Ordinal data, matrix algebra & factor analysis

Sarah Medland – Boulder 2008

Thursday morning

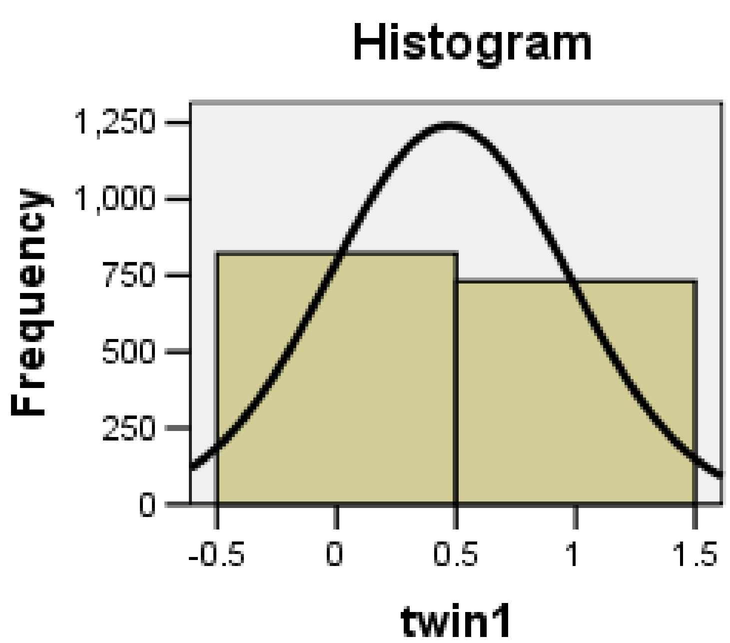
 

This morning

- Fitting the regression model with ordinal data
- Factor Modelling
  - Continuous
  - Ordinal

Binary Data… 1 variable

- Thresholds T ; $[t_{11}]$



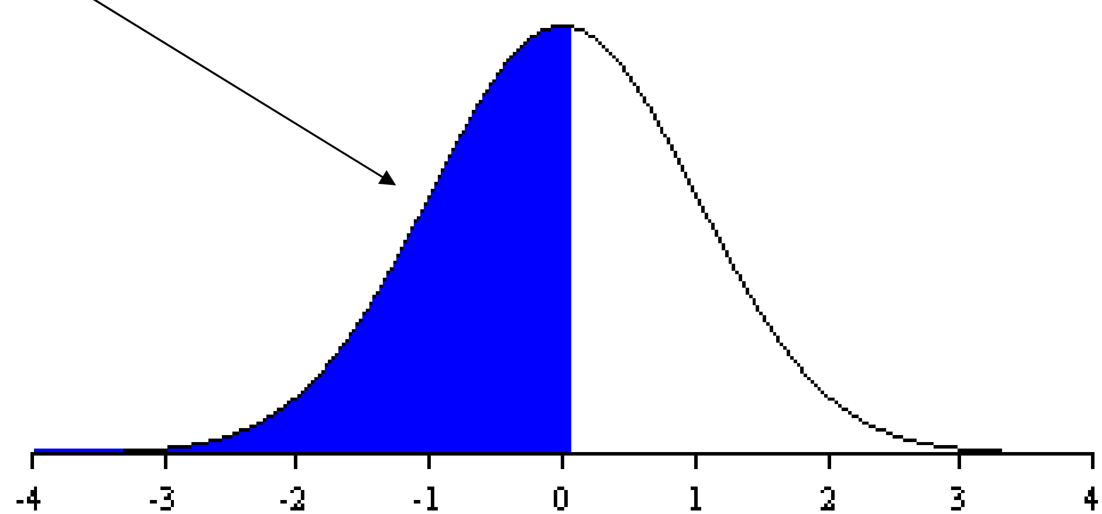
Standard normal distribution

Mean = 0

SD =1

Non Smokers =53%

Threshold =.074



Binary Data… adding a regression

- Thresholds T + D*B ;

$$=[t_{11}]+\begin{bmatrix} Age \\ Sex \end{bmatrix}*\begin{bmatrix}\beta_{age} & \beta_{sex}\end{bmatrix}$$

