Threshold Liability Models
(Ordinal Data Analysis)

Frühling Rijsdijk

MRC SGDP Centre, Institute of Psychiatry,
King’s College London
Ordinal data

• Measuring instrument discriminates between two or a few ordered categories e.g.:
  – Absence (0) or presence (1) of a disorder
  – Score on a single Q item e.g. : 0 - 1, 0 - 4

• In such cases the data take the form of counts, i.e. the number of individuals within each category of response
Analysis of ordinal variables

• The session aims to show how we estimate correlations from count data (with the ultimate goal to estimate $h^2$, $c^2$, $e^2$)
• For this we need to introduce the concept of ‘Liability’ or ‘liability threshold models’
• This is followed by a more mathematical description of the model
Liability is a *theoretical* construct. It’s the assumption we make about the distribution of a variable which we were only able to measure in terms of a few ordered categories.

Assumptions:

(1) Categories reflect an imprecise measurement of an underlying *normal distribution* of liability.

(2) The liability distribution has 1 or more *thresholds* (cut-offs) to discriminate between the categories.
The risk or liability to a disorder is normally distributed, only when a certain threshold is exceeded will someone have the disorder. Prevalence: proportion of affected individuals.

For a single questionnaire item score e.g:

0 = not at all
1 = sometimes
2 = always

Does not make sense to talk about prevalence: we simply count the endorsements of each response category.
The Standard Normal Distribution

Liability is a latent variable, the scale is arbitrary, distribution is assumed to be a Standard Normal Distribution (SND) or z-distribution:

- Mathematically described by the SN Probability Density function ($\Phi = \text{phi}$), a bell-shaped curve with:
  - mean = 0 and SD = 1
  - z-values are the number of SD away from the mean
- Convenience: area under curve = 1, translates directly to probabilities

![Diagram of the Standard Normal Distribution with 68% between -1 and 1]
Standard Normal Cumulative Probability in right-hand tail
(For negative z values, areas are found by symmetry)

Area = P(z ≥ z_T)
Standard Normal Cumulative Probability in right-hand tail
(For negative $z$ values, areas are found by symmetry)

\[ \text{Area} = P(z \geq z_T) \]

\[ \int_{Z_T}^{\infty} \Phi(L_1; \mu = 0, \sigma^2 = 1) \, dL_1 \]
Standard Normal Cumulative Probability in right-hand tail
(For negative z values, areas are found by symmetry)

<table>
<thead>
<tr>
<th>Z₀</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.50</td>
</tr>
<tr>
<td>.2</td>
<td>.42</td>
</tr>
<tr>
<td>.4</td>
<td>.35</td>
</tr>
<tr>
<td>.6</td>
<td>.27</td>
</tr>
<tr>
<td>.8</td>
<td>.21</td>
</tr>
<tr>
<td>1</td>
<td>.16</td>
</tr>
<tr>
<td>1.2</td>
<td>.12</td>
</tr>
<tr>
<td>1.4</td>
<td>.08</td>
</tr>
<tr>
<td>1.6</td>
<td>.06</td>
</tr>
<tr>
<td>1.8</td>
<td>.036</td>
</tr>
<tr>
<td>2</td>
<td>.023</td>
</tr>
<tr>
<td>2.2</td>
<td>.014</td>
</tr>
<tr>
<td>2.4</td>
<td>.008</td>
</tr>
<tr>
<td>2.6</td>
<td>.005</td>
</tr>
<tr>
<td>2.8</td>
<td>.003</td>
</tr>
<tr>
<td>2.9</td>
<td>.002</td>
</tr>
</tbody>
</table>

\[
\text{Area} = P(z \geq z_T)
\]

\[
\int_{Z_T}^{\infty} \Phi(L_1; \mu = 0, \sigma^2 = 1) \, dL_1
\]
Two ordinal traits: Data from twins

> Contingency Table with 4 observed cells:

cell a: pairs concordant for unaffected

cell d: pairs concordant for affected

cell b/c: pairs discordant for the disorder

<table>
<thead>
<tr>
<th>Twin1</th>
<th>Twin2</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>c</td>
<td>d</td>
<td></td>
</tr>
</tbody>
</table>

0 = unaffected
1 = affected
Joint Liability Model for twin pairs

- Assumed to follow a **bivariate normal distribution**, where both traits have a mean of 0 and standard deviation of 1, but the **correlation** between them is variable.

