

Measuring latent factors with categorical observed variables

This has been called

- Item response theory
- Factor analysis for categorical data
- Latent trait modeling

Slide 1

There are basically two general approaches which can be considered if we want to take into account the categorical nature of the observed variables

- Underlying response variable approach
- Response function approach (item response theory approach)

The difference in these two approaches boils down to where the categorical nature of the observed data is taken into account.

Underlying variable approach

Assumes that underlying each of the categorically observed variables x_j is a continuous variable x_j^* which is actually measuring the ultimate underlying latent factors \mathbf{f} , but we were only able to “partially observe” it through x_j .

We assume the traditional linear factor analysis model for the “partially observed” variables \mathbf{x}_j^* , i.e.

$$\mathbf{x}^* = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon}$$

and we assume that for $x_j = s$, with $s = 0, 1, \dots, m_j$ that

$$x_j = \begin{cases} 1 & \text{if } -\infty < x_j^* \leq \tau_{j(1)} \\ 2 & \text{if } \tau_{j(1)} < x_j^* < \tau_{j(2)} \\ \vdots & \\ s & \text{if } \tau_{j(s-1)} < x_j^* < \tau_{j(s)} \\ \vdots & \\ m_j & \text{if } \tau_{j(m_j-1)} < x_j^* < \infty \end{cases}$$

with $-\infty < x_j^* < \tau_{j(1)} < \tau_{j(1)} < \dots < \tau_{j(m_j-1)} < \infty$ are parameters called threshold values.

Slide 2

Underlying variable approach

Since only ordinal information is available about x_j^* , the mean and variance of it are not identified and are therefore set to zero and one, respectively. It is assumed that $\mathbf{f} \sim N(\mathbf{0}, \mathbf{I})$ and $\epsilon \sim N(\mathbf{0}, \text{diag}(\Psi))$.

Several methods for fitting this model:

- underlying multivariate normality technique (requires p dimensional integral)
- underlying bivariate normality technique (single step uses only bivariate distributions)
- three step approach
 - Method implemented in LISREL and MPLUS
 - first step: estimate thresholds based on univariate distributions
 - second step: estimate correlation of \mathbf{x}^* using polychoric correlations. Note if two variables are dichotomous, it is called tetrachoric correlation.
 - Use the polychoric correlations as input into a usual factor analysis routine

Limited information techniques (except for the multivariate normality technique) because they are only based on univariate and bivariate moments. For categorical data, the first and second moments are not sufficient statistics for the entire distribution.

Slide 3

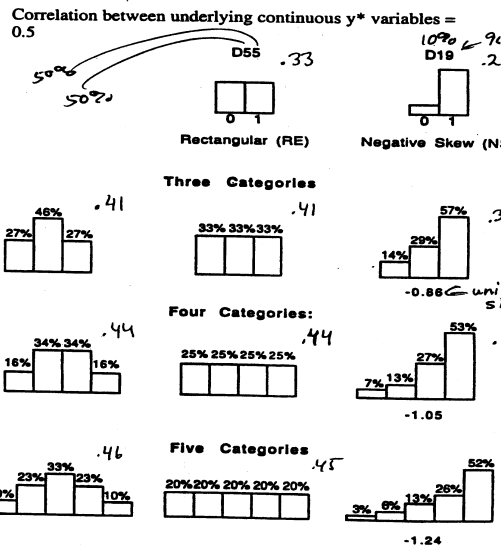
Polychoric correlation - How does it work Draw picture

Slide 4

Pearson correlation for categorical - How bad is it?

From Muthen Muthen Mplus shortcourse

Correlational Attenuation



Slide 5

Pearson correlation for categorical - How bad is it?

From Muthen Muthen Mplus shortcourse

Table 1 (Part 2)
Pearson Correlations for True Correlations = 0.50

	D19	D28	D37	D46	D55	D64	D73	D82	D91	SSY	SRE	SNS	SPS
D19	25												
D28	26	30											
D37	26	30	32										
D46	24	30	32	33									
D55	23	28	31	33	33								
D64	20	26	30	23	33	33							
D73	18	23	27	30	31	32	32						
D82	15	20	23	26	28	30	30	30					
D91	10	15	18	20	22	24	26	26	25				
SSY	26	32	35	36	37	36	35	32	26	41			
SRE	25	31	35	36	37	36	35	31	25	41	41		
SNS	29	33	35	36	35	33	30	26	20	39	39	39	
SPS	20	26	30	33	35	36	35	33	29	39	39	34	39
4SY	27	33	36	38	38	38	36	33	27	43	43	40	40
4RE	26	33	36	38	38	38	36	33	26	42	42	40	40
4NS	30	35	36	36	35	33	30	27	20	40	39	40	34
4PS	20	27	31	34	35	36	35	30	40	39	34	40	
SSY	28	34	37	38	39	38	37	34	28	44	43	41	41
SRE	27	33	37	38	39	38	37	33	27	43	43	41	41
SNS	31	35	36	36	35	33	30	26	20	40	39	40	34
SPS	20	26	30	33	35	36	35	31	40	39	34	40	
CCN	29	35	38	39	40	39	38	35	29	45	45	42	42