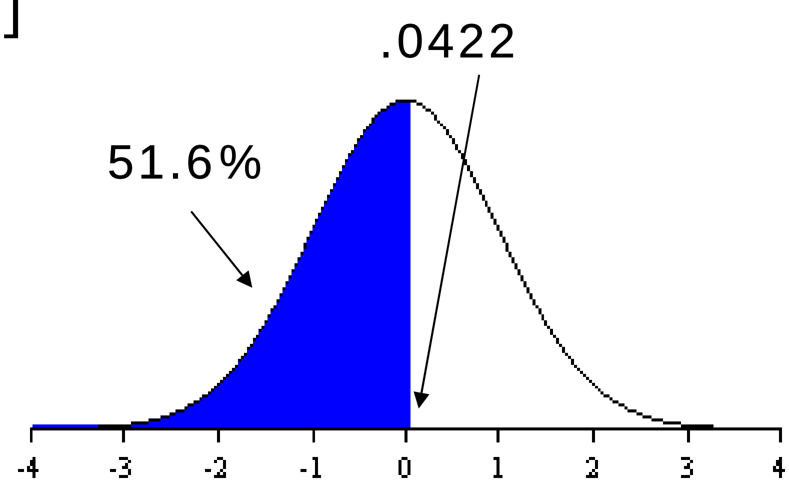
$$=\left[t_{11}+Age*\beta_{age}+Sex*\beta_{sex}\right]$$

$$=\left[-.1118+Age*.007+Sex*-.050\right]$$

if Age = 22 and Sex =1 (Male)

$$=\left[-.1118+(22*.007)+(1*-.050)\right]$$

$$=[.0422]$$

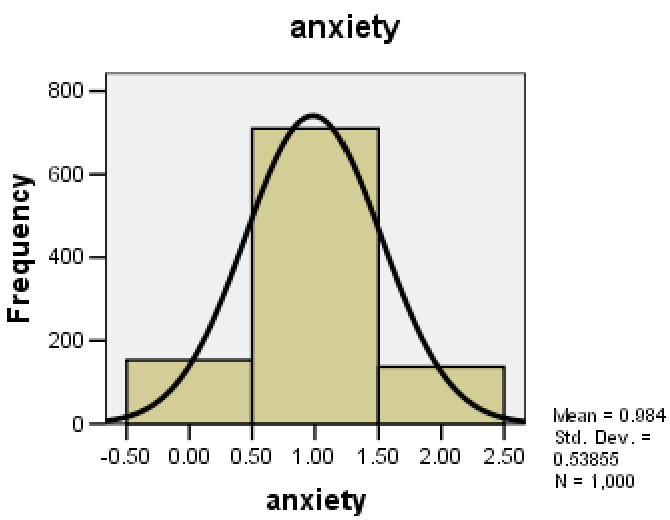


What about more than 2 categories?

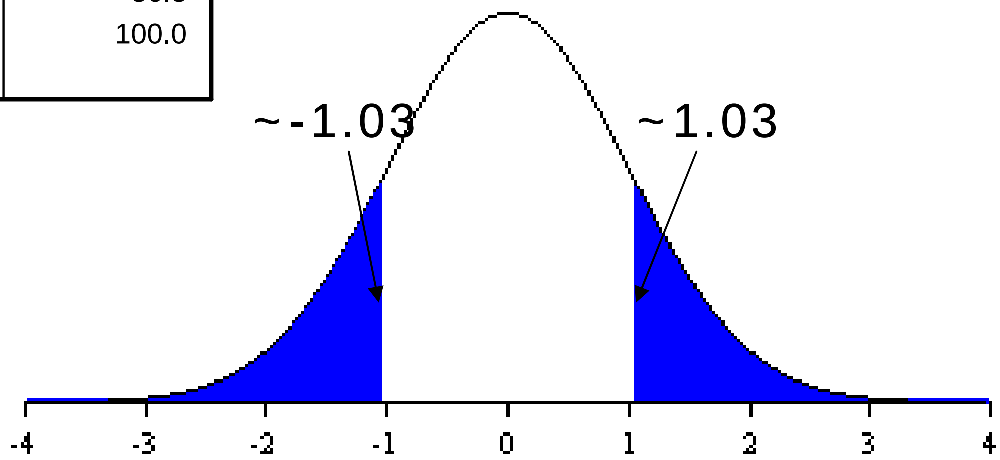
- Thresholds = L*T;

anxiety

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	153	15.3	15.3	15.3
	1.00	710	71.0	71.0	86.3
	2.00	137	13.7	13.7	100.0
	Total	1000	100.0	100.0	



~15% in each tail
Thresholds:



Displacement = ~2.06

What about more than 2 categories?

