Vandana Agrawal

Discriminant Analysis – Depression Data Set

The objective of Discriminant analysis is to identify the variable which will classify the data into distinct classes. Discriminant analysis is used when we already have class defined and we want to build a model that will help us to classify any new observation into a class.

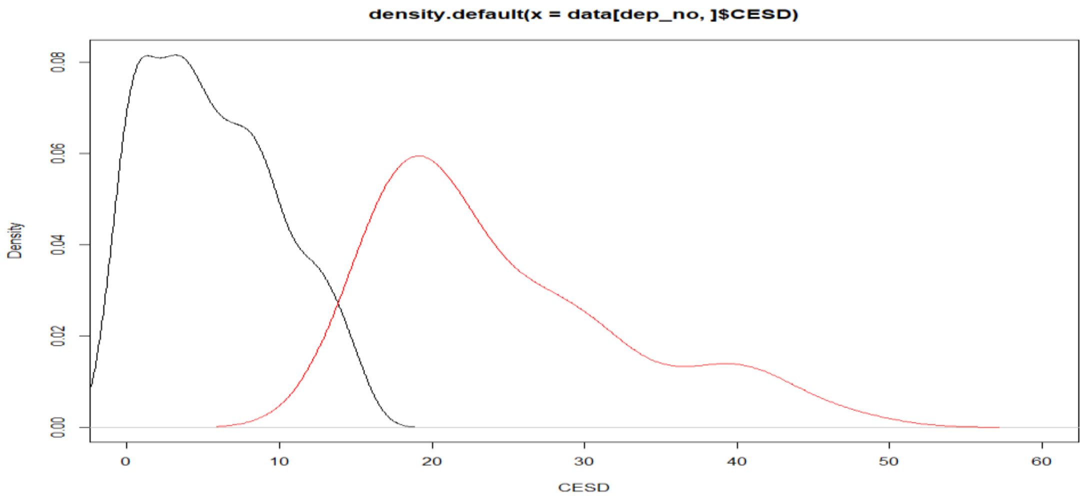
Step 1. Import the data

Step 2. Data Exploration

Data Summary:

```
> summary(data)
      SEX             AGE            MARITAL          EDUCAT          EMPLOY          INCOME
 Min.   :1.000   Min.   :18.00   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   : 2.00
 1st Qu.:1.000   1st Qu.:28.00   1st Qu.:2.000   1st Qu.:3.00   1st Qu.:1.000   1st Qu.: 9.00
 Median :2.000   Median :42.50   Median :2.000   Median :3.00   Median :1.000   Median :15.00
 Mean   :1.622   Mean   :44.41   Mean   :2.374   Mean   :3.48   Mean   :2.109   Mean   :20.57
 3rd Qu.:2.000   3rd Qu.:59.00   3rd Qu.:3.000   3rd Qu.:4.00   3rd Qu.:3.000   3rd Qu.:28.00
 Max.   :2.000   Max.   :89.00   Max.   :5.000   Max.   :7.00   Max.   :7.000   Max.   :65.00
     RELIG             C1               C2               C3               C4               C5
 Min.   :1.000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
 Median :1.000   Median :0.0000   Median :0.000   Median :0.0000   Median :0.0000   Median :0.000
 Mean   :1.983   Mean   :0.3639   Mean   :0.568   Mean   :0.5442   Mean   :0.1939   Mean   :0.551
 3rd Qu.:3.000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.000
 Max.   :6.000   Max.   :3.0000   Max.   :3.000   Max.   :3.0000   Max.   :3.0000   Max.   :3.000
       C6               C7               C8               C9              C10              C11
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000   Median :0.0000   Median :0.0000
 Mean   :0.2483   Mean   :0.2449   Mean   :0.3503   Mean   :0.568   Mean   :0.4626   Mean   :0.3605
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :3.0000   Max.   :3.0000   Max.   :3.0000   Max.   :3.000   Max.   :3.0000   Max.   :3.0000
      C12              C13              C14              C15              C16              C17
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :1.0000   Median :0.000
 Mean   :0.5136   Mean   :0.3401   Mean   :0.7211   Mean   :0.6735   Mean   :0.7449   Mean   :0.619
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000
 Max.   :3.0000   Max.   :3.0000   Max.   :3.0000   Max.   :3.0000   Max.   :3.0000   Max.   :3.000

      C18              C19              C20              CESD             CASES            DRINK
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 0.000   Min.   :0.0000   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 3.000   1st Qu.:0.0000   1st Qu.:1.000
 Median :0.0000   Median :0.0000   Median :0.0000   Median : 7.000   Median :0.0000   Median :1.000
 Mean   :0.3095   Mean   :0.2551   Mean   :0.2483   Mean   : 8.884   Mean   :0.1701   Mean   :1.204
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:12.000   3rd Qu.:0.0000   3rd Qu.:1.000
 Max.   :3.0000   Max.   :3.0000   Max.   :3.0000   Max.   :47.000   Max.   :1.0000   Max.   :2.000
     HEALTH           REGDOC           TREAT           BEDDAYS          ACUTEILL         CHRONILL
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :2.000   Median :1.000   Median :1.000   Median :0.0000   Median :0.0000   Median :1.0000
 Mean   :1.772   Mean   :1.187   Mean   :1.497   Mean   :0.2143   Mean   :0.2959   Mean   :0.5068
 3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :4.000   Max.   :2.000   Max.   :2.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```



In above density plot , black line represents people with Depression and red colored line represents people with out depression. CESD value clearly separate out people with and with out depression. Removing this column from the data set to analyze how other variables contribute to classifying the Cases into depression and no depression class.

Step 3: Variable Selection and t- test:

Divide the data set into population of people having depression and population of people who do not have depression. Perform t-test to determine whether the mean of a population of people with depression significantly differs from the mean of population of people who do not have depression for a given variable.

