	*	0.2469	0.7531
97	**0**	**1**	*	0.2469	0.7531
102	**0**	**1**	*	0.2469	0.7531
111	**0**	**1**	*	0.2469	0.7531
119	**0**	**1**	*	0.2469	0.7531
120	**0**	**1**	*	0.2469	0.7531
127	**0**	**1**	*	0.2469	0.7531
132	**0**	**1**	*	0.2469	0.7531

Discriminant analysis using only beddays

The DISCRIM Procedure
Classification Results for Calibration Data: WORK.DEPRESS
Cross-validation Results using Linear Discriminant Function

Posterior Probability of Membership in cases					
Obs	**From cases**	**Classified into cases**		**0**	**1**
151	**0**	**1**	*	0.2469	0.7531
156	**0**	**1**	*	0.2469	0.7531
159	**0**	**1**	*	0.2469	0.7531
169	**0**	**1**	*	0.2469	0.7531
174	**0**	**1**	*	0.2469	0.7531
181	**0**	**1**	*	0.2469	0.7531
185	**0**	**1**	*	0.2469	0.7531
196	**0**	**1**	*	0.2469	0.7531
197	**0**	**1**	*	0.2469	0.7531
198	**0**	**1**	*	0.2469	0.7531
202	**0**	**1**	*	0.2469	0.7531
207	**0**	**1**	*	0.2469	0.7531
209	**0**	**1**	*	0.2469	0.7531
211	**0**	**1**	*	0.2469	0.7531
246	**1**	**0**	*	0.6176	0.3824
249	**1**	**0**	*	0.6176	0.3824
251	**1**	**0**	*	0.6176	0.3824
252	**1**	**0**	*	0.6176	0.3824
254	**1**	**0**	*	0.6176	0.3824
255	**1**	**0**	*	0.6176	0.3824
258	**1**	**0**	*	0.6176	0.3824
260	**1**	**0**	*	0.6176	0.3824
261	**1**	**0**	*	0.6176	0.3824
263	**1**	**0**	*	0.6176	0.3824
265	**1**	**0**	*	0.6176	0.3824
266	**1**	**0**	*	0.6176	0.3824
267	**1**	**0**	*	0.6176	0.3824
270	**1**	**0**	*	0.6176	0.3824

Discriminant analysis using only beddays

The DISCRIM Procedure
Classification Results for Calibration Data: WORK.DEPRESS
Cross-validation Results using Linear Discriminant Function

Posterior Probability of Membership in cases					
Obs	**From cases**	**Classified into cases**		**0**	**1**
271	**1**	**0**	*	0.6176	0.3824
274	**1**	**0**	*	0.6176	0.3824
275	**1**	**0**	*	0.6176	0.3824
276	**1**	**0**	*	0.6176	0.3824
278	**1**	**0**	*	0.6176	0.3824
279	**1**	**0**	*	0.6176	0.3824
280	**1**	**0**	*	0.6176	0.3824
281	**1**	**0**	*	0.6176	0.3824
282	**1**	**0**	*	0.6176	0.3824
284	**1**	**0**	*	0.6176	0.3824
287	**1**	**0**	*	0.6176	0.3824
288	**1**	**0**	*	0.6176	0.3824
289	**1**	**0**	*	0.6176	0.3824
291	**1**	**0**	*	0.6176	0.3824
293	**1**	**0**	*	0.6176	0.3824

**** Misclassified observation***

The posterior probability of belonging to each group is calculated. The case or individual is assigned to the class with the greatest probability value.

Discriminant analysis using only beddays

The DISCRIM Procedure
Classification Results for Calibration Data: WORK.DEPRESS
Cross-validation Results using Linear Discriminant Function

Number of Observations and Percent Classified into cases			
From cases	**0**	**1**	**Total**
0	202 82.79	42 17.21	244 100.00
1	29 58.00	21 42.00	50 100.00
Total	231 78.57	63 21.43	294 100.00
Priors	0.5	0.5	

Error Count Estimates for cases			
	0	**1**	**Total**
Rate	0.1721	0.5800	0.3761
Priors	0.5000	0.5000	