- Thresholds = L*T;

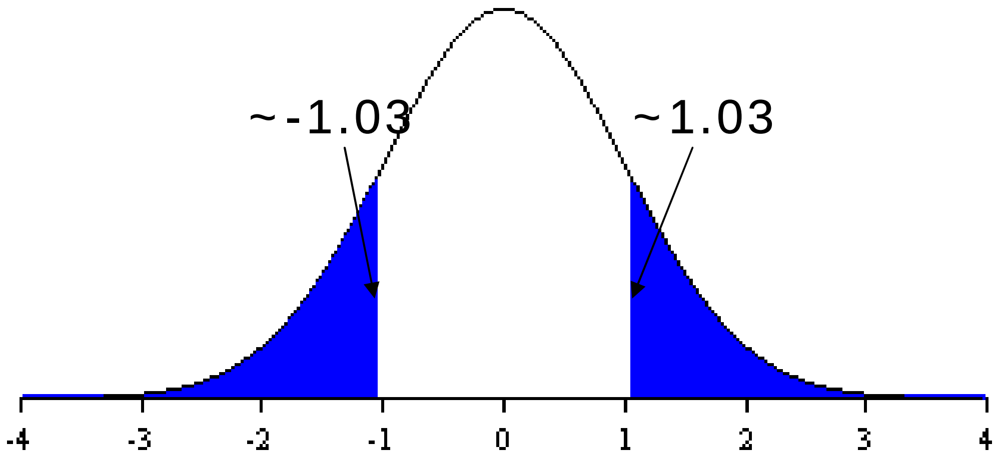
$$= \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} * \begin{bmatrix} t_{11} \\ t_{21} \end{bmatrix}$$

$$= \begin{bmatrix} 1*t_{11} + 0*t_{21} \\ 1*t_{11} + 1*t_{21} \end{bmatrix}$$

$$= \begin{bmatrix} -1.03 \\ -1.03 + 2.06 \end{bmatrix}$$

$$= \begin{bmatrix} -1.03 \\ 1.03 \end{bmatrix}$$

~15% in each tail
Thresholds:



Displacement = ~2.06

Adding a regression

- L*T + G@(D*B);

```
T Full maxth nthr Free   ! Thresholds
B Full nvar ndef Free    ! Regression betas
L lower maxth maxth      ! For converting incremental to cumulative thresholds
G Full maxth 1           ! For duplicating regression betas across thresholds
D Full ndef nsib         ! Contains definition variables
```

- maxth =2, ndef=2, nsib=1, nthr=2

$$G = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad D = \begin{bmatrix} sex \\ age \end{bmatrix} \quad B = \begin{bmatrix} \beta sex & \beta age \end{bmatrix}$$

Adding a regression

$$G = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad D = \begin{bmatrix} sex \\ age \end{bmatrix} \quad B = \begin{bmatrix} \beta sex & \beta age \end{bmatrix}$$

$$\mathrm{B{*}D} = \begin{bmatrix} \beta sex * sex1 + \beta age * age1 \end{bmatrix}$$

$$\mathrm{G@(B{*}D)} = \begin{bmatrix} \beta sex * sex1 + \beta age * age1 \\ \beta sex * sex1 + \beta age * age1 \end{bmatrix}$$

Adding a regression

L*T + G@(B*D) =

$$\begin{bmatrix} t11 + \beta sex * sex1 + \beta age * age1 \\ (t11 + t21) + \beta sex * sex1 + \beta age * age1 \end{bmatrix}$$

Multivariate Threshold Models

Specification in Mx

Thanks Kate Morley for these slides

```
#define nsib 1    ! number of siblings = 1
#define maxth 2   ! Maximum number of thresholds
#define nvar 2    ! Number of variables
#define ndef 1    ! Number of definition variables
#define nthr 2    ! nsib x nvar


T Full maxth nthr Free    ! Thresholds
B Full nvar ndef Free     ! Regression betas
L lower maxth maxth       ! For converting incremental to cumulative thresholds
G Full maxth 1            ! For duplicating regression betas across thresholds
K Full ndef nsib          ! Contains definition variables
```

Thresholds = L*T +G@((\vec(B*K))’)

