Microsoft Excel でする判別分析

大阪市立大学経済学部
中川　満
2009/06/08

(1) **（データのダウンロード）**データを中川研究室の HP の中にある「Microsoft Excel でする判別分析」をダウンロード。

(2) **（ブックを開く）**そのファイル（DiscriminantAnalysis.xls）を開く。データの中身は以下の通り。

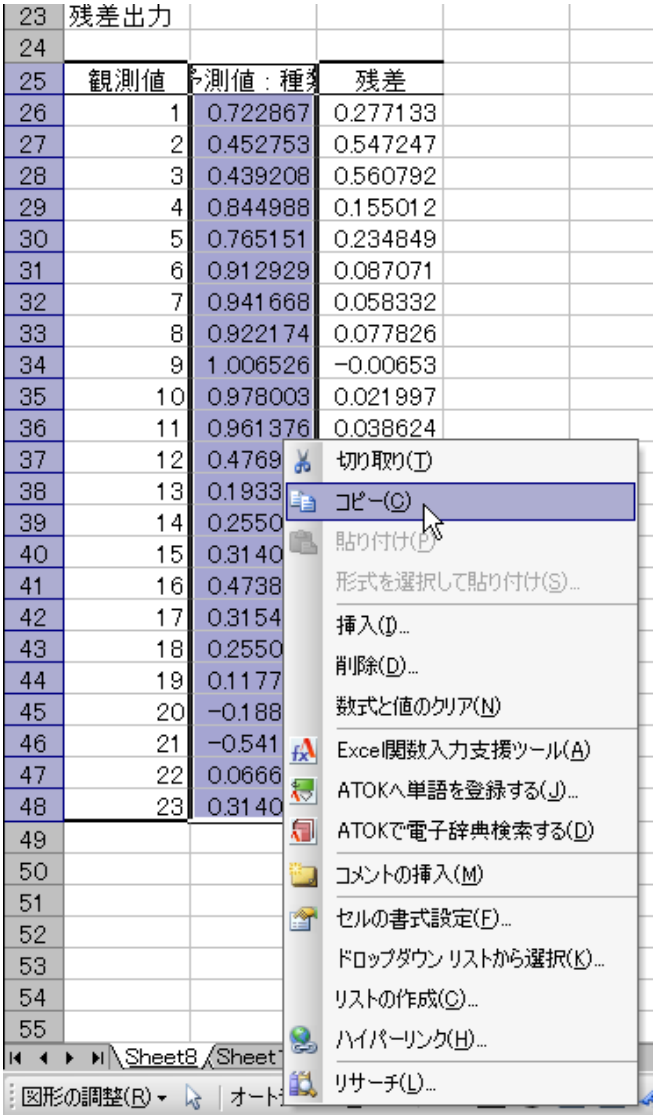
	A	B	C
1	x1	x2	種類
2	6.32	5.24	1
3	5.92	5.12	1
4	5.92	5	1
5	6.44	5.64	1
6	6.4	5.16	1
7	6.56	5.56	1
8	6.64	5.36	1
9	6.68	4.96	1
10	6.72	5.48	1
11	6.76	5	1
12	6.72	5.08	1
13	6	4.88	0
14	5.6	4.64	0
15	5.64	4.96	0
16	5.76	4.8	0
17	5.96	5.08	0
18	5.72	5.04	0
19	5.64	4.96	0
20	5.44	4.88	0
21	5.04	4.44	0
22	4.56	4.04	0
23	5.48	4.2	0
24	5.76	4.8	0

(3) **（回帰分析の実行）**まず回帰分析を行う。被説明変数は、種類。説明変数は x1,x2。このとき、残差を出力するように「残差」のチェックボックスにチェックを入れておく。