```
dep_no=which(depress==0)
dep_yes=which(depress==1)
```

Variable Name	t-Value	p-value	Population mean with depression	Population mean without depression	Difference in population mean
Sex	3.2764	0.001543	1.8	1.586066	0.213934
Age	-1.7866	0.07817	40.38	45.2418	-4.8618
Marital	-0.2007	0.8415	2.34	2.381148	-0.041148
Education	-2.0734	0.04146	3.16	3.545082	-0.385082
Employment	1.7628	0.0825	2.48	2.032787	0.447213
Income	-3.7507	0.000283	15.2	21.67623	-6.47623
Religion	1.9609	0.05417	2.32	1.913934	0.406066
C1	7.0751	3.66E-09	1.32	0.1680328	1.1519672
C2	9.7814	9.65E-14	1.68	0.3401639	1.3398361
C3	9.3065	6.52E-13	1.76	0.295082	1.464918

C4	4.9184	9.82E-06	0.84	0.06147541	0.77852459
C5	9.7205	1.28E-13	1.68	0.3196721	1.3603279
C6	5.3847	1.76E-06	0.9	0.1147541	0.7852459
C7	6.9727	6.7E-09	1.16	0.05737705	1.10262295
C8	3.0179	0.003697	0.7	0.2786885	0.4213115
C9	5.5573	6.7E-07	1.32	0.4139344	0.9060656
C10	10.353	1.69E-14	1.64	0.2213115	1.4186885
C11	6.9567	4.83E-09	1.16	0.1967213	0.9632787
C12	5.9189	1.87E-07	1.22	0.3688525	0.8511475
C13	3.869	0.00029	0.82	0.2418033	0.5781967
C14	6.2477	4.47E-08	1.58	0.545082	1.034918
C15	5.7742	3.15E-07	1.46	0.5122951	0.9477049
C16	6.6053	1.21E-08	1.6	0.5696721	1.0303279
C17	6.79	6.12E-09	1.5	0.4385246	1.0614754
C18	4.0799	0.000152	0.82	0.204918	0.615082
C19	2.9822	0.004285	0.58	0.1885246	0.3914754
C20	5.8536	3.37E-07	0.92	0.1106557	0.8093443
Drink	-0.4775	0.6344	1.18	1.209016	-0.029016
Health	2.3545	0.02168	2.06	1.713115	0.346885
RegDoc	1.308	0.1955	1.26	1.172131	0.087869
Treat	-1.512	0.135	1.4	1.516393	-0.116393
Bed Days	3.3248	0.0015	0.42	0.1721311	0.2478689
Acutelll	1.3496	0.1817	0.38	0.2786885	0.1013115
Chronill	1.7854	0.07843	0.62	0.4836066	0.1363934

From the above t test we see that almost all the variable mean values overlap each other. There is no significant mean difference. We do not have clearly defined separation in variable mean through which distinguish the classes clearly. Hence going ahead utilizing all the variable for Analysis. We do not want to drop the variable a variable or its combinations might have significant effect on class separation analysis.

Step 4: Summary Statistics:

<u>Column mean of population with no depression.</u>