Threshold model for multivariate, multiple category data with definition variables:

Part 2 Part 1

L*T +G@((\vec(B*K))')

We will break the algebra into two parts:
1 - Definition variables;
2 - Uncorrected thresholds;
and go through it in detail.

$$\mathbf{L} * \mathbf{T} + \mathbf{G} \otimes ((vec(\mathbf{B} * \mathbf{K}))')$$

$$\mathbf{B} * \mathbf{K} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} * \begin{array}{c} \begin{array}{cc} \text{Twin 1} & \text{Twin 2} \end{array} \\ \begin{bmatrix} sex_A & sex_B \\ age_A & age_B \end{bmatrix} \end{array}$$

$$= \begin{bmatrix} b_{11} \times sex_A + b_{12} \times age_A & b_{11} \times sex_B + b_{12} \times age_B \\ b_{21} \times sex_A + b_{22} \times age_A & b_{21} \times sex_B + b_{22} \times age_B \end{bmatrix}$$

Definition variables

Threshold correction
Twin 1
Variable 1

Threshold correction
Twin 2
Variable 1

Threshold correction
Twin 1
Variable 2

Threshold correction
Twin 2
Variable 2

$$\mathbf{L} * \mathbf{T} + \mathbf{G} \otimes ((vec(\mathbf{B} * \mathbf{K}))')$$

$$vec(\mathbf{B} * \mathbf{K}) = \begin{bmatrix} b_{11} \times sex_A + b_{12} \times age_A \\ b_{21} \times sex_A + b_{22} \times age_A \\ b_{11} \times sex_B + b_{12} \times age_B \\ b_{21} \times sex_B + b_{22} \times age_B \end{bmatrix}$$

Transpose:

$$\begin{bmatrix} b_{11} \times sex_A + b_{12} \times age_A & b_{21} \times sex_A + b_{22} \times age_A & b_{11} \times sex_B + b_{12} \times age_B & b_{21} \times sex_B + b_{22} \times age_B \end{bmatrix}$$

$$\mathbf{L} * \mathbf{T} + \mathbf{G} \otimes ((vec(\mathbf{B} * \mathbf{K}))')$$

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} b_{11}sex_A + b_{12}age_A & b_{21}sex_A + b_{22}age_A & b_{11}sex_B + b_{12}age_B & b_{21}sex_B + b_{22}age_B \end{bmatrix}$$
$$= \begin{bmatrix} b_{11}sex_A + b_{12}age_A & b_{21}sex_A + b_{22}age_A & b_{11}sex_B + b_{12}age_B & b_{21}sex_B + b_{22}age_B \\ b_{11}sex_A + b_{12}age_A & b_{21}sex_A + b_{22}age_A & b_{11}sex_B + b_{12}age_B & b_{21}sex_B + b_{22}age_B \end{bmatrix}$$

$$\mathbf{L} * \mathbf{T} + \mathbf{G} \otimes ((vec(\mathbf{B} * \mathbf{K}))')$$

Thresholds 1 & 2
Twin 2
Variable 2

$$\begin{aligned}\mathbf{L} * \mathbf{T} &= \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} * \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{13} \\ t_{21} & t_{22} & t_{23} & t_{24} \end{bmatrix} \\ &= \begin{bmatrix} \boxed{\begin{matrix} t_{11} \\ t_{11} + t_{21} \end{matrix}} & \boxed{\begin{matrix} t_{12} \\ t_{12} + t_{22} \end{matrix}} & \boxed{\begin{matrix} t_{13} \\ t_{13} + t_{23} \end{matrix}} & \boxed{\begin{matrix} t_{14} \\ t_{14} + t_{24} \end{matrix}} \end{bmatrix}\end{aligned}$$

Thresholds 1 & 2
Twin 1
Variable 1

Thresholds 1 & 2
Twin 1
Variable 2

Thresholds 1 & 2
Twin 2
Variable 1

$$\mathbf{L} * \mathbf{T} + \mathbf{G} \otimes ((vec(\mathbf{B} * \mathbf{K}))')$$