- SAS commands for Discriminant Analysis using a single classifying variable

```
proc stepdisc method=forward;
class cases;
var sex age marital educat employ income relig drink health regdoc treat beddays acuteill
        chronill;
title 'Stepwise discriminant analysis';
run;
proc discrim mahalanobis;
class cases;
var beddays income relig sex age health;
title 'Discriminant analysis following stepwise';
run;
```

  - The Proc Stepdisc command performs stepwise discriminant analysis.
  - This allows you to identify which variables significantly contribute to the maximization of D^2.
  - I chose the method=forward, which is Forward Stepwise Selection. This allows a systematic method where you start with no variables in the model and then keep adding one to increase D^2.
  - The analysis tested 14 variables and six were found to contribute significantly.
  - These six variables were then used to run a new discriminant analysis.

Stepwise discriminant analysis

The STEPDISC Procedure

• The Method for Selecting Variables is FORWARD			
Total Sample Size	294	**Variable(s) in the Analysis**	14
Class Levels	2	**Variable(s) Will Be Included**	0
		Significance Level to Enter	0.15

Number of Observations Read	294
Number of Observations Used	294

Class Level Information					
cases	**Variable Name**	**Frequency**	**Weight**	**Proportion**
0	_0	244	244.0000	0.829932
1	_1	50	50.0000	0.170068

Stepwise discriminant analysis

The STEPDISC Procedure
Forward Selection: Step 1

Statistics for Entry, DF = 1, 292				
Variable	**R-Square**	**F Value**	**Pr > F**	**Tolerance**
sex	0.0275	8.25	0.0044	1.0000
age	0.0102	3.02	0.0833	1.0000
marital	0.0001	0.04	0.8344	1.0000
educat	0.0122	3.61	0.0583	1.0000
employ	0.0118	3.48	0.0631	1.0000
income	0.0254	7.61	0.0062	1.0000
relig	0.0155	4.60	0.0327	1.0000
drink	0.0007	0.21	0.6442	1.0000
health	0.0243	7.26	0.0074	1.0000
regdoc	0.0072	2.11	0.1476	1.0000
treat	0.0076	2.25	0.1346	1.0000
beddays	0.0515	15.86	<.0001	1.0000
acuteill	0.0070	2.04	0.1538	1.0000
chronill	0.0105	3.10	0.0793	1.0000

Variable beddays will be entered.

Variable(s) That Have Been Entered
beddays

Multivariate Statistics					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.948495	15.86	1	292	<.0001
Pillai's Trace	0.051505	15.86	1	292	<.0001
Average Squared Canonical Correlation	0.051505				

Stepwise discriminant analysis

The STEPDISC Procedure
Forward Selection: Step 2

Statistics for Entry, DF = 1, 291				
Variable	**Partial R-Square**	**F Value**	**Pr > F**	**Tolerance**
sex	0.0208	6.18	0.0135	0.9865
age	0.0049	1.44	0.2317	0.9780
marital	0.0000	0.00	0.9450	0.9987
educat	0.0161	4.75	0.0302	0.9969
employ	0.0105	3.10	0.0795	0.9985
income	0.0306	9.18	0.0027	0.9978
relig	0.0144	4.25	0.0401	0.9988
drink	0.0015	0.42	0.5162	0.9981
health	0.0128	3.78	0.0529	0.9552
regdoc	0.0044	1.29	0.2579	0.9920
treat	0.0026	0.75	0.3862	0.9709
acuteill	0.0002	0.06	0.8034	0.8201
chronill	0.0045	1.32	0.2510	0.9721

Variable income will be entered.

Variable(s) That Have Been Entered	
income	beddays

Multivariate Statistics					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.919498	12.74	2	291	<.0001
Pillai's Trace	0.080502	12.74	2	291	<.0001
Average Squared Canonical Correlation	0.080502				

Stepwise discriminant analysis

The STEPDISC Procedure
Forward Selection: Step 3

Statistics for Entry, DF = 1, 290				
Variable	Partial R-Square	F Value	Pr > F	Tolerance
sex	0.0133	3.90	0.0492	0.9520
age	0.0113	3.31	0.0698	0.9438
marital	0.0022	0.63	0.4264	0.9437
educat	0.0034	0.99	0.3203	0.8147
employ	0.0038	1.11	0.2939	0.9350
relig	0.0176	5.21	0.0232	0.9941
drink	0.0032	0.94	0.3335	0.9877
health	0.0066	1.94	0.1652	0.9177
regdoc	0.0035	1.03	0.3113	0.9894
treat	0.0023	0.66	0.4163	0.9685
acuteill	0.0002	0.06	0.8058	0.8186
chronill	0.0028	0.81	0.3676	0.9646

Variable relig will be entered.