```
> xbar1
        SEX         AGE     MARITAL      EDUCAT      EMPLOY      INCOME       RELIG
 1.58606557 45.24180328  2.38114754  3.54508197  2.03278689 21.67622951  1.91393443
         C1          C2          C3          C4          C5          C6          C7
 0.16803279  0.34016393  0.29508197  0.06147541  0.31967213  0.11475410  0.05737705
         C8          C9         C10         C11         C12         C13         C14
 0.27868852  0.41393443  0.22131148  0.19672131  0.36885246  0.24180328  0.54508197
        C15         C16         C17         C18         C19         C20       DRINK
 0.51229508  0.56967213  0.43852459  0.20491803  0.18852459  0.11065574  1.20901639
     HEALTH      REGDOC       TREAT     BEDDAYS    ACUTEILL    CHRONILL
 1.71311475  1.17213115  1.51639344  0.17213115  0.27868852  0.48360656
```

.

Column mean of population with Depression.

```
> xbar2
     SEX      AGE  MARITAL   EDUCAT   EMPLOY   INCOME    RELIG       C1       C2 
    1.80    40.38     2.34     3.16     2.48    15.20     2.32     1.32     1.68 
      C3       C4       C5       C6       C7       C8       C9      C10      C11 
    1.76     0.84     1.68     0.90     1.16     0.70     1.32     1.64     1.16 
     C12      C13      C14      C15      C16      C17      C18      C19      C20 
    1.22     0.82     1.58     1.46     1.60     1.50     0.82     0.58     0.92 
   DRINK   HEALTH   REGDOC    TREAT  BEDDAYS ACUTEILL CHRONILL 
    1.18     2.06     1.26     1.40     0.42     0.38     0.62 
> |
```

Step 5: Model

Above, we separated the population mean of each class. For classification of data point on each class we wanted to project the data point to one dimensional vector such that the difference of population mean of both class is maximum and variance of projected data point should be minimum.

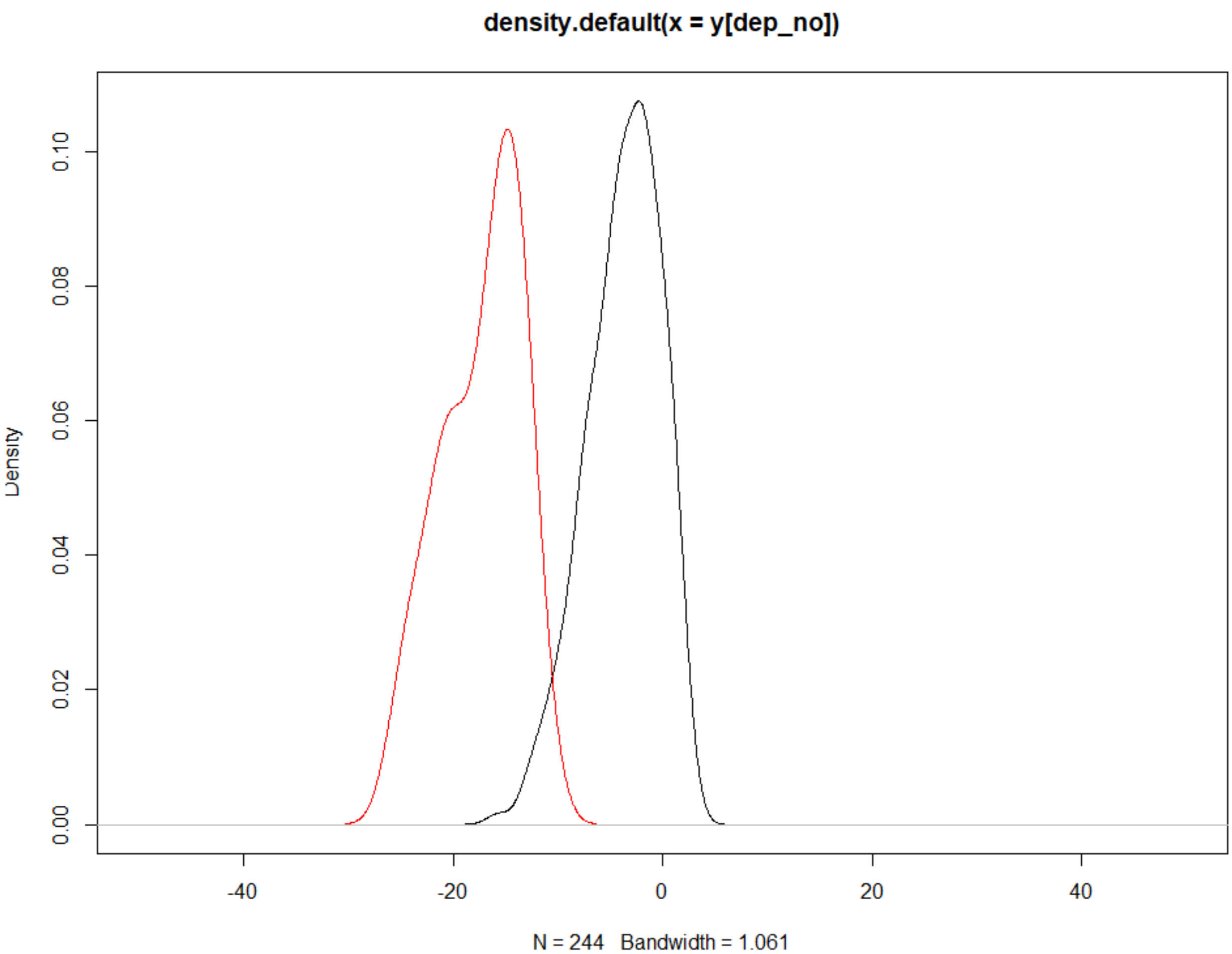
```
> S1=cov(data[dep_no,])
> S2=cov(data[dep_yes,])
> Sp=(2*S1+2*S2)/4
> y=(xbar1-xbar2)%*%solve(Sp)%*%t(as.matrix(data))
> y=as.vector(y)
> y
  [1]   0.86428583  -5.45901762  -3.24492355   0.23542687  -1.91231651  -5.08
441003
  [7]  -1.90982880  -4.97628975 -14.29587955  -0.27004754 -13.35155820  -4.35
862346
 [13]  -3.54613849  -1.36246553  -8.32812581 -19.34499798 -24.47548626  -5.08
336087
 [19]  -0.70232771   1.81536247  -2.17691505  -2.36108660  -1.25480711  -2.56
490040
 [25]  -4.51844094  -4.83460621   2.65751753 -10.21320414 -21.32829941  -3.87
860060
 [31]  -1.95414492  -4.73914646  -4.63458417  -5.32525958  -3.29979637  -1.04
257534
 [37]  -7.23540319 -13.97092648  -1.36785227  -0.42797447  -2.02946716  -3.14
533485
 [43] -12.27634031  -1.69133496  -0.68738592  -1.51559246 -13.97916415  -8.41
296468
```