- The **shape** of a bivariate normal distribution is determined by the **correlation** between the traits.

\[ r = 0.00 \quad r = 0.90 \]
• The observed cell proportions relate to the proportions of the BND with a certain correlation between the latent variables ($y_1$ and $y_2$), each cut at a certain threshold

• i.e. the joint probability of a certain response combination is the volume under the BND surface bounded by appropriate thresholds on each liability
Expected cell proportions

**Numerical integration** of the BND over the two liabilities e.g. the probability that both twins are above $T_c$:

$$\int_{T_{c1}}^{\infty} \int_{T_{c2}}^{\infty} \Phi(y_1, y_2; \mu = 0, \Sigma) dy_1 dy_2$$

$\Phi$ is the bivariate normal probability density function, $y_1$ and $y_2$ are the liabilities of twin1 and twin2, with means of 0, and $\Sigma$ the correlation between the two liabilities $T_{c1}$ is threshold (z-value) on $y_1$, $T_{c2}$ is threshold (z-value) on $y_2$. 
Expected cell proportions

\[
\begin{align*}
&T_{c1} \quad T_{c2} \\
&\int_{-\infty}^{T_{c1}} \int_{-\infty}^{T_{c2}} \Phi(y_1, y_2; \mu = 0, \Sigma) \, dy_1 \, dy_2 \\
&T_{c1} \quad \infty \\
&\int_{-\infty}^{T_{c1}} \int_{T_{c2}}^{\infty} \Phi(y_1, y_2; \mu = 0, \Sigma) \, dy_1 \, dy_2 \\
&\infty \quad T_{c2} \\
&\int_{T_{c1}}^{\infty} \int_{T_{c1}}^{T_{c2}} \Phi(y_1, y_2; \mu = 0, \Sigma) \, dy_1 \, dy_2 \\
&T_{c1} \quad -\infty \\
&\int_{T_{c1}}^{\infty} \int_{-\infty}^{T_{c1}} \Phi(y_1, y_2; \mu = 0, \Sigma) \, dy_1 \, dy_2 
\end{align*}
\]
Estimation of Correlations and Thresholds

• Since the BN distribution is a known mathematical distribution, for each correlation ($\sum$) and any set of thresholds on the liabilities we know what the expected proportions are in each cell.

• Therefore, observed cell proportions of our data will inform on the most likely correlation and threshold on each liability.

<table>
<thead>
<tr>
<th></th>
<th>y1</th>
<th>y2</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>.87</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.05</td>
<td>.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$\text{r} = 0.60$

$T_{c1} = T_{c2} = 1.4$ (z-value)
Bivariate Ordinal Likelihood

• The likelihood for each observed ordinal response pattern is computed by the expected proportion in the corresponding cell of the BN distribution.

• The maximum-likelihood equation for the whole sample is 
\[-2 \log \text{likelihood of each vector of observation, and summing across all observations (pairs)}\]

• This -2LL is minimized to obtain the maximum likelihood estimates of the correlation and thresholds.

• Tetra-choric correlation if \(y_1\) and \(y_2\) reflect 2 categories (1 Threshold); Poly-choric when >2 categories per liability.
Twin Models

- Estimate correlation in liabilities separately for MZ and DZ pairs from their Count data
- Variance decomposition (A, C, E) can be applied to the liability of the trait
- Correlations in liability are determined by path model
- Estimate of the heritability of the liability
ACE Liability Model

\[
\begin{align*}
E & \rightarrow C \\
C & \rightarrow A \\
A & \rightarrow C
\end{align*}
\]

\[
\begin{align*}
1 & \\
1/0.5 & \\
1 & \\
1 & \\
1 & \\
1 & \\
1 & \\
1 & \\
\end{align*}
\]

\[
\begin{align*}
\text{Unaf} & \\
\text{Aff} & \\
\text{Unaf} & \\
\text{Aff}
\end{align*}
\]

\{ Variance constraint \}
\{ Threshold model \}

Twin 1

Twin 2
Summary

• OpenMx models ordinal data under a threshold model
• Assumptions about the (joint) distribution of the data (Standard Bivariate Normal)
• The relative proportions of observations in the cells of the Contingency Table are translated into proportions under the SBN
• The most likely thresholds and correlations are estimated
• Genetic/Environmental variance components are estimated based on these correlations derived from MZ and DZ data
Power issues

• Ordinal data / Liability Threshold Model: less power than analyses on continuous data

Neale, Eaves & Kendler 1994

• Solutions:
  1. Bigger samples
  2. Use more categories
Practical

R Script: ThreshLiab.R
Data File: CASTage8.csv
Sample & Measures

- Simulated data based on CAST data collected at age 8 in the TEDS sample
- Parent report of CAST: Childhood Autism Spectrum Test (Scott et al., 2002)
- Twin pairs: 501 MZ & 503 DZ males
The CAST score dichotomized at around 98% (i.e. scores of >15), is the clinical cut-off point for children at risk for Autism Spectrum Disorder.