Variable(s) That Have Been Entered		
income	relig	beddays

Multivariate Statistics					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.903279	10.35	3	290	<.0001
Pillai's Trace	0.096721	10.35	3	290	<.0001
Average Squared Canonical Correlation	0.096721				

Stepwise discriminant analysis

The STEPDISC Procedure
Forward Selection: Step 4

Statistics for Entry, DF = 1, 289				
Variable	**Partial R-Square**	**F Value**	**Pr > F**	**Tolerance**
sex	0.0183	5.39	0.0210	0.9351
age	0.0083	2.42	0.1208	0.9290
marital	0.0015	0.43	0.5146	0.9404
educat	0.0056	1.62	0.2042	0.8041
employ	0.0048	1.40	0.2376	0.9330
drink	0.0012	0.34	0.5606	0.9574
health	0.0072	2.10	0.1487	0.9174
regdoc	0.0020	0.59	0.4444	0.9771
treat	0.0030	0.86	0.3540	0.9670
acuteill	0.0000	0.00	0.9903	0.8094
chronill	0.0031	0.89	0.3469	0.9644

Variable sex will be entered.

Variable(s) That Have Been Entered			
sex	income	relig	beddays

Multivariate Statistics					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.886743	9.23	4	289	<.0001
Pillai's Trace	0.113257	9.23	4	289	<.0001
Average Squared Canonical Correlation	0.113257				

Stepwise discriminant analysis

The STEPDISC Procedure
Forward Selection: Step 5

Statistics for Entry, DF = 1, 288				
Variable	**Partial R-Square**	**F Value**	**Pr > F**	**Tolerance**
age	0.0088	2.54	0.1119	0.9289
marital	0.0030	0.86	0.3557	0.9201
educat	0.0052	1.51	0.2199	0.7958
employ	0.0022	0.63	0.4278	0.9040
drink	0.0020	0.58	0.4483	0.9301
health	0.0065	1.88	0.1713	0.9160
regdoc	0.0030	0.86	0.3558	0.9311
treat	0.0009	0.26	0.6124	0.9020
acuteill	0.0000	0.00	0.9510	0.8004
chronill	0.0012	0.35	0.5539	0.9122

Variable age will be entered.

Variable(s) That Have Been Entered				
sex	age	income	relig	beddays

Multivariate Statistics					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.878982	7.93	5	288	<.0001
Pillai's Trace	0.121018	7.93	5	288	<.0001
Average Squared Canonical Correlation	0.121018				

Stepwise discriminant analysis

The STEPDISC Procedure
Forward Selection: Step 6

Statistics for Entry, DF = 1, 287				
Variable	**Partial R-Square**	**F Value**	**Pr > F**	**Tolerance**
marital	0.0000	0.01	0.9395	0.6560
educat	0.0072	2.09	0.1492	0.7861
employ	0.0083	2.40	0.1228	0.7655
drink	0.0012	0.34	0.5627	0.9163
health	0.0138	4.01	0.0461	0.8210
regdoc	0.0015	0.43	0.5145	0.8993
treat	0.0038	1.10	0.2948	0.8414
acuteill	0.0002	0.05	0.8251	0.7841
chronill	0.0037	1.07	0.3016	0.8687

Variable health will be entered.

Variable(s) That Have Been Entered					
sex	age	income	relig	health	beddays

Multivariate Statistics					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.866864	7.35	6	287	<.0001
Pillai's Trace	0.133136	7.35	6	287	<.0001
Average Squared Canonical Correlation	0.133136				

Stepwise discriminant analysis

The STEPDISC Procedure
Forward Selection: Step 7

Statistics for Entry, DF = 1, 286				
Variable	**Partial R-Square**	**F Value**	**Pr > F**	**Tolerance**
marital	0.0000	0.01	0.9202	0.5901
educat	0.0040	1.15	0.2850	0.7592
employ	0.0047	1.36	0.2452	0.7328
drink	0.0026	0.74	0.3898	0.8061
regdoc	0.0022	0.62	0.4314	0.8139
treat	0.0019	0.55	0.4601	0.7811
acuteill	0.0000	0.00	0.9801	0.7327
chronill	0.0007	0.21	0.6476	0.7466

No variables can be entered.