```
 [49]  -9.56347761  -7.09461050  -0.52185303 -11.39499013  -7.11828083   0.74
811444
 [55]  -7.20912842  -3.64920933  -1.79455934 -25.54876643 -12.89831693 -16.98
794113
 [61]  -8.87193461  -2.71624231  -5.48598025  -4.55855184  -3.50647875  -0.60
745186
 [67]  -1.33992005 -12.42957746 -17.19182723  -5.75708926  -1.80870211  -1.98
239859
 [73] -23.73235918 -14.90080087  -5.18902000 -18.03335805  -4.22998149  -4.52
050082
 [79]  -0.74900061 -16.89708547  -6.47359900  -6.35096202 -12.84931600  -4.06
403966
 [85]   1.43687720   0.01938207  -9.49613100  -4.62045754  -7.35539764   0.92
709708
 [91]  -7.74462873  -3.30708833  -4.15989894  -3.80782992  -3.13444736  -2.34
803771
 [97]  -6.62497589  -1.08542677 -15.77188649   0.75618276  -3.93307814  -6.36
027905
[103]  -4.72363412 -15.14861887  -9.12661188 -11.11592107 -11.44468367  -9.17
558789
[109]  -3.76527695  -5.22785660 -19.04535995 -15.43045121 -15.59159178 -19.56
048146
[115] -11.67573830  -6.00090817 -14.15749087  -3.81101075  -2.37696711  -0.75
873049
[121]  -3.71169218  -8.26886181  -2.79530036 -22.60500988 -23.41496365 -15.39
205346
[127] -15.71567101  -0.13812702  -8.20091156  -3.56335384 -15.15086605 -21.13
450777
[133]   1.22308869  -9.31079035   0.77259487   0.31075689  -1.76641753  -1.34
977767
[139]   1.92750541 -14.29635548  -7.51166817 -17.63086747   1.60818003 -22.87
809987
[145]  -2.87310552   0.14331824 -15.73332148  -3.45430761  -6.02305786  -2.11
693116
[151] -12.95852420  -9.83991874  -7.72039072  -5.67319693  -1.93422467  -2.13
957072
[157]  -3.09601426  -3.03946802  -4.39631026  -0.98005021  -2.55513616   1.20
673443
```

