Psychology 205: Research Methods in Psychology Advanced Statistical Procedures Data= Model + Residual

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

November, 2016

Outline

The basic data frame

Table: The basic data frame organizes data by subjects (rows) and variables (columns)

Subject	DV	IV_1	IV_2	IV_3	SV_1	SV_2	SV_2	CV_1	CV_2	...	CV_n
1	Y_1	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	...	X_{1n}
2	Y_2	X_{21}	X_{22}	X_{23}	X_{24}	X_{25}	X_{26}	X_{27}	X_{28}	...	X_{2n}
...	...	...	...	...	...	...	...	...	...	...	...
N	Y_N	X_{N1}	X_{N2}	X_{N3}	X_{N4}	X_{N5}	X_{N6}	X_{N7}	X_{N8}	...	X_{Nn}

Preliminary Steps – see prior handouts

1. Make sure that the psych package is active library(psych)
2. Read in the data
    - Copy to the clipboard
    - my.data <- read.clipboard()
3. Describe the data
    - `describe(my.data)`
4. Multivariate plots to examine the data more carefully
    - `pairs.panels(my.data=[2:9])` #specify which columns to plot

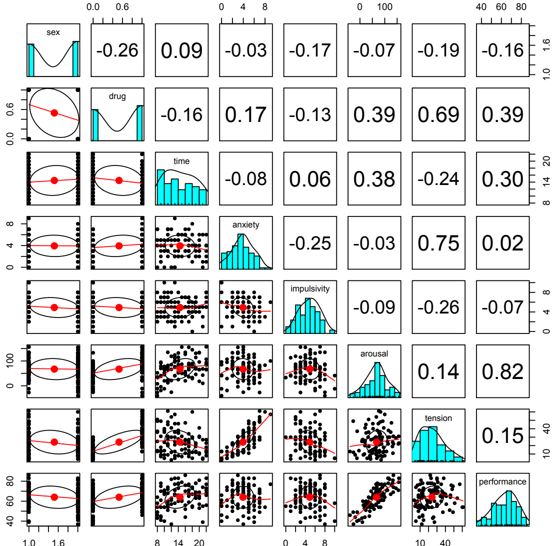
Descriptive statistics using describe

```
> library(psych)
> my.data <- read.clipboard()
> describe(my.data)
```

	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
snum	1	100	50.50	29.01	50.5	50.50	37.06	1	100	99	0.00	-1.24	2.90
sex	2	100	1.52	0.50	2.0	1.52	0.00	1	2	1	-0.08	-2.01	0.05
drug	3	100	0.53	0.50	1.0	0.54	0.00	0	1	1	-0.12	-2.01	0.05
time	4	100	14.47	4.29	14.0	14.34	5.93	8	22	14	0.18	-1.19	0.43
anxiety	5	100	3.94	1.97	4.0	3.96	1.48	0	9	9	-0.03	-0.41	0.20
impulsivity	6	100	4.90	2.04	5.0	4.89	2.97	0	10	10	0.04	-0.45	0.20
arousal	7	100	67.15	42.59	69.5	67.64	38.55	-40	157	197	-0.13	-0.20	4.26
tension	8	100	24.14	14.05	23.0	22.95	14.83	3	61	58	0.63	-0.34	1.41
performance	9	100	63.99	11.19	66.0	64.50	11.86	37	86	49	-0.33	-0.54	1.12
cost	10	100	1.00	0.00	1.0	1.00	0.00	1	1	0	NaN	NaN	0.00

A Scatter Plot Matrix (SPLOM) plot

pairs.panels(my.data[2:9]) #omit the first and last variables



Types of models

1. Y = bX (X is continuous) Regression
2. Y = bX (X has two levels) t-test
3. Y = bX (X has $>$ 2 levels) F-test
4. $Y = b_1X_1 + b_2X_2 + b_3X_3$ (X_i is continuous) Multiple regression
5. $Y = b_1X_1 + b_2X_2 + b_3X_{12}$ (X_i is continuous) Multiple regression with an interaction term
   - In this case, we need to zero center the X_i so that the product is independent of the Xs.
6. $Y = b_1X_1 + b_2X_2 + b_3X_{12}$ (X_i is is categorical) Analysis of Variance
7. $Y = b_1X_1 + b_2X_2 + b_3X_{12} + Z$ (X_i and Z are continuous) Analysis of Covariance