No further steps are possible.

Stepwise discriminant analysis

The STEPDISC Procedure

Forward Selection Summary									
Step	Number In	Entered	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	beddays	0.0515	15.86	<.0001	0.94849478	<.0001	0.05150522	<.0001
2	2	income	0.0306	9.18	0.0027	0.91949826	<.0001	0.08050174	<.0001
3	3	relig	0.0176	5.21	0.0232	0.90327861	<.0001	0.09672139	<.0001
4	4	sex	0.0183	5.39	0.0210	0.88674270	<.0001	0.11325730	<.0001
5	5	age	0.0088	2.54	0.1119	0.87898173	<.0001	0.12101827	<.0001
6	6	health	0.0138	4.01	0.0461	0.86686402	<.0001	0.13313598	<.0001

- This table is showing the six variables out of 14 that contributed significantly to increasing D^2.
- These variables should be used in a new discriminant analysis.

Discriminant analysis following stepwise

The DISCRIM Procedure

Total Sample Size	294	**DF Total**	293
Variables	6	**DF Within Classes**	292
Classes	2	**DF Between Classes**	1

Number of Observations Read	294
Number of Observations Used	294

Class Level Information					
cases	**Variable Name**	**Frequency**	**Weight**	**Proportion**	**Prior Probability**
0	_0	244	244.0000	0.829932	0.500000
1	_1	50	50.0000	0.170068	0.500000

Pooled Covariance Matrix Information	
Covariance Matrix Rank	**Natural Log of the Determinant of the Covariance Matrix**
6	7.60929

Discriminant analysis following stepwise

The DISCRIM Procedure

Squared Distance to cases		
From cases	0	1
0	0	1.08072
1	1.08072	0

F Statistics, NDF=6, DDF=287 for Squared Distance to cases		
From cases	0	1
0	0	7.34641
1	7.34641	0

Prob > Mahalanobis Distance for Squared Distance to cases		
From cases	0	1
0	1.0000	<.0001
1	<.0001	1.0000

Generalized Squared Distance to cases		
From cases	0	1
0	0	1.08072
1	1.08072	0

Discriminant analysis following stepwise

The DISCRIM Procedure

Linear Discriminant Function for cases		
Variable	0	1
Constant	-14.92578	-16.69696
beddays	-0.13224	1.12853
income	0.17499	0.14407
relig	1.94142	2.26781
sex	8.00153	8.81869
age	0.14707	0.12506
health	1.76348	2.20638

Discriminant analysis following stepwise

The DISCRIM Procedure

Number of Observations and Percent Classified into cases			
From cases	**0**	**1**	**Total**
0	180 73.77	64 26.23	244 100.00
1	15 30.00	35 70.00	50 100.00
Total	195 66.33	99 33.67	294 100.00
Priors	0.5	0.5	

Error Count Estimates for cases			
	0	**1**	**Total**
Rate	0.2623	0.3000	0.2811
Priors	0.5000	0.5000	

Previous Results – Single Factor (beddays)

Number of Observations and Percent Classified into cases			
From cases	**0**	**1**	**Total**
0	202 82.79	42 17.21	244 100.00
1	29 58.00	21 42.00	50 100.00
Total	231 78.57	63 21.43	294 100.00
Priors	0.5	0.5	

- SAS with prior probability considered
  - Non-depressed = 80% (0.80)
  - Depressed = 20% (0.20)

```
proc discrim mahalanobis crosslisterr;
class cases;
var beddays income relig sex age health;
priors '0'=0.8 '1'=0.2;
title 'Discriminant analysis following stepwise, with prior probability statement';
run;
```

  - The statement priors ‘0’=0.8 and ‘1’=0.2 will make adjustment to *C* to minimize the probability of misclassification.
  - You will see in the results that the number of individuals classified as depressed (1) was reduced. This results shows up in the final table that shows the proportion of non-depressed (0) and depressed (1) that were misclassified.