```
[163]  -1.67761050  -1.83257218  -1.16070514  -0.15624812  -2.10382254   0.78
693131
[169]  -0.58341473  -5.20754781  -0.09681513  -0.34201335  -7.60303600 -15.86
523893
[175]  -1.56631577  -7.54813841 -13.54531740  -6.50812603  -5.08059883 -10.87
659790
[181]  -3.04003597 -20.87594653  -1.75904540  -0.28114540  -1.95446688 -13.48
192247
[187]  -3.41861518 -13.11480124 -21.08457498  -6.16807477  -5.22958848  -6.73
934389
[193]   1.19076078   0.39590052  -4.34840858   0.75803154  -6.23578822   0.70
928195
[199]  -2.63308918  -4.21009999 -20.08983054  -2.70155521  -3.67582261  -0.40
613250
[205]  -6.33655275   0.84833331   1.61737895  -1.93814394  -3.67972909 -13.18
128302
[211] -16.91452725   2.27714607   1.03652431  -0.59355215   1.24612734  -7.64
000271
[217]  -0.88776368  -4.29815743  -4.60174668  -1.45896679  -2.60192687  -0.93
319455
[223]  -4.02510083  -9.94678623 -21.26758582  -0.39258374  -6.23255996  -7.84
934075
[229]  -1.86375712 -10.87777710  -2.53299990   0.31211051  -2.52496882  -3.43
584176
[235] -20.54358701  -0.52312744 -10.62605173  -4.07538409  -6.68482252  -2.98
921230
[241]  -7.64088236  -4.77035916  -5.43343380  -0.81651692   1.00187178  -6.66
417033
[247]  -7.46884452  -5.17159437  -4.11839832  -7.11200641 -18.02676074  -2.08
243757
[253]  -1.93887460   2.18839590  -5.71324813 -16.17618521 -11.52500476 -13.81
154558
[259]  -9.39962874   1.20672288  -4.57121851  -1.51677842  -7.16515628  -3.19
257288
[265]  -2.47901411  -3.98821585 -12.41486220  -2.45031683  -4.04765986   0.32
931594
[271]  -5.86191357  -2.14334582  -7.18567820  -6.84617589  -9.28359144  -6.14
616813
```

```
[277]  -0.20320612  -7.96986211 -10.46755912   0.13158803   0.19052051  -5.14
675500
[283]  -8.79207848   1.14092844  -7.87991497   0.49594801  -4.11318009 -19.78
257858
[289] -24.98116074   2.44415978  -1.64784827  -5.33082130   1.45586730  -7.61
941214
```

Plotting the projected data set values of both the class population

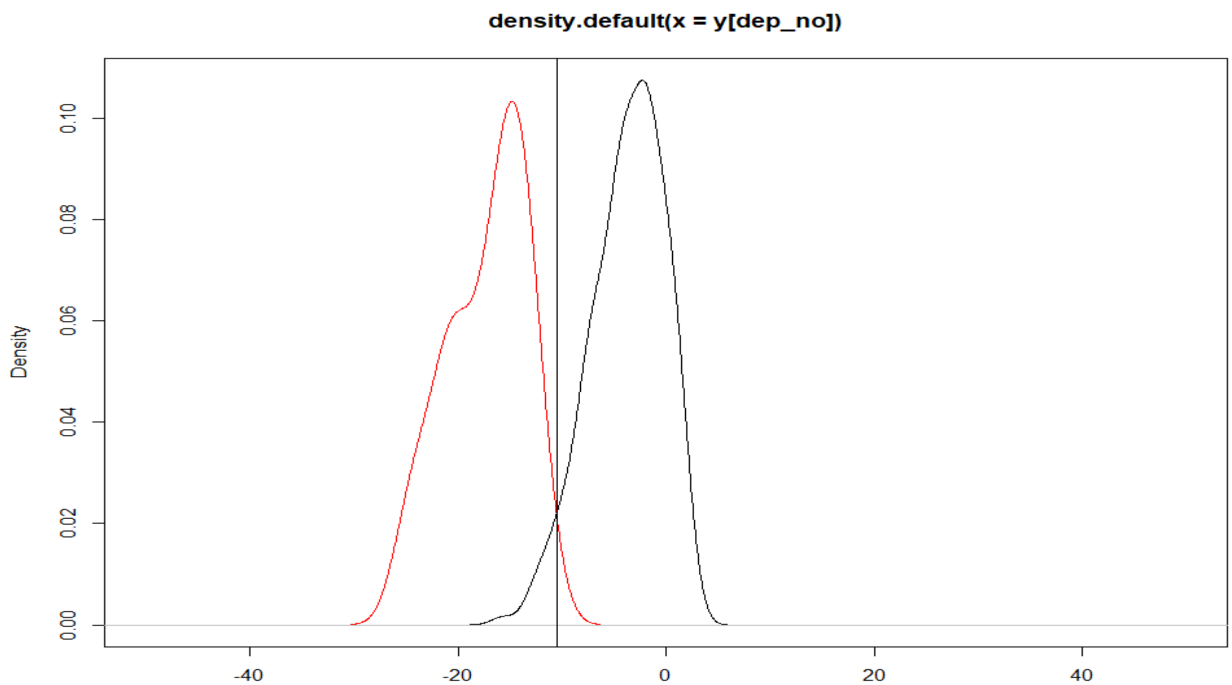


In above plot Black lines represents projected population density of People who do not have depression. And red line represents the Population of people having depression. The intersection point/Cutoff between the two would be the decision point for classification of each class .