The General Linear Model

```
model = lm(y ~ x1 + x2 + x1*x2,data=my.data)
```

But the product term is correlated with X_1 and X_2 and so we need to zero center (subtract out the mean) from the predictors.

```
cen.data.df <- data.frame(scale(my.data,scale=FALSE))
model = lm(y ~ x1 + x2 + x1*x2,data=cen.data.df)
summary(model)  #to show the results
```

Analysis of Variance

If X_i are really categorical, we can make them into "factors" to do the ANOVA

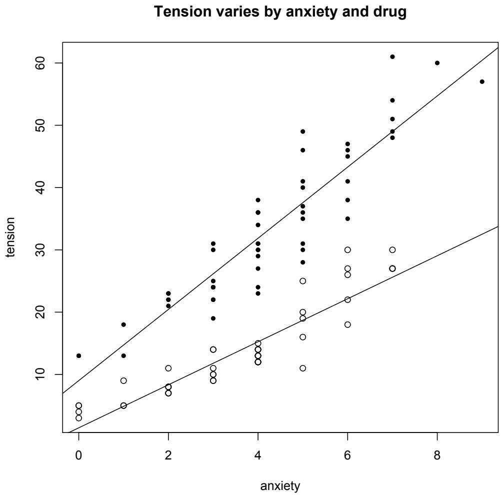
```
X1cat  <- as.factor(my.data$X1)
x2cat <- as.factor(my.data$X2)
model <- aov(my.data$Y ~ X1cat + x2cat + X1cat*X2cat)
summary(model)  #to show the results
print(model.tables(model, 'means'),digits=2)
```

A simple multiple regression

```
> model <- lm(tension~drug + anxiety,data=my.data)
> summary(model)


Call:
lm(formula = tension ~ drug + anxiety, data = my.data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3561 -2.7237 -0.6611  3.0246 14.9750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.8057     1.1018  -2.546   0.0125 *
drug         16.4042     0.9480  17.304   <2e-16 ***
anxiety       4.6324     0.2409  19.226   <2e-16 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 4.667 on 97 degrees of freedom
Multiple R-squared:  0.8919,        Adjusted R-squared:  0.8897
F-statistic: 400.4 on 2 and 97 DF,  p-value: < 2.2e-16
```

Plot this result

```
> with(my.data,plot(tension ~ anxiety,pch=21-drug,
     main= 'Tension varies by anxiety and drug'))
> by(my.data,my.data$drug,function(x) abline(lm(tension ~anxiety,data=x)))
```



Multivariate Analysis

1. Suppose we have multiple predictors and we want to understand their structure.
2. We can find the sum of all the predictors to get a total score, or we can find the sum of some subset of predictors to get total scores on subsets or factors of the data.
3. How many factors are there in the data?

Factor Analysis and Principal Components

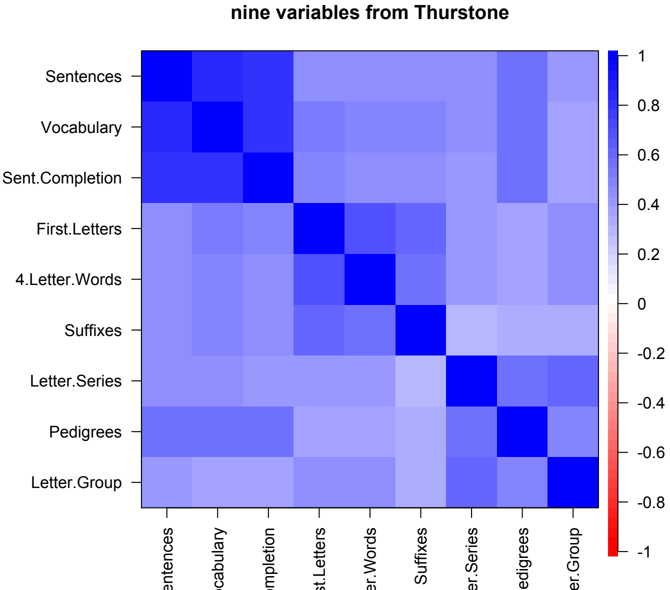
1. Trying to approximate a data matrix or a correlation matrix with one of "lower rank"
   - The data are a matrix of N x n but the rank of the matrix is the smaller (n)
   - Can we approximate this with a matrix of N x k where $k < n$
2. $R = FF' + U^2$ Factor analysis
   - F is the matrix of factor "loadings" or correlations between the variable and the latent factors
   - U^2 is a fudge factor to account for the residual variance
3. $R = CC'$ (The components model).