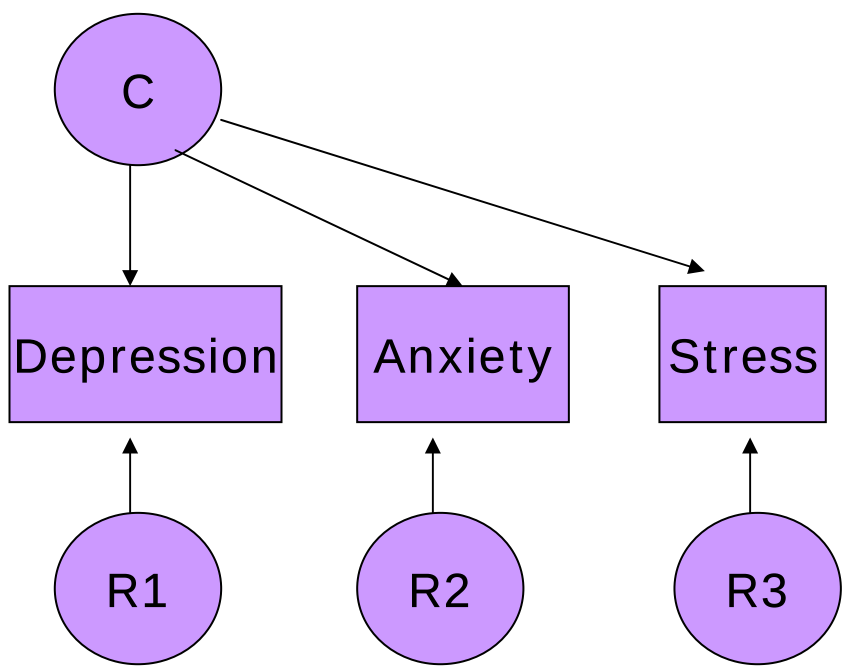
$$\begin{bmatrix} t_{11} & t_{12} & \dots \\ t_{11} + t_{21} & t_{12} + t_{22} & \dots \end{bmatrix} + \begin{bmatrix} b_{11} sex_A + b_{12} age_A & b_{11} sex_B + b_{12} age_B & \dots \\ b_{11} sex_A + b_{12} age_A & b_{11} sex_B + b_{12} age_B & \dots \end{bmatrix}$$

=

$$\begin{bmatrix} t_{11} + b_{11} sex_A + b_{12} age_A & t_{12} + b_{21} sex_A + b_{22} age_A & \dots \\ t_{11} + t_{21} + b_{11} sex_A + b_{12} age_A & t_{12} + t_{22} + b_{21} sex_A + b_{22} age_A & \dots \end{bmatrix}$$

Factor Analysis

- Suppose we have a theory that the covariation between self reports of depression, anxiety and stress levels is due to one underlying factor