```
> cutoff=.5*(xbar1-xbar2)%*%solve(Sp)%*%(xbar1+xbar2)
> cutoff=as.vector(cutoff)
> cutoff
[1] -10.54708
```

Vandana Agrawal

Plot the Cut off lines:

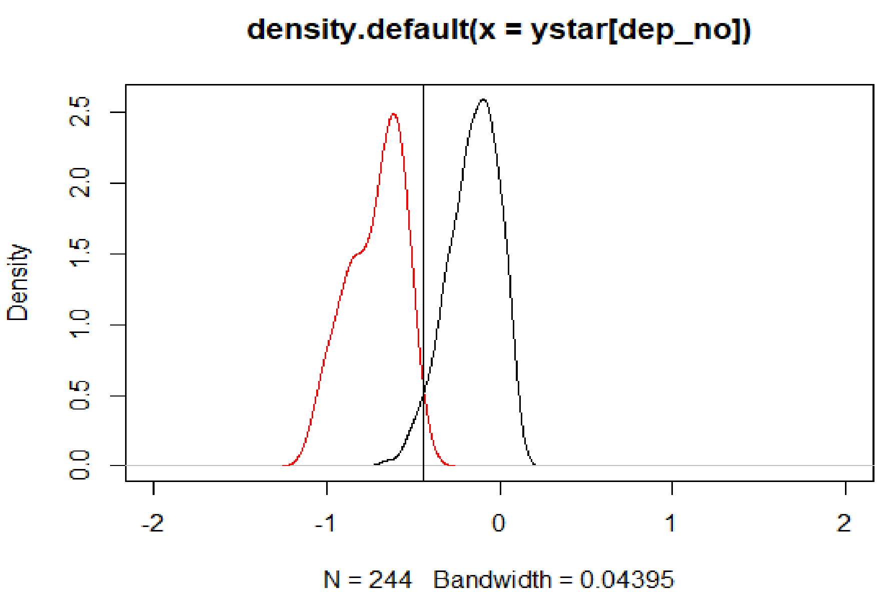



```
> a=t((xbar1-xbar2)%*%solve(Sp))
> a
                [,1]
SEX      -1.54934437
AGE      -0.03131146
MARITAL   0.05463296
EDUCAT    0.19417368
EMPLOY    0.74631447
INCOME    0.01185779
RELIG     0.10545222
C1       -0.19682607
C2        1.05816343
C3       -2.68830997
C4       -0.40064815
C5       -0.08022393
C6       -0.50950045
C7        0.04198721
C8        0.36764626
C9       -0.20778627
C10      -2.95245375
```

```
C11       -0.35071696
C12       -1.55482886
C13       -0.62319204
C14       -0.41128870
C15       -0.21457688
C16       -1.14814460
C17       -0.86343710
C18        0.09805581
C19       -0.51095526
C20       -0.92966926
DRINK      1.08266229
HEALTH    -0.04747901
REGDOC     1.40477972
TREAT     -1.15104084
BEDDAYS   -2.07052252
ACUTEILL   0.45382855
CHRONILL  -0.03160519
```

From the above weight we can deduce that the separation for variable sex,C2,C3,C10,C12,C16, Drink, Treat, Bed, regdoc, days projected values gives more separation . Also weight multiplied by the data point value will give the score and assuming their posterior probability is .5 we will get the score which will tell us which class out case falls in.

Standardizing the data and weights the separation plots still looks the same.



Results:

```
> c_accuracy(depress,classify)
     recall   precision    accuracy         tpr         fpr    fmeasure
 1.00000000  0.81967213  0.96258503  1.00000000  0.04508197  0.90090090
         tp          tn          fp          fn
50.00000000 233.00000000 11.00000000  0.00000000
```

```
> upper=(xbar1-xbar2)%*%solve(Sp)%*%((xbar1-xbar2))
> upper
         [,1]
[1,] 13.60461

>
```

Upper gives us the difference of projected population mean separation.