9 Mental Tests from Thurstone (built into the psych package as demonstration

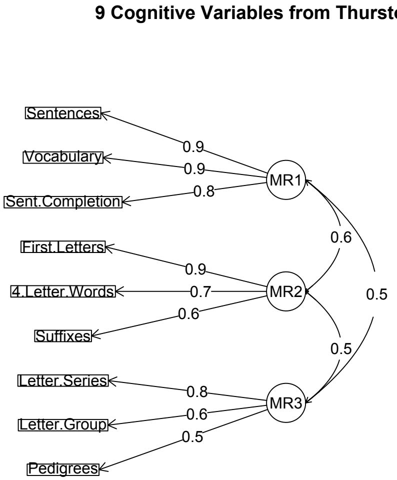
```
> lowerMat(Thurstone)
```

	Sntnc	Vcblr	Snt.C	Frs.L	4.L.W	Sffxs	Ltt.S	Pdgrs	Ltt.G
Sentences	1.00								
Vocabulary	0.83	1.00							
Sent.Completion	0.78	0.78	1.00						
First.Letters	0.44	0.49	0.46	1.00					
4.Letter.Words	0.43	0.46	0.42	0.67	1.00				
Suffixes	0.45	0.49	0.44	0.59	0.54	1.00			
Letter.Series	0.45	0.43	0.40	0.38	0.40	0.29	1.00		
Pedigrees	0.54	0.54	0.53	0.35	0.37	0.32	0.56	1.00	
Letter.Group	0.38	0.36	0.36	0.42	0.45	0.32	0.60	0.45	1.00

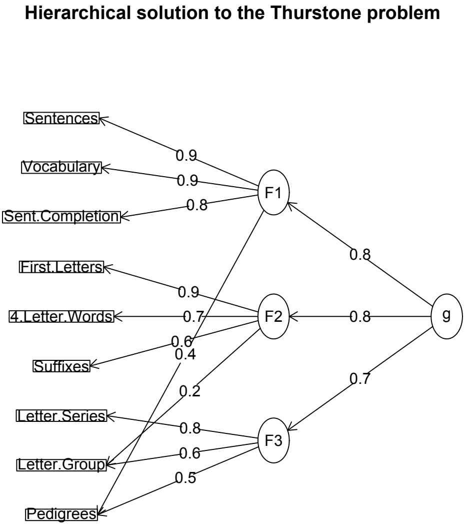
9 Cognitive variables from Thurstone; cor.plot(thurstone)



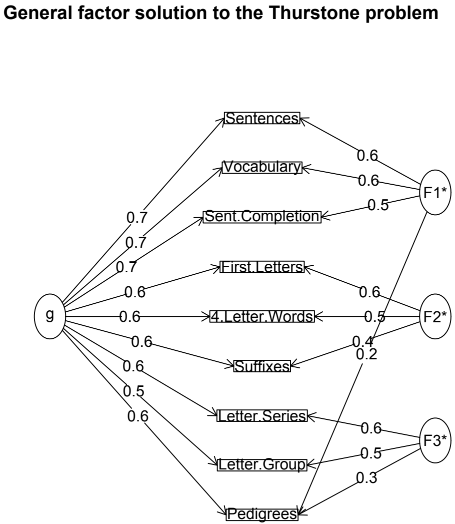
3 factors of the Thurstone variables: f3 <- fa(Thurstone,3)