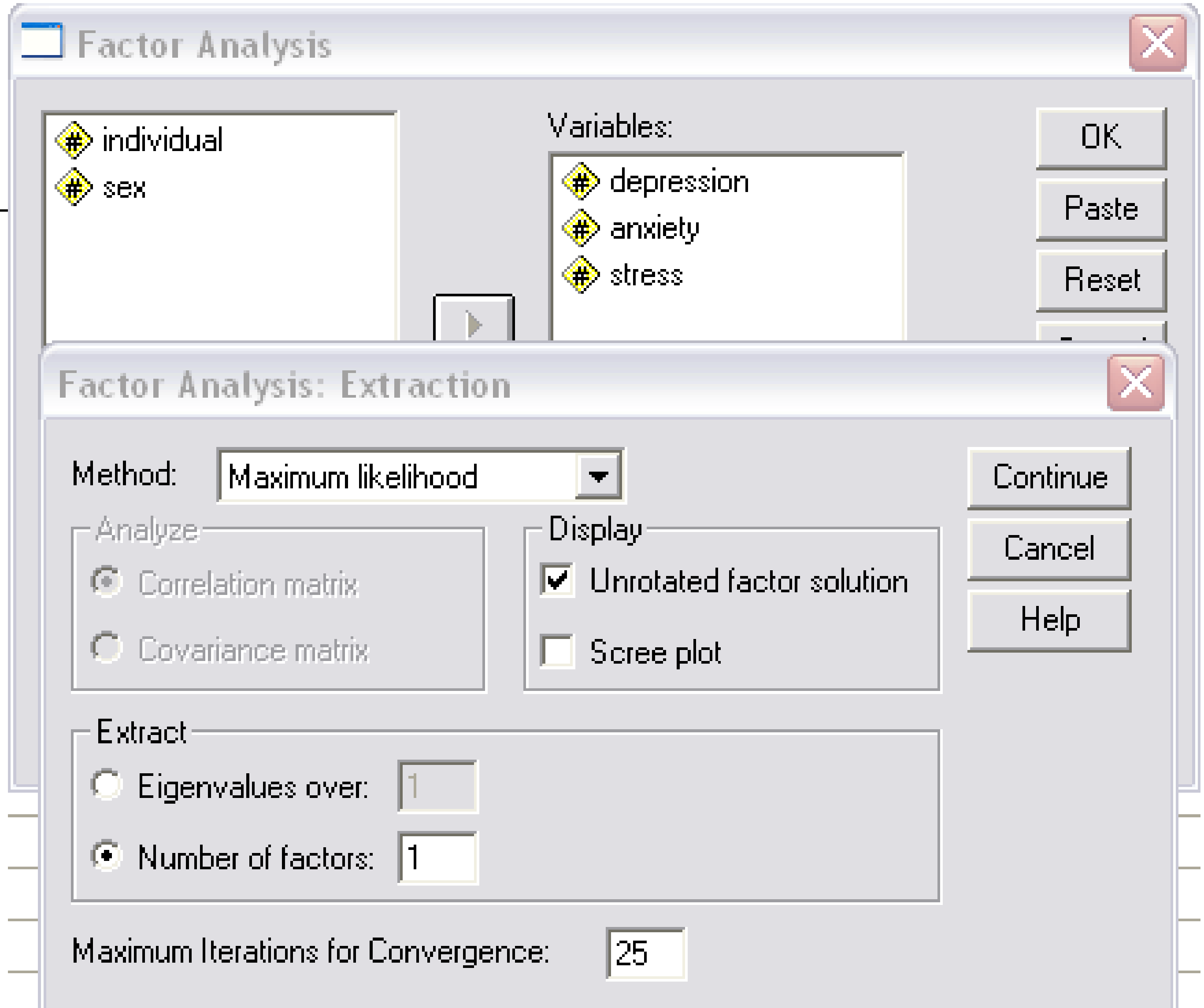


Factor Analysis….

- Our data (simulated)
  - Five variables – Three traits
  - Depression, Anxiety & Stress
  - Transformed to Z-scores

	individual	depression	anxiety	stress	sex
1	1.0	.87	-.49	.52	0
2	2.0	-1.08	.38	-.05	0
3	3.0	-.83	-.21	-1.14	0
4	4.0	-.15	-1.16	-.61	0
5	5.0	1.06	.57	.42	0
6	6.0	-.53	-1.45	-1.71	0
7	7.0	-.53	.33	.68	0
8	8.0	.31	.64	-.52	0
9	9.0	-1.38	-.47	-1.80	0

In Spss…


Factor Analysis
individual
sex
Variables:
depression
anxiety
stress
OK
Paste
Reset
Factor Analysis: Extraction
Method:
Maximum likelihood
Continue
Cancel
Help
Analyze
Correlation matrix
Covariance matrix
Display
Unrotated factor solution
Scree plot
Extract
Eigenvalues over:
1
Number of factors:
1
Maximum Iterations for Convergence:
25

Factor Analysis

Communalities

	Initial	Extraction
depression	.415	.774
anxiety	.325	.408
stress	.257	.319

Extraction Method: Maximum Likelihood.

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.951	65.045	65.045	1.501	50.037	50.037
2	.644	21.466	86.511			
3	.405	13.489	100.000			

Extraction Method: Maximum Likelihood.

Factor Matrix[a]

	Factor
	1
depression	.880
anxiety	.639
stress	.565

Extraction Method: Maximum Likelihood.

a. 1 factors extracted. 5 iterations required.

c_factor.mx

```
#define nvar 3                         ! n dependent variables per individual

G1: Singleton (non-pair) data
Data Ninput_vars=5 NGroups=2                   ! Number of variables per family

Rectangular file=SarahData.txt                 ! read raw data
Labels  Id depression anxiety stress sex


select depression anxiety stress ;

Begin matrices;
    M Full 1 nvar free                 ! mean
    L Full nvar 1 free                 ! factor loadings
    R Diag nvar nvar free              ! Residual Variance
End matrices;

Begin Algebra;
    C=L*L'+R*R' ;
End Algebra;

! start values
Start 0 M 1 1 M 1 2 M 1 3
Start .5 L 1 1 L 2 1 L 3 1
Start .5 R 1 1 R 2 2 R 3 3

Means M ;       ! means model
Covariances C ;        ! variance/covariance model
end
```

```
    L Full nvar 1 free          ! factor loadings
    R Diag nvar nvar free       ! Residual Variance
End matrices;

Begin Algebra;
    C=L*L'+R*R' ;
End Algebra;
```