```
> sy=(sum((y[dep_no]-mean(y[dep_no]))^2)+sum((y[dep_yes]-mean(y[dep_yes]))^2))/
(length(dep_no)+length(dep_yes)-2)
> sy
[1] 12.89031
> upper/sy
         [,1]
[1,] 1.055414
```

Separation in this case is quite good 1.055414

In above case we got the accuracy of almost 96% which is quite high and the reason could be we have used C1 to C20 variable there sum is directly related to people having depression or not.

In order to compare Discriminant analysis with logistic regression and to analyses the what all question from C1 to C20 are significant ,removed the variable which are highly correlated. Considered the same variables as we have used in Logistic regression .i.e
SEX,AGE,MARITAL,EDUCAT,EMPLOY,INCOME,RELIG,DRINK,HEALTH,REGDOC,TREAT,BEDDAYS,ACUTEILL,
CHRONILL,C8,C9,C12,C13,C14,C16,C17,C18

```
> xbar1
      SEX       AGE   MARITAL    EDUCAT    EMPLOY    INCOME     RELIG
1.5860656 45.2418033 2.3811475 3.5450820 2.0327869 21.6762295 1.9139344
       C8        C9       C12       C13       C14       C16       C17
0.2786885 0.4139344 0.3688525 0.2418033 0.5450820 0.5696721 0.4385246
      C18     DRINK    HEALTH    REGDOC     TREAT   BEDDAYS  ACUTEILL
0.2049180 1.2090164 1.7131148 1.1721311 1.5163934 0.1721311 0.2786885
 CHRONILL
0.4836066
```

Vandana Agrawal

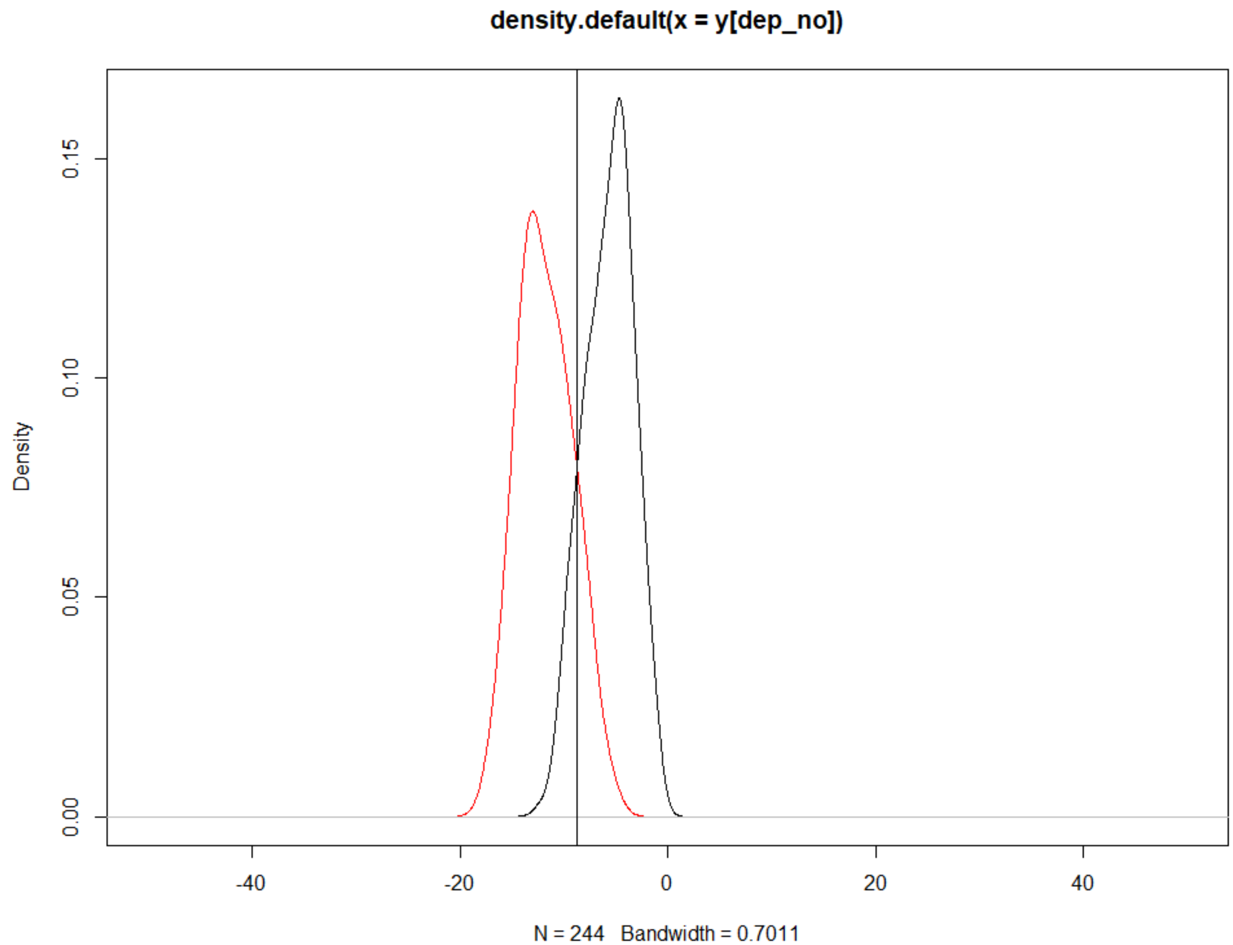
```
> xbar2
     SEX      AGE  MARITAL   EDUCAT   EMPLOY   INCOME    RELIG       C8       C9 
    1.80    40.38     2.34     3.16     2.48    15.20     2.32     0.70     1.32 
     C12      C13      C14      C16      C17      C18    DRINK   HEALTH   REGDOC 
    1.22     0.82     1.58     1.60     1.50     0.82     1.18     2.06     1.26 
   TREAT  BEDDAYS ACUTEILL CHRONILL 
    1.40     0.42     0.38     0.62 
```

Weights:

```
> a
                [,1]
SEX      -1.61931088
AGE      -0.01160252
MARITAL   0.03748113
EDUCAT   -0.10983603
EMPLOY    0.13662683
INCOME    0.04476452
RELIG    -0.33251989
C8       -0.47029878
C9       -1.08041720
C12      -0.77922322
C13      -0.56989733
C14      -0.88312429
C16      -0.88374928
C17      -0.94553297
C18      -0.74376666
DRINK     0.48679559
HEALTH    0.25771353
REGDOC    0.02705919
TREAT    -0.80488440
BEDDAYS  -0.71139183
ACUTEILL -0.16931948
CHRONILL -0.04003868
```

```
> cutoff
[1] -8.732985

>
```



In Above plot red line represents people who have depression and black represents people with no depression. In this case we find quite much more area where the two line overlap on each other. And that might be because we have removed some of the variable which are helping to separate out the two classes.

Results:

```
> c_accuracy(depress,classify)
     recall   precision    accuracy         tpr         fpr    fmeasure          tp
  0.8400000   0.5915493   0.8741497   0.8400000   0.1188525   0.6942149  42.0000000
         tn          fp          fn
215.0000000  29.0000000   8.0000000
```

```
> upper
         [,1]
[1,] 6.096174
> sy
[1] 5.680868
> upper/sy
         [,1]
[1,] 1.073106
```

From the above result we can conclude that the accuracy of the model is reduced to 87%. But Important to note here is that there is false negative in this case. And our previous is good because in this model we are saying people no when they do have depression is not a good classification.

Logistic Regression vs Discriminant Analysis:

In given sample for discriminant analysis let say We want to develop a model to predict the outcome depression or no depression for a new patient. If a person have serious issue with alcohol consumption age is between 30-40 and Bed days score 3or more than in which category it will fall ? it indicates which of the predictors are the most differentiating (highest discriminant weights), in other words, which predictor distinguish best among these patients and why they fall into one class versus another class. In summary, it is a technique for classification, differentiation, and profiling.

logistic regression is very similar to discriminant analysis, the primary question addressed by LR is “How likely is the case to belong to each class”. (What is the probability of person having Depression) In contrast, the primary question addressed by discriminant analysis, is “Which class is the case most likely to belong to”. So, logistic regression estimates the probability of each case to belong to two groups (on the dependent variable) or the probability of occurrence if the predictor changes. As the focus is on probability the goal of analyses is to create a linear combination of the log of the odds of a case being in one group or another. An odds ratio is estimated for each of the predictor variables in the model.

Also, in Discriminate Analysis we assume the sample population has same covariance and variables are normally distributed. But in logistics regression that is not required.