A hierarchical representation of the solution



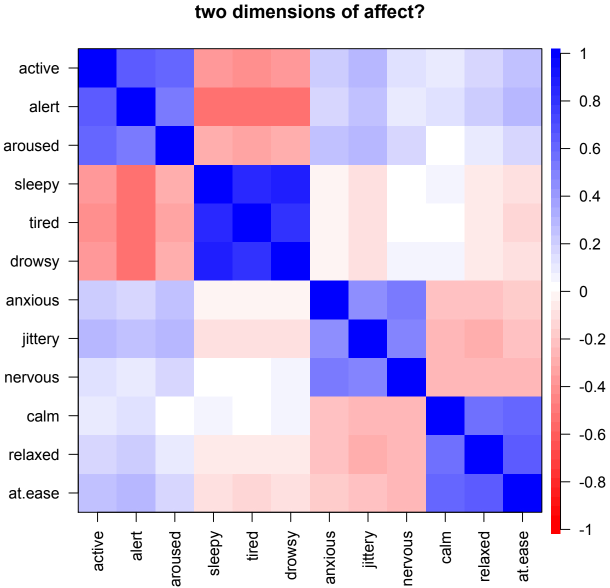
A general factor representation of the solution



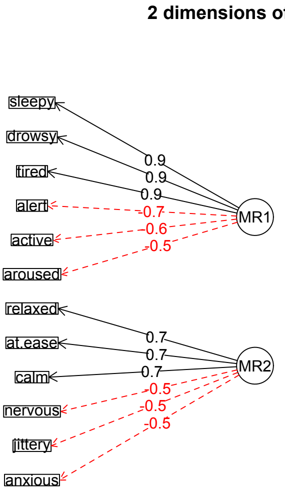
frame

```
> EA.TA <- msq[c("active","alert","aroused","sleepy","tired",
       "drowsy","anxious","jittery","nervous","calm","relaxed","at.ease")]
>  affect <- lowerCor(EA.TA)

        activ alert arosd slepy tired drwsy anxis jttry nervs calm  relxd at.es
active   1.00
alert    0.62  1.00
aroused  0.60  0.53  1.00
sleepy  -0.40 -0.53 -0.33  1.00
tired   -0.42 -0.53 -0.35  0.81  1.00
drowsy  -0.39 -0.53 -0.32  0.85  0.78  1.00
anxious  0.19  0.17  0.22 -0.04 -0.05 -0.03  1.00
jittery  0.27  0.23  0.29 -0.12 -0.12 -0.11  0.45  1.00
nervous  0.11  0.09  0.17  0.02  0.01  0.02  0.51  0.47  1.00
calm     0.06  0.11  0.01  0.03  0.01  0.05 -0.25 -0.28 -0.27  1.00
relaxed  0.16  0.18  0.09 -0.07 -0.08 -0.07 -0.24 -0.30 -0.28  0.54  1.00
at.ease  0.23  0.28  0.15 -0.12 -0.14 -0.10 -0.19 -0.22 -0.27  0.58  0.61  1.00
```