Discriminant analysis following stepwise, with prior probability statement

The DISCRIM Procedure

Total Sample Size	294	DF Total	293
Variables	6	DF Within Classes	292
Classes	2	DF Between Classes	1

Number of Observations Read	294
Number of Observations Used	294

Class Level Information					
cases	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	_0	244	244.0000	0.829932	0.800000
1	_1	50	50.0000	0.170068	0.200000

Pooled Covariance Matrix Information	
Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
6	7.60929

Discriminant analysis following stepwise, with prior probability statement

The DISCRIM Procedure

Squared Distance to cases		
From cases	0	1
0	0	1.08072
1	1.08072	0

F Statistics, NDF=6, DDF=287 for Squared Distance to cases		
From cases	0	1
0	0	7.34641
1	7.34641	0

Prob > Mahalanobis Distance for Squared Distance to cases		
From cases	0	1
0	1.0000	<.0001
1	<.0001	1.0000

Generalized Squared Distance to cases		
From cases	0	1
0	0.44629	4.29960
1	1.52701	3.21888

Discriminant analysis following stepwise, with prior probability statement

The DISCRIM Procedure

Linear Discriminant Function for cases		
Variable	**0**	**1**
Constant	-15.14892	-18.30640
beddays	-0.13224	1.12853
income	0.17499	0.14407
relig	1.94142	2.26781
sex	8.00153	8.81869
age	0.14707	0.12506
health	1.76348	2.20638

Discriminant analysis following stepwise, with prior probability statement

The DISCRIM Procedure

Classification Summary for Calibration Data: WORK.DEPRESS

Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into cases			
From cases	**0**	**1**	**Total**
0	232 95.08	12 4.92	244 100.00
1	40 80.00	10 20.00	50 100.00
Total	272 92.52	22 7.48	294 100.00
Priors	0.8	0.2	

Error Count Estimates for cases			
	0	**1**	**Total**
Rate	0.0492	0.8000	0.1993
Priors	0.8000	0.2000	

Discriminant analysis following stepwise, with prior probability statement

The DISCRIM Procedure
Classification Results for Calibration Data: WORK.DEPRESS
Cross-validation Results using Linear Discriminant Function

Posterior Probability of Membership in cases					
Obs	From cases	Classified into cases		0	1
11	1	0	*	0.6591	0.3409
17	1	0	*	0.5937	0.4063
29	1	0	*	0.7963	0.2037
47	1	0	*	0.9362	0.0638
50	0	1	*	0.1793	0.8207
58	1	0	*	0.8709	0.1291
59	1	0	*	0.5174	0.4826
60	1	0	*	0.8334	0.1666
68	1	0	*	0.9638	0.0362
69	1	0	*	0.6133	0.3867
74	1	0	*	0.8239	0.1761
76	1	0	*	0.6181	0.3819
80	1	0	*	0.9454	0.0546
81	0	1	*	0.4181	0.5819
97	0	1	*	0.4943	0.5057
99	1	0	*	0.9636	0.0364
104	1	0	*	0.6214	0.3786
106	1	0	*	0.8005	0.1995
107	0	1	*	0.4795	0.5205
108	0	1	*	0.2788	0.7212
112	1	0	*	0.8675	0.1325
113	1	0	*	0.8216	0.1784
114	1	0	*	0.6470	0.3530
125	1	0	*	0.7035	0.2965
126	1	0	*	0.8583	0.1417
131	1	0	*	0.5421	0.4579
132	1	0	*	0.6453	0.3547

Discriminant analysis following stepwise, with prior probability statement

The DISCRIM Procedure
Classification Results for Calibration Data: WORK.DEPRESS
Cross-validation Results using Linear Discriminant Function