C=L*L'+R*R'

$$= \begin{bmatrix} l_{dep.} \\ l_{anx.} \\ l_{stress} \end{bmatrix} * \begin{bmatrix} l_{dep.} & l_{anx.} & l_{stress} \end{bmatrix} + \begin{bmatrix} r_{dep.} & 0 & 0 \\ 0 & r_{dep.} & 0 \\ 0 & 0 & r_{dep.} \end{bmatrix} * \begin{bmatrix} r_{dep.} & 0 & 0 \\ 0 & r_{dep.} & 0 \\ 0 & 0 & r_{dep.} \end{bmatrix}$$

$$= \begin{bmatrix} l_{dep.}{}^2 & l_{dep.}.l_{anx.} & l_{dep.}.l_{stress} \\ l_{dep.}.l_{anx.} & l_{anx.}{}^2 & l_{anx.}.l_{stress} \\ l_{dep.}.l_{stress} & l_{anx.}.l_{stress} & l_{stress}{}^2 \end{bmatrix} + \begin{bmatrix} r_{dep.}{}^2 & 0 & 0 \\ 0 & r_{dep.}{}^2 & 0 \\ 0 & 0 & r_{dep.}{}^2 \end{bmatrix}$$

$$= \begin{bmatrix} l_{dep.}{}^2 + r_{dep.}{}^2 & l_{dep.}.l_{anx.} & l_{dep.}.l_{stress} \\ l_{dep.}.l_{anx.} & l_{anx.}{}^2 + r_{anx.}{}^2 & l_{anx.}.l_{stress} \\ l_{dep.}.l_{stress} & l_{anx.}.l_{stress} & l_{stress}{}^2 + r_{anx.}{}^2 \end{bmatrix}$$

c_factor.mx

- Plus a standardisation group so that our estimates can be compared to those from spss

```
Begin Matrices = Group 1 ;
End Matrices ;

Begin Algebra;
 V = \sqrt(\d2v(C)') ;          !compute the standard deviation
 S = L%V | (\d2v(R)')%V ;       !compute the standardised factor loadings
End Algebra;
End
```

What do we get?

Factor Matrix[a]

	Factor
	1
depression	.880
anxiety	.639
stress	.565

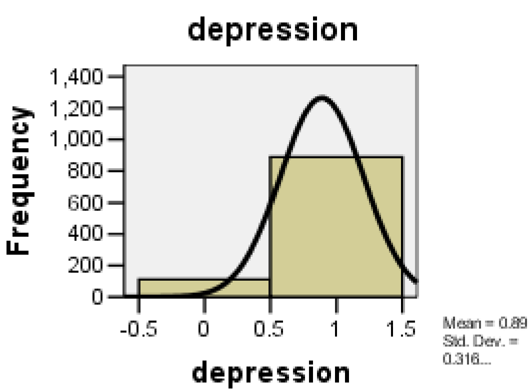
Extraction Method: Maximum Likelihood.

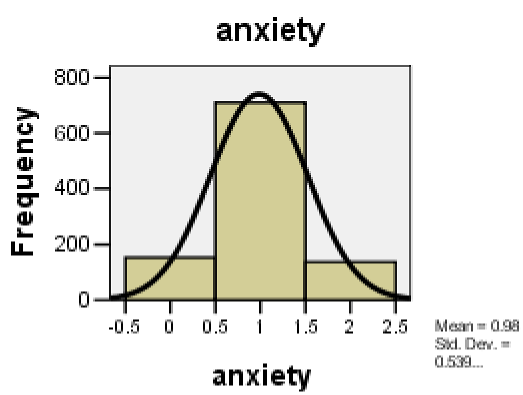
a. 1 factors extracted. 5 iterations required.

