THE FACTOR ANALYSIS BASIC TOUR

FACTOR PATTERNS IN CORRELATIONAL DATA

Note: Assume that r's of .3 or .4 = variables go together

Matrix 1

	1	2	3	4	5	6
1	-	72	62	58	74	82
2		-	53	62	47	58
3			-	63	71	58
4				-	53	54
5					-	68
6						-

Matrix 2

	1	2	3	4	5	6
1	-	17	56	72	21	03
2		-	14	03	53	64
3			-	63	01	09
4				-	17	09
5					-	72
6						-

Matrix 3

	1	2	3	4	5	6
1	-	03	19	02	11	07
2		-	53	63	16	01
3			-	51	08	20
4				-	22	08
5					-	17
6						-

Matrix 4

	1	2	3	4	5	6
1	-	03	72	01	19	15
2		-	17	03	14	59
3			-	20	08	13
4				-	68	09
5					-	17
6						-

Matrix 5

	1	2	3	4	5	6
1	-	03	07	02	13	09
2		-	21	17	09	17
3			-	03	14	07
4				-	02	09
5					-	16
6						-

So, here are some data from 4 variables with n=10

MTB > prin c16-c19 <--- Cols in MTB where data are

ROW	M1	M2	M3	M4
1	36	41	45	45
2	52	58	57	41
3	51	51	48	65
4	47	46	40	63
5	56	49	48	36
6	66	72	65	56
7	48	48	44	43

```
   8    61    55    50    70
   9    50    39    50    55
  10    33    40    37    46
 MTB > corr c16-c19
              M1        M2        M3
M2         0.805
M3         0.782     0.843
M4         0.329     0.178     0.055
```

If you examine the intercorrelations, M1-M3 correlate highly with each other but, M4 seems not to correlate with M1-M3.

A FACTOR IS A RULE THAT COMBINES THE VARIABLES TOGETHER IN SOME LINEAR FASHION

Note Assume that we define Factor A as .8M1 + .7M2 + .8M3 and give 0 weight to M4 ... since it seems not to fit.

```
MTB > let c20=(.8*c16)+(.7*c17)+(.8*c18)
MTB > prin c16-c20
 ROW    M1    M2    M3    M4       A

   1    36    41    45    45    93.5
   2    52    58    57    41   127.8
   3    51    51    48    65   114.9
   4    47    46    40    63   101.8
   5    56    49    48    36   117.5
   6    66    72    65    56   155.2
   7    48    48    44    43   107.2
   8    61    55    50    70   127.3
   9    50    39    50    55   107.3
  10    33    40    37    46    84.0
```

THE VALUES IN A ... THE RESULTS OF APPLYING THE FACTOR RULE ... ARE CALLED <u>FACTOR SCORES</u>.

How does our set of <u>factor scores</u> correlate back with each of the original measures ... M1-M4?

```
MTB > corr c16-c20

              M1        M2        M3        M4
M2         0.805
M3         0.782     0.843
M4         0.329     0.178     0.055
A          0.933     0.942     0.928     0.211
```

Well, the way we defined Factor A and the resulting r of the factor A scores and M1-M4 shows that our definiton worked well for M1-M3 but, didn't do much for predicting or explaining M4. Residuals give you some idea about how good the "fit" was (Factor A rule and factor scores TO the data ... M1-M4). But, to see those, you have to do regressions and then examine the residuals. The A factor scores variable is used to predict M1-M4 ... then residuals are examined.

```
MTB > regr c16 1 c20;
SUBC> resi c21. <---- leftover M1 after Fact A scores removed

MTB> regr c17 1 c20;
SUBC> resi c22. <--- leftover M2 after Fact A scores removed

MTB > regr c18 1 c20;
SUBC> resi c23. <--- leftover M3 after Fact A scores removed

MTB > regr c19 1 c20;
SUBC> resi c24.   <---- leftover M4 after Fact A scores removed

MTB > name c21='Er1A' c22='Er2A' c23='Er3A' c24='Er4A'
MTB > prin c21-c24
```