(4) **（判別関数）**結果のシートの以下の部分から、判別関数は $-3.92+0.64x_1+0.11x_2$ である。次に判別の閾値を求めよう。

16		係数
17	切片	-3.92239
18	x1	0.641422
19	x2	0.112875

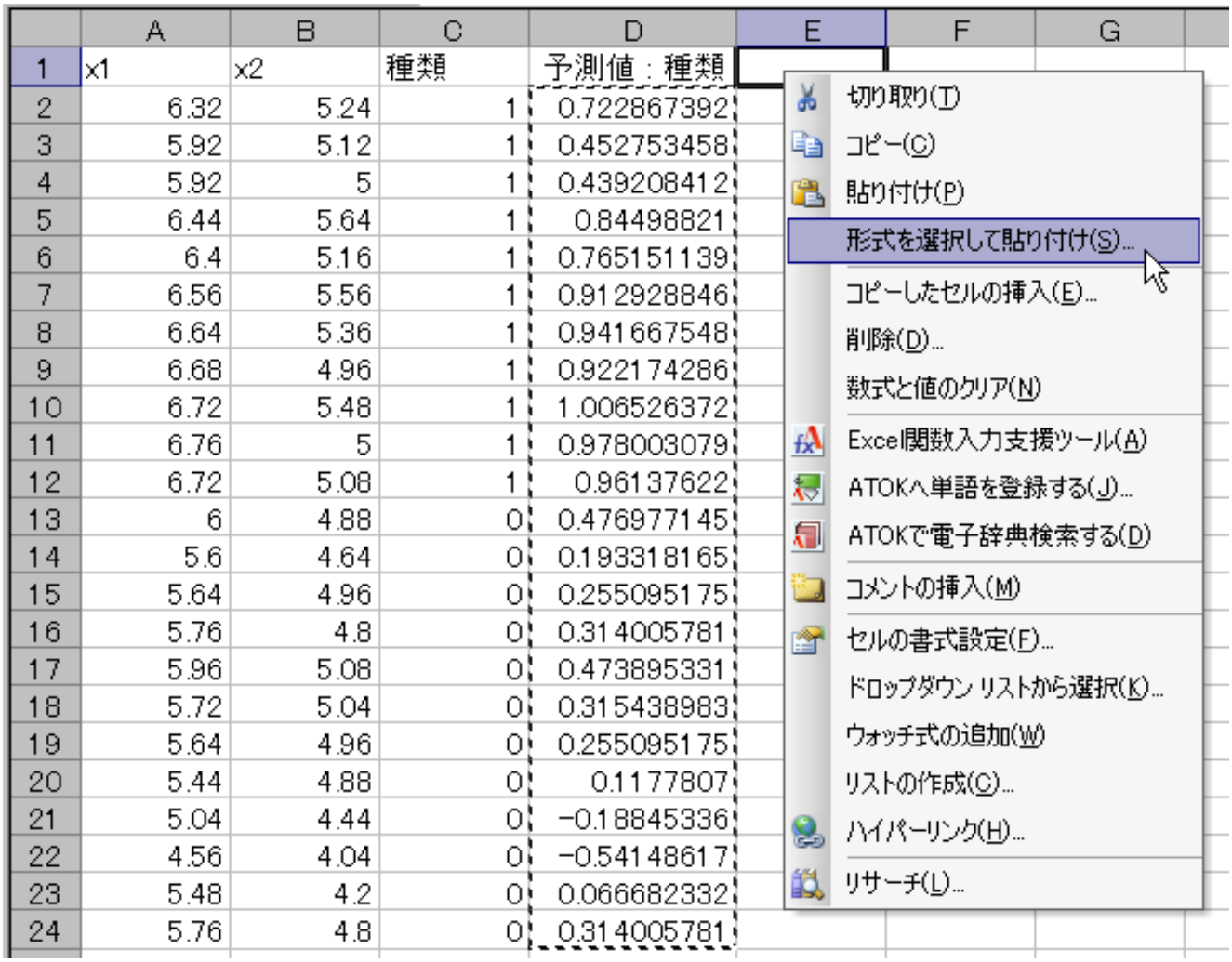
(5) **（予測値のコピー）**次に、結果のシートの予測値の範囲（B25 から B48）をコピーする。コピーは以下のように範囲をドラッグで指定した後、右クリックし、「コピー」をクリックするか、[Ctrl-c]を利用する。