Show the correlations graphically



Show the factor structure



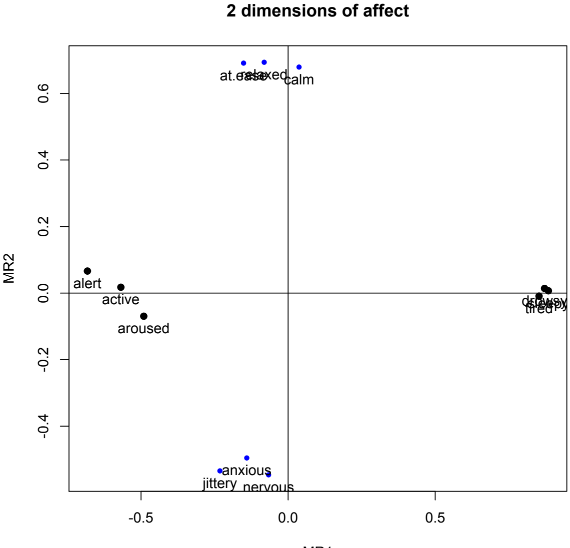
Examine the factor structure

```
f2 <- fa(EA.TA,2)
f2
```

```
Call: fa(r = EA.TA, nfactors = 2)
Standardized loadings (pattern matrix) based upon correlation matrix
          MR1   MR2   h2   u2 com
active  -0.57  0.02 0.32 0.68 1.0
alert   -0.68  0.07 0.47 0.53 1.0
aroused -0.49 -0.07 0.24 0.76 1.0
sleepy   0.88  0.01 0.78 0.22 1.0
tired    0.85 -0.01 0.73 0.27 1.0
drowsy   0.87  0.01 0.76 0.24 1.0
anxious -0.14 -0.50 0.26 0.74 1.2
jittery -0.23 -0.53 0.33 0.67 1.4
nervous -0.07 -0.55 0.30 0.70 1.0
calm     0.04  0.68 0.46 0.54 1.0
relaxed -0.08  0.69 0.49 0.51 1.0
at.ease -0.15  0.69 0.51 0.49 1.1

                       MR1  MR2
SS loadings           3.40 2.26
Proportion Var        0.28 0.19
Cumulative Var        0.28 0.47
Proportion Explained  0.60 0.40
Cumulative Proportion 0.60 1.00

 With factor correlations of
      MR1   MR2
MR1  1.00 -0.06
MR2 -0.06  1.00
```

Yet another 2 dimensional plot fa

plot(f2,title="2 dimensions of affect",labels=colnames(EA.TA))



Prepare the scoring keys matrix

```
> keys <- make.keys(EA.TA,list(EA=c("alert","active","aroused","-drowsy",
         "-tired","-sleepy"),
          TA=c("anxious","jittery","nervous","-calm","-relaxed", "-at.ease")))
> keys
```

```
        EA TA
active   1  0
alert    1  0
aroused  1  0
sleepy  -1  0
tired   -1  0
drowsy  -1  0
anxious  0  1
jittery  0  1
nervous  0  1
calm     0 -1
relaxed  0 -1
at.ease  0 -1
```

The EA.TA scores

```
ea.ta.scores <- score.items(keys,EA.TA)
ea.ta.scores

Call: score.items(keys = keys, items = EA.TA)

(Unstandardized) Alpha:
        EA   TA
alpha 0.87 0.75

Average item correlation:
            EA   TA
average.r 0.54 0.34

 Guttman 6* reliability:
          EA   TA
Lambda.6 0.9 0.77

Scale intercorrelations corrected for attenuation
 raw correlations below the diagonal,
 alpha on the diagonal
 corrected correlations above the diagonal:
       EA     TA
EA  0.874 -0.021
TA -0.017  0.751
```

```
Item by scale correlations:
 corrected for item overlap and scale reliability
           EA    TA
active   0.65 -0.02
alert    0.73 -0.07
aroused  0.56  0.07
sleepy  -0.84  0.02
tired   -0.82  0.03
drowsy  -0.83  0.01
anxious  0.06  0.36
jittery  0.25  0.53
nervous  0.06  0.55
calm     0.01 -0.66
relaxed  0.14 -0.69
at.ease  0.22 -0.69
```

The output from scoreItems

```
 names(ea.ta.scores)
describe(ea.ta.scores$scores)
```

```
 names(ea.ta.scores)
 [1] "scores"          "missing"         "alpha"           "av.r"            "sn"              "n.items"
 [7] "item.cor"        "cor"             "corrected"       "G6"              "item.corrected"  "response.freq"
[13] "raw"             "ase"             "Call"
```

```
   vars    n mean   sd median trimmed  mad min max range skew kurtosis   se
EA    1 3896 1.35 0.76   1.33    1.35 0.74   0   3     3 0.00    -0.81 0.01
TA    2 3896 0.91 0.55   0.83    0.88 0.49   0   3     3 0.61     0.44 0.01
```