However, for the purpose of this exercise, we use 2 cut-offs to create 3 categories:

- <9: unaffected (0)
- 9-15: sub-clinical (1)
- >15: ASD (2)
Inspection of the data

CAST score categorized (0,1,2), the proportions:

<table>
<thead>
<tr>
<th>CAST</th>
<th>Freq.</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>804</td>
<td>80.08%</td>
</tr>
<tr>
<td>1</td>
<td>158</td>
<td>15.74%</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>4.18%</td>
</tr>
<tr>
<td>Total</td>
<td>1,004</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Z-values

Z-value Th1 = .84
Z-value Th2 = 1.75
CTs of the MZ and DZ group

table(mzData$Ocast1, mzData$Ocast2 )

table(dzData$Ocast1, dzData$Ocast2 )
Castdata <- read.table('CASTage8.csv', header=T, sep='"', na.strings='.

selVars <- c('Ocast1', 'Ocast2')

# Declare variables to be ordered Factors for OpenMx
Castdata$Ocast1  <- mxFactor(Castdata$Ocast1, levels=c(0:2) )
Castdata$Ocast2  <- mxFactor(Castdata$Ocast2, levels=c(0:2) )

# Select Data for Analysis
mzData <- subset(Castdata, zyg==1, selVars)
dzData <- subset(Castdata, zyg==2, selVars)

# get CT for Ordinal variable
table(mzData$Ocast1, mzData$Ocast2)
table(dzData$Ocast1, dzData$Ocast2)
# 1) Specify Saturated Model (max number of parameters: 2 cor, 8 TH)
# Matrices for expected Means (SND) & Tetrachoric correlations

```r
meanL <- mxMatrix( type="Zero", nrow=1, ncol=ntv, name="M" )
```

```r
corMZ <- mxMatrix(type="Stand", nrow=ntv, ncol=ntv, free=T, values=.8,
                  lbound=-.99, ubound=.99, name="expCorMZ")
```

```r
CorDZ <- mxMatrix(type="Stand", nrow=ntv, ncol=ntv, free=T, values=.8,
                  lbound=-.99, ubound=.99, name="expCorDZ")
```
Matrices & Algebra for expected Thresholds

Tmz <- mxMatrix (type="Full", nrow=nth, ncol=ntv, free=TRUE,
values=c(.8, 1, .8, 1),
lbound=c(-3, .001, -3, .001 ),
ubound=(3),
labels=c("Tmz11","imz11", "Tmz12","imz12"),
name="ThMZ" )
A multiplication is used to ensure that any threshold is higher than the previous one. This is necessary for the optimization procedure involving numerical integration over the MVN.

**Expected Thresholds:**

\[
\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad \left( \begin{bmatrix} T_{MZ11} & T_{MZ12} \\ i_{MZ11} & i_{MZ12} \end{bmatrix} \right) = \left( \begin{bmatrix} T_{MZ11} \\ T_{MZ21} \end{bmatrix} + i_{MZ11} \right) \left( \begin{bmatrix} T_{MZ12} \\ T_{MZ22} \end{bmatrix} + i_{MZ12} \right)
\]

Note: this only works if the increments are **POSITIVE values**, therefore a **BOUND** statement around the increments are necessary.
\[
\begin{pmatrix}
\text{T}_{\text{MZ11}} & \text{T}_{\text{MZ12}} \\
\text{i}_{\text{MZ11}} & \text{i}_{\text{MZ12}}
\end{pmatrix} =
\begin{pmatrix}
.8 & .8 \\
1 & 1
\end{pmatrix}
\begin{pmatrix}
\text{(-3 to 3)} & \text{(-3 to 3)} \\
\text{(.001 to 3)} & \text{(.001 to 3)}
\end{pmatrix}
\]

The positive bounds on the increments stop the thresholds going ‘backwards’, i.e. they preserve the ordering of the categories.