Posterior Probability of Membership in cases					
Obs	From cases	Classified into cases		0	1
140	1	0	*	0.9187	0.0813
141	0	1	*	0.4459	0.5541
142	1	0	*	0.8641	0.1359
144	1	0	*	0.6812	0.3188
147	1	0	*	0.5947	0.4053
151	1	0	*	0.6899	0.3101
153	0	1	*	0.3739	0.6261
154	0	1	*	0.3600	0.6400
164	0	1	*	0.4724	0.5276
174	1	0	*	0.9533	0.0467
177	1	0	*	0.6513	0.3487
182	1	0	*	0.8411	0.1589
186	1	0	*	0.7001	0.2999
188	1	0	*	0.5663	0.4337
189	1	0	*	0.5875	0.4125
191	0	1	*	0.3799	0.6201
211	1	0	*	0.7287	0.2713
223	0	1	*	0.4332	0.5668
225	1	0	*	0.8191	0.1809
228	0	1	*	0.4235	0.5765
235	1	0	*	0.5895	0.4105
251	1	0	*	0.7746	0.2254
257	1	0	*	0.8777	0.1223
258	1	0	*	0.6982	0.3018
259	0	1	*	0.4102	0.5898
279	0	1	*	0.3832	0.6168

Discriminant analysis following stepwise, with prior probability statement

The DISCRIM Procedure

Classification Results for Calibration Data: WORK.DEPRESS

Cross-validation Results using Linear Discriminant Function

Posterior Probability of Membership in cases					
Obs	From cases	Classified into cases		0	1
288	1	0	*	0.9272	0.0728
289	1	0	*	0.7142	0.2858

** Misclassified observation*

Discriminant analysis following stepwise, with prior probability statement

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.DEPRESS
Cross-validation Summary using Linear Discriminant Function

Number of Observations and Percent Classified into cases			
From cases	**0**	**1**	**Total**
0	230 94.26	14 5.74	244 100.00
1	41 82.00	9 18.00	50 100.00
Total	271 92.18	23 7.82	294 100.00
Priors	0.8	0.2	

Error Count Estimates for cases			
	0	**1**	**Total**
Rate	0.0574	0.8200	0.2099
Priors	0.8000	0.2000	

The percentage of depressed individuals misclassified as non-depressed was greatly reduced vs. the analysis without prior probability consideration (0.2623 vs. 0.0574). This is good because people needing help will be able to get the necessary medications.

The percentage of misclassified depressed individuals was greatly increased vs. the analysis without prior probability consideration (0.30 vs. 0.82). This increase is good because medication will not be needlessly given to this group.

Summary of the Different Discriminant Analyses Based on Comparisons of Misclassified Cases

- Two types of misclassification are possible:
    - Classifying a case as non-depressed (0) when it should be classified as depressed (1) (under medicating).
    - Classifying a cased as depressed (1) when it should be classified as non-depressed (0) (over medicating).

1. Discriminant analysis with one independent variable, bed days:

Number of Observations and Percent Classified into cases			
From cases	0	1	Total
0	202 82.79	42 17.21	244 100.00
1	29 58.00	21 42.00	50 100.00
Total	231 78.57	63 21.43	294 100.00
Priors	0.5	0.5	

- 17.21 % of cases identified as non-depressed are will now receive the necessary care.
- 58% of cases identified as depressed are misdiagnosed; they should have been classified as normal. It is good that these individuals are identified because they will not be given medication they don't need.

2. Discriminant analysis based on six variables selected using stepwise discriminant analysis:

Number of Observations and Percent Classified into cases			
From cases	0	1	Total
0	180 73.77	64 26.23	244 100.00
1	15 30.00	35 70.00	50 100.00
Total	195 66.33	99 33.67	294 100.00
Priors	0.5	0.5	

- The number and percentage of cases misdiagnosed as being non-depressed went up (17.21 vs. 26.23%). This may be good because more people are getting the needed treatment.
- The number and percentage of cases misdiagnosed as depressed went down (58.0 vs. 30.0%). This is good because fewer people (29 vs. 15) are receiving unnecessary treatment.

3. Discriminant analysis when *a priori* proportions used for the original classifying pre-discriminant analysis are considered.

Number of Observations and Percent Classified into cases			
From cases	**0**	**1**	**Total**
0	230 94.26	14 5.74	244 100.00
1	41 82.00	9 18.00	50 100.00
Total	271 92.18	23 7.82	294 100.00
Priors	0.8	0.2	

- Using *a priori* values, the number and percentage of cases misdiagnosed as being non-depressed went down considerably (26.23 vs. 5.74%). This means fewer will be receiving unnecessary medication.

- Percentage of cases misdiagnosed as depressed went up (82.0 vs. 30.0%), but this is not a bad result because 41 fewer people will be receiving unnecessary medication.