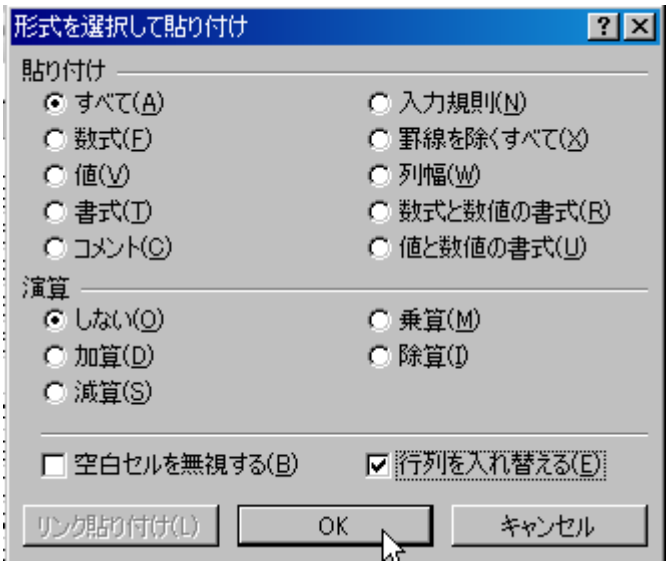
(6) **（予測値のペースト）**次に、もとのデータがあるシート 1 にもどり、D1 以下にペーストする。ペーストは、[Ctrl-v]を利用する。

	A	B	C	D
1	x1	x2	種類	予測値：種類
2	6.32	5.24	1	0.722867392
3	5.92	5.12	1	0.452753458
4	5.92	5	1	0.439208412
5	6.44	5.64	1	0.84498821
6	6.4	5.16	1	0.765151139
7	6.56	5.56	1	0.912928846
8	6.64	5.36	1	0.941667548
9	6.68	4.96	1	0.922174286
10	6.72	5.48	1	1.006526372
11	6.76	5	1	0.978003079
12	6.72	5.08	1	0.96137622
13	6	4.88	0	0.476977145
14	5.6	4.64	0	0.193318165
15	5.64	4.96	0	0.255095175
16	5.76	4.8	0	0.314005781
17	5.96	5.08	0	0.473895331
18	5.72	5.04	0	0.315438983
19	5.64	4.96	0	0.255095175
20	5.44	4.88	0	0.1177807
21	5.04	4.44	0	−0.18845336
22	4.56	4.04	0	−0.541486 17
23	5.48	4.2	0	0.066682332
24	5.76	4.8	0	0.314005781

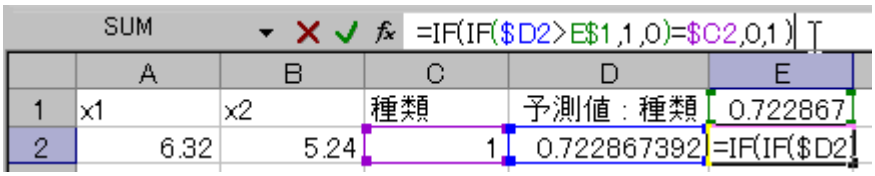
(7) **（予測値を閾値として利用）**D2 から D24 までのデータを判別ルールの閾値として利用するために、E1 からの列にならべる。つまり、1 行目には E 列以降に閾値の候補を入力するのである。そのために、D2,D24 までをドラッグして[Ctrl-c]でコピーし、以下の図のように E1 を右クリックし、「形式を選択して貼り付け」を左クリックする。



(8) **（列データの行データへの変換とペースト）**以下のウインドウが表示されたら、「行列を入れ替える」のチェックボックスを左クリックしてチェックを入れる。その後、「OK」を左クリック。