Z-value Th1 = .84
Z-value Th2 = 1.75
# RUN SUBMODELS

# SubModel 1: Thresholds across Twins within zyg group are equal

```r
Sub1Model <- mxModel(SatModel, name="sub1")
Sub1Model <- omxSetParameters( Sub1Model, 
labels=c("Tmz11", "imz11", "Tmz12", "imz12"), newlabels=c("Tmz11", "imz11", "Tmz11", "imz11"), ...
```

# SubModel 3: Thresholds across Twins & zyg group are equal

```r
Sub3Model <- mxModel(Sub1Model, name="sub3")
Sub3Model <- omxSetParameters( Sub3Model, 
labels=c("Td1z11", "idz11", "Td1z12", "idz12"), newlabels=c("Tmz11", "imz11", "Tmz11", "imz11"), ...
```

**omxSetParameters**: function to modify the attributes of parameters in a model
Without having to re-specify the model
# ACE MODEL with one overall set of Thresholds
pathA <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values =.6, label="a11", name="a" )
pathC <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values =.6, label="c11", name="c" )
pathE <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values =.6, label="e11", name="e" )

# Algebra for Matrices to hold A, C, and E Variance Components
covA <- mxAlgebra( expression=a %*% t(a), name="A" )
covC <- mxAlgebra( expression=c %*% t(c), name="C" )
covE <- mxAlgebra( expression=e %*% t(e), name="E" )
covP <- mxAlgebra( expression=A+C+E, name="V" )

# Constrain Total variance of the liability to 1
matUnv <-mxMatrix( type="Unit", nrow=nv, ncol=1, name="Unv" )
varL <-mxConstraint( expression=diag2vec(V)==Unv, name="VarL" )
Practical

• Run first part of the script up to sub3Model
  • What are the conclusions about the thresholds, i.e. what is the best model?
  • What kind of Genetic model would you run on this data given the correlations?

• Run the ACE model and check the parameter estimates (with 95% CI)
<table>
<thead>
<tr>
<th>MODEL</th>
<th>ep</th>
<th>-2LL</th>
<th>df</th>
<th>Δχ²(df)</th>
<th>P-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 All TH free</td>
<td>10</td>
<td>2202.7</td>
<td>1998</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2 Sub1: TH tw1=tw2 in MZ</td>
<td>8</td>
<td>2203.8</td>
<td>2000</td>
<td>1.01 (2)</td>
<td>.61 ns</td>
</tr>
<tr>
<td>3 Sub2: TH tw1=tw2 in DZ</td>
<td>8</td>
<td>2206.0</td>
<td>2000</td>
<td>3.24 (2)</td>
<td>.20 ns</td>
</tr>
<tr>
<td>4 Sub3: One overall TH</td>
<td>4</td>
<td>2211.1</td>
<td>2004</td>
<td>8.40 (6)</td>
<td>.21 ns</td>
</tr>
</tbody>
</table>

1 Thresh/Inc: MZ tw1 = .94, .84  
DZ tw1 = .75, .91

2 Thresh/Inc: MZ = .95, .78  
DZ tw1 = .75, .91

3 Thresh/Inc: MZ tw1 = .94, .84  
DZ = .80, .81

4 Thresh/Inc: .86, .79

The Twin correlations for model 4 are:

\[ r_{MZM} = 0.82 \, (0.74 - 0.87) \quad r_{DZM} = 0.44 \, (0.30 - 0.56) \]
ACE Estimates for the ordinalized CAST score in Boys at age 8

<table>
<thead>
<tr>
<th></th>
<th>h²</th>
<th>c²</th>
<th>e²</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACE</td>
<td>0.76</td>
<td>0.06</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>(0.48/0.87)</td>
<td>(0/0.31)</td>
<td>(0.13/0.26)</td>
</tr>
</tbody>
</table>

Model 1: Name  ep  -2LL  df  AIC
ACE   5  2211.14  2004  -1796.86
Age Effects on Thresholds

• The effect of covariates like Age can be modelled in the Threshold model, similarly to the means model.
• An example script is added to the folder in which Age is incorporated and its effects modelled in the thresholds (Age Regression on TH.R).

\[
\begin{pmatrix}
T_{11} & T_{12} \\
T_{21} & T_{22}
\end{pmatrix} + \begin{pmatrix}
\text{BageTH}*\text{Age1} \\
\text{BageTH}*\text{Age2}
\end{pmatrix}
\]

\[
\begin{pmatrix}
T_{11} \\
T_{21} \\
T_{12} \\
T_{22}
\end{pmatrix}
\]