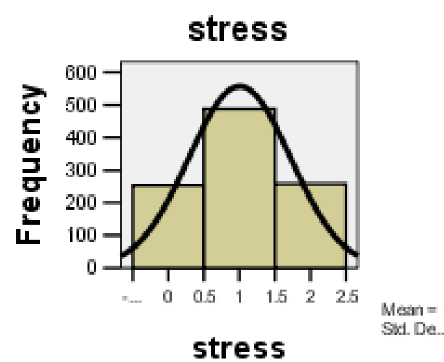
```
MATRIX S
This is a computed FULL matrix
 [=L%V|(\D2V(R)')%V]
              1          2
1        0.8795     0.4758
2        0.6390     0.7692
3        0.5649     0.8251
```

What if our data was ordinal?

- Depression
  - Yes/No 0/1
- Anxiety and Stress
  - Low / Average / High 0/1/2







Spss says no

⊟ Hide details

Data. The variables should be quantitative at the **interval** or **ratio** level. Categorical data (such as religion or country of origin) are not suitable for factor analysis. Data for which Pearson correlation coefficients can sensibly be calculated should be suitable for factor analysis.

Assumptions. The data should have a bivariate normal distribution for each pair of variables, and observations should be independent. The factor analysis model specifies that variables are determined by common factors (the factors estimated by the model) and unique factors (which do not overlap between observed variables); the computed estimates are based on the assumption that all unique factors are uncorrelated with each other and with the common factors.

Table 2.2: Classification of correlations according to their observed distribution.

Measurement	Two Categories	Three or more Categories	Continuous
Two	Tetrachoric	Polychoric	Biserial
Three or more	Polychoric	Polychoric	Polyserial
Continuous	Biserial	Polyserial	Product Moment

Mx can do this

34	0	0	0	0
69	0	0	0	0
76	0	0	0	0
106	0	0	0	0
108	0	0	0	0
199	0	0	0	0
201	0	0	0	0

- Data file: ord.dat
  - Five variables
  - ID, Depression, Anxiety, Stress, Sex
  - Data is sorted to make it run faster!!!
- Script file: o_factor.mx

O_factor.mx

```
#define nvar 3                      ! n dependent variables per individual
#define nthresh 2                   ! maximum number of thresholds

G1: Singleton (non-pair) data
Data Ninput_vars=5 NGroups=2                ! Number of variables per family

Ordinal file=ord.dat                ! read raw data
Labels  Id depression anxiety stress sex


select depression anxiety stress ;

Begin matrices;
    T Full nthresh nvar free                ! thresholds
    L Lower nthresh nthresh         ! for adding up thresholds
    F Full nvar 1 free              ! factor loadings
    R Diag nvar nvar free           ! Residual Variance
End matrices;

Begin Algebra;
    C=F*F'+R*R' ;
End Algebra;
```

O_factor.mx

```
Value 1 L 1 1 to L nthresh nthresh

Sp T
100 101 102
0   103 104

! start values
Start -1 T 1 1 T 1 2 T 1 3
Start  1 T 2 2 T 2 3
Start .5 F 1 1 F 2 1 F 3 1
Start .5 R 1 1 R 2 2 R 3 3

!Setting the 1st threshold to be negative as less than 50% in the 0 category
Bound -4 0 T 1 1 T 1 2 T 1 3
!Setting the 2nd threshold to be possitive for anxiety and stress
Bound 0.01 4 T 2 2 T 2 3
```

Set to 0 because depression has 2 categories

O_factor.mx

```
Thresholdss L*T ;                 ! thresholdss model
Covariances C ;          ! variance/covariance model
end

Standardize
Constraint

Begin Matrices = Group 1 ;
U unit 1 3                        !to constrain the total variance of the 3 variables
End Matrices ;

Begin Algebra;
 V = \d2v(C) ;                    !extract the variance
 S = F | \d2v(R)' ;               ! summary matrix containing standardised factor loadings
End Algebra;

Constraint V=U ;                  !constrain the variance
End
```

Answer

Ordinal data

```
 MATRIX S
This is a computed FULL
 [=F|\D2V(R)']
          1        2
1    0.8952   0.4457
2    0.6194   0.7850
3    0.5570   0.8305
```

Continuous data

```
 MATRIX S
This is a computed FULL matrix
 [=L%V|(\D2V(R)')%V]
          1        2
1    0.8795   0.4758
2    0.6390   0.7692
3    0.5649   0.8251
```

Difference due to loss of information with ordinal data & slightly different fit function

If we have time

- Test to see if adding another factor improves the fit