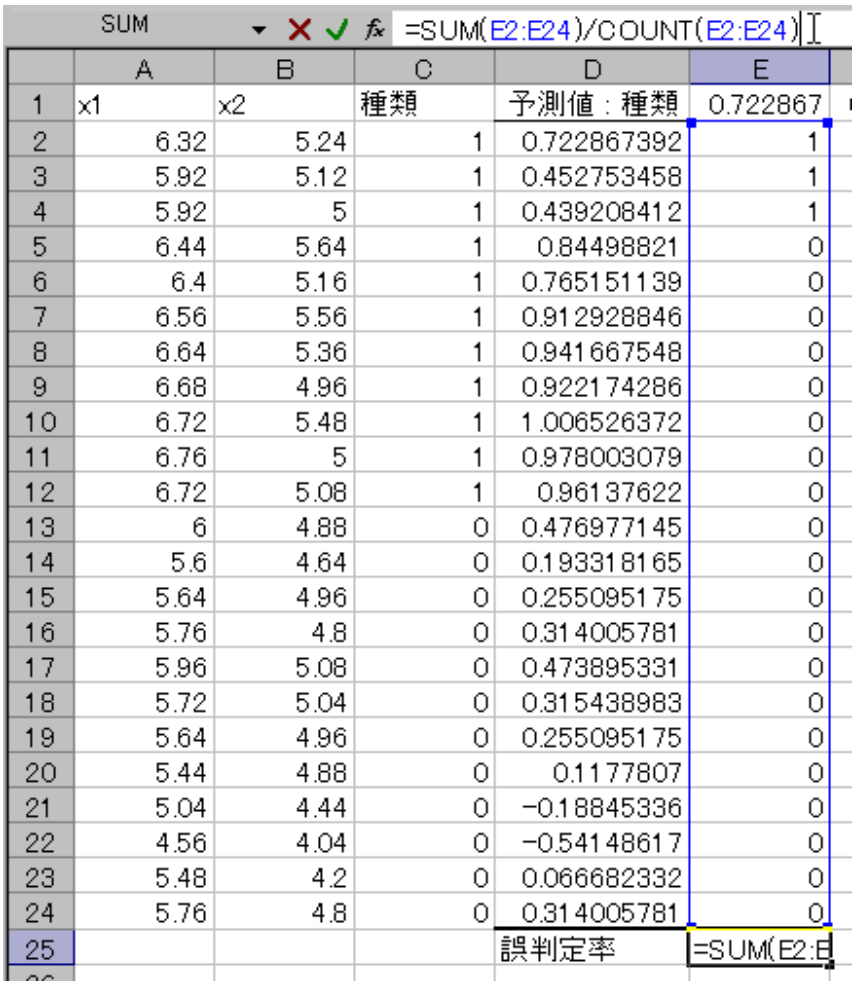


(9) **（判別結果の計算式入力）**それぞれの予測値による判別結果と実際の種類が一致しているかを、一致していれば０、一致していなければ１として E2 以下に入力する。これは誤判別率を計算するために、誤判別を１、そうでない場合を 0 とするのが都合がいいからである。そのために、E2 に「=IF(IF($D2>E$1,1,0)=$C2,0,1)」と入力する。この式の意味は、「IF($D2>E$1,1,0)」によって E1 の閾値をつかって予測値と比較するという判別ルールによる種類の判定結果をあたえ、その結果を「IF(　　=$C2,」によって実際の種類と比較し、「,0,1)」で一致すれば誤判別ではないので前者の 0、一致しなければ誤判別なので 1 を与えよという命令である。なお、$記号は絶対参照の記号で、数式をコピーした先で参照先が変わらないように固定するための記号である。予測値の列（D 列）と実際の種類の列（C 列）、閾値のある行（1）は固定しなければならないので、それぞれ「$D」、「$C」、「$1」で絶対参照としている。