ROW	Er1A	Er2A	Er3A	Er4A
1	-4.56888	0.52419	4.11021	-4.5705
2	-4.62285	1.48202	3.32608	-12.7061
3	0.41494	0.51537	-0.86589	12.8493
4	2.54634	1.64227	-3.98333	12.4288
5	4.19802	-2.70065	-1.83495	-16.4642
6	-3.44730	2.66698	1.11369	-1.0098
7	1.01890	1.11668	-1.99599	-8.2223
8	4.61118	-1.28413	-3.48756	16.3542
9	2.97209	-7.93010	3.96674	3.7656
10	-3.12245	3.96736	-0.34900	-2.4250

Above is called first residual matrix. Note that the errors are relatively smallER using factor A scores to predict M1-M3 but, the errors are much larger for M4. This means that when we applied the Factor A rule, not that much was left over (residuals) when partialing it from M1-M3 but, substantial residuals were there when partialing it from M4. **BACK TO THE DRAWING BOARD FOR M4!!**

We now need to start the cycle over again but, remember, we now have a first residual matrix so ... this is the point at which we need to work. Taking Factor A out of the original M1-N4 data matrix left us with the first set of residuals: how much of THAT can we now explain or eliminate by now defining a second Factor B?

Working with the first residual matrix, let's define Factor B as: .7 * Er4A

```
 MTB > let c25=.7*c24
```

Now we use new FACTOR scores to predict first residual matrix

```
 MTB > regr c21 1 c25;
 SUBC> resi c26. <--- leftover for resid M1 after Fact B taken
out

 MTB > regr c22 1 c25;
 SUBC> resi c27. <--- leftover for resid M2 after Fact B removed

 MTB > regr c23 1 c25;
 SUBC> resi c28. <-- leftover for resid M3 after Fact B removed

 MTB > regr c24 1 c25; <- leftover for resid M4 after Fact B
taken
 SUBC> resi c29.

 MTB > name c26='Er1B' c27='Er2B' c28='Er3B' c29='Er4B'
```

How do the factor scores from the Factor B rule related to the first residual matrix?

```
 MTB > corr c21-c25
```

	Er1A	Er2A	Er3A	Er4A
Er2A	-0.601			
Er3A	-0.624	-0.249		
Er4A	0.373	-0.065	-0.388	
C25	0.373	-0.065	-0.388	1.000

Note that our Factor B rule and factor scores don't correlate too well with residuals on M1-M3 but perfectly (and this clearly is not usually the case) with the residuals on M4. What about the second residual matrix?

```
 MTB > prin c26-c29
```

ROW	Er1B	Er2B	Er3B	Er4B
1	-4.01530	0.43479	3.63486	0
2	-3.08386	1.23347	2.00458	0
3	-1.14138	0.76672	0.47050	0
4	1.04095	1.88539	-2.69066	0
5	6.19220	-3.02272	-3.54732	0
6	-3.32499	2.64723	1.00866	0
7	2.01480	0.95584	-2.85115	0
8	2.63033	-0.96422	-1.78664	0
9	2.51599	-7.85643	4.35839	0
10	-2.82872	3.91993	-0.60121	0

So, how do the Factor A scores correlate with the variables, M1-M4? On the second page, you saw how A correlated with M1-M4. On the page above, you see how Factor B scores correlate with the first residual matrix. We can put this information in a table.

	Fac A	Fac B
M1	.933	.373
M2	.942	-.065
M3	.928	-.388
M4	.211	1.00

CORRELATIONS BETWEEN FACTOR SCORES AND VARIABLES OR MEASURES ARE CALLED FACTOR LOADINGS.

For our data, a two factor solution seems reasonable and, the variables M1-M3 load highly on Factor A, but M4 does not ... while M4 loads highly on Factor B ... but M1-M3 don't. Keep in mind of course that we could have defined the factors differently and the numbers we got above would not be the same.