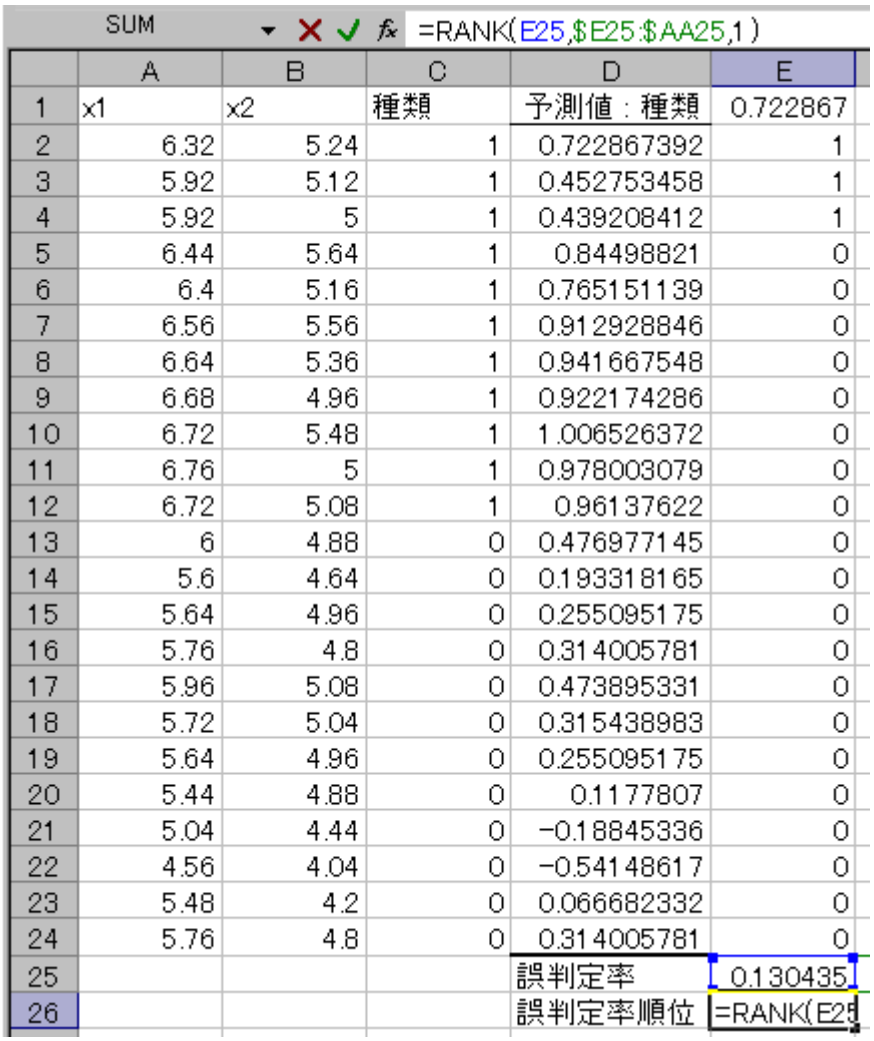
(10)　　**（判別結果の計算）**E2 をオートフィルで E2 から E24 へコピーする。これによってすべての観測値についてそれが誤判別かどうかが１，０で入力できた。

(11)　　**（誤判別確率式の入力）**25 行目には誤判別確率を入力する。D25 に「誤判定率」と入力し、E25 に「=SUM(E2:E24)/COUNT(E2:E24)」と入力する。この式の分子は、誤判別数を合計し(SUM)、分母は観測個数を数える(COUNT)。



	A	B	C	D	E
1	x1	x2	種類	予測値：種類	0.722867
2	6.32	5.24	1	0.722867392	1
3	5.92	5.12	1	0.452753458	1
4	5.92	5	1	0.439208412	1
5	6.44	5.64	1	0.84498821	0
6	6.4	5.16	1	0.765151139	0
7	6.56	5.56	1	0.912928846	0
8	6.64	5.36	1	0.941667548	0
9	6.68	4.96	1	0.922174286	0
10	6.72	5.48	1	1.006526372	0
11	6.76	5	1	0.978003079	0
12	6.72	5.08	1	0.96137622	0
13	6	4.88	0	0.476977145	0
14	5.6	4.64	0	0.193318165	0
15	5.64	4.96	0	0.255095175	0
16	5.76	4.8	0	0.314005781	0
17	5.96	5.08	0	0.473895331	0
18	5.72	5.04	0	0.315438983	0
19	5.64	4.96	0	0.255095175	0
20	5.44	4.88	0	0.1177807	0
21	5.04	4.44	0	-0.18845336	0
22	4.56	4.04	0	-0.541486117	0
23	5.48	4.2	0	0.066682332	0
24	5.76	4.8	0	0.314005781	0
25				誤判定率	=SUM(E2:E

(12)　**（誤判別確率順位計算式の入力）**26 行目には誤判定率順位を入力する。D26 に「誤判定率順位」と入力し、E26 に「=RANK(E25,$E25:$AA25,1)」と入力する。この式は、E25 から AA25 までの間（$E25:$AA25）の中で E25 の小さい順での(1)順位を計算せよというものである。これで、どの閾値のときの誤判別確率が最小かを見つけやすくなる。なお、絶対参照は、順位を決める範囲（この場合は 25 行目の E 列から AA 列まで）を固定するために、「$E」、「$AA」という形で使用されている。



SUM　=RANK(E25,$E25:$AA25,1)

	A	B	C	D	E
1	x1	x2	種類	予測値：種類	0.722867
2	6.32	5.24	1	0.722867392	1
3	5.92	5.12	1	0.452753458	1
4	5.92	5	1	0.439208412	1
5	6.44	5.64	1	0.84498821	0
6	6.4	5.16	1	0.765151139	0
7	6.56	5.56	1	0.912928846	0
8	6.64	5.36	1	0.941667548	0
9	6.68	4.96	1	0.922174286	0
10	6.72	5.48	1	1.006526372	0
11	6.76	5	1	0.978003079	0
12	6.72	5.08	1	0.96137622	0
13	6	4.88	0	0.476977145	0
14	5.6	4.64	0	0.193318165	0
15	5.64	4.96	0	0.255095175	0
16	5.76	4.8	0	0.314005781	0
17	5.96	5.08	0	0.473895331	0
18	5.72	5.04	0	0.315438983	0
19	5.64	4.96	0	0.255095175	0
20	5.44	4.88	0	0.1177807	0
21	5.04	4.44	0	-0.18845336	0
22	4.56	4.04	0	-0.541486 17	0
23	5.48	4.2	0	0.066682332	0
24	5.76	4.8	0	0.314005781	0
25				誤判定率	0.130435
26				誤判定率順位	=RANK(E25

(13)　**（判別結果、誤判別確率、誤判別確率順位の計算）**E2 から E26 までを E 列から AA 列（データの最後の列）までオートフィルでコピーする。

P	Q	R	S	T	U
0.476977	0.193318	0.255095	0.314006	0.473895	0.315439
0	0	0	0	0	0
1	0	0	0	1	0
1	0	0	0	1	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	1	1	1	1	1
0	0	0	0	0	0
0	1	0	0	0	0
0	1	1	0	0	0
0	1	1	1	0	1
0	1	1	1	0	0
0	1	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	1	1	0	0	0
0.086957	0.304348	0.217391	0.130435	0.130435	0.086957
1	14	10	3	3	1

(14)　**（閾値の探索）**最後の行の順位 1 を探す。これにより、閾値は 0.476977 か 0.315439 であることがわかる。（それ以外の値も可能性があるが、とりあえずこれでいいでしょう。）

練習問題

中川研究室の HP から判別分析練習問題のデータをダウンロードする。このデータは、ホーエルの入門数理統計学の p.187 の問題からとったデータで、IT は知能検査の点数、RR は読書力の点数である。ある試験の合格に関しては、1 が合格、0 が不合格として表されている。まず、IT、RR の両方をつかって判別分析を行え。次に、IT のみをつかって判別分析をおこない、最後に RR のみをつかって判別分析をおこない、それぞれの誤判別確率を比較し、最小の誤判別確率の場合はどの変数をつかって判別分析を行った場合か調べよ。