Pearson Correlations for True Correlations = .50

	4SY	4RE	4NS	4PS	SSY	SRE	SNS	SPS	CCN
4SY	44								
4RE	44	44							
4NS	41	41	41						
4PS	41	41	35	41					
SSY	45	45	42	42	46				
SRE	45	45	41	41	46	45			
SNS	41	40	42	34	42	41	42		
SPS	41	41	34	42	42	41	34	42	
CCN	47	46	43	43	48	47	44	44	50

Slide 6

Alcoholism example

Questionnaire fill out by patients presenting for alcoholism treatment. Part of a large assessment/evaluation study to examine the impact of alcoholism treatment on health related outcomes. The present data are made up of 8227 individuals (34.2% are female) all are over 18 yrs old (average age 37).

hangover	Had hangover/headach (past 6 months)	56.10
nausea	Had nausea/vomiting (past 6 months)	32.79
sweats	Had sweats (past 6 months)	30.59
seizures	Had seizures (past 6 months)	2.80
cravings	Had cravings (past 6 months)	48.24
passout	Passed out (past 6 months)	24.12
blackout	Blacked out (past 6 months)	41.91
shakes	Had shakes/tremors (past 6 months)	30.67
sleeping	Had trouble sleeping (past 6 months)	46.71
dts	Had the DTs (past 6 months)	2.77
halluc	Had hallucinations (past 6 months)	6.82
othsym	Other physical symptoms (past 6 months)	5.35

Slide 7

There are $2^{12} = 4096$ possible response profiles. From these 8227 individuals there were only a total of 794 different response profiles given. 18.1% of individuals said no to all 12 items. Some examples of other response profiles are: 3.8% who said yes only to the hangover question, 2.1% who said yes to everything except seizures, dts and halluc, 1.9% said yes to just trouble sleeping and cravings.

Output from Mplus

```

SAMPLE STATISTICS
ESTIMATED SAMPLE STATISTICS

      SAMPLE THRESHOLDS
      HANGOVER    NAUSEA$1    SWEATS$1    SEIZURES    CRAVINGS
      -----
      1    -0.153    0.446    0.507    1.912    0.044

      SAMPLE THRESHOLDS
      PASSOUT$    BLACKOUT    SHAKES$1    SLEEPING    DTSS$1
      -----
      1    0.703    0.204    0.505    0.083    1.916

      SAMPLE THRESHOLDS
      HALLUC$1    OTHSYM$1
      -----
      1    1.489    1.612
    
```

Slide 8

These are just cumulates of the normal distribution describing the overall marginal proportion of a “NO” response

e.g. $F(-.153) = .439$, Hence the proportion with a hangover is $1 - .439 = .561$ or $F(1.612) = .9465$, Hence the proportion with other symptoms is $1 - .9465 = .0535$

Example association between Hangover and Nausea

Frequency		NAUSEA		
Percent				
Row Pct				
Col Pct	0	1	Total	
	0	3179	433	3612
		38.64	5.26	43.90
		88.01	11.99	
		57.50	16.05	
HANGOVER	1	2350	2265	4615
		28.56	27.53	56.10
		50.92	49.08	
		42.50	83.95	
Total		5529	2698	8227
		67.21	32.79	100.00

Slide 9

$$\text{odds ratio} = \frac{ad}{bc} = \frac{3179 * 2265}{2350 * 433} = 7.0762$$

The polychoric correlation is estimated by maximum likelihood finding the bivariate correlation needed so that if a bivariate normal were dichotomized using the univariate thresholds that the 2 by 2 table above would result. It is 0.623 (Next page).

Approximation to the polychoric correlation

- Yule's Q = $\frac{or-1}{or+1} = 0.7523$
- Digby's approximation = $\frac{(or)^{3/4}-1}{(or)^{3/4}+1} = 0.6254$

Output from Mplus

SAMPLE TETRACHORIC CORRELATIONS					
	HANGOVER	NAUSEA	SWEATS	SEIZURES	CRAVINGS
HANGOVER					
NAUSEA	0.623				
SWEATS	0.449	0.503			
SEIZURES	0.098	0.305	0.258		
CRAVINGS	0.429	0.434	0.540	0.217	
PASSOUT	0.481	0.449	0.349	0.265	0.385
BLACKOUT	0.539	0.483	0.364	0.240	0.385
SHAKES	0.437	0.495	0.652	0.411	0.499
SLEEPING	0.391	0.409	0.558	0.165	0.512
DTS	0.232	0.356	0.464	0.433	0.339
HALLUC	0.242	0.354	0.453	0.468	0.427
OTHSYM	0.055	0.152	0.111	0.044	0.139

Slide 10

SAMPLE TETRACHORIC CORRELATIONS					
	PASSOUT	BLACKOUT	SHAKES	SLEEPING	DTS
BLACKOUT	0.680				
SHAKES	0.449	0.482			
SLEEPING	0.304	0.335	0.505		
DTS	0.310	0.360	0.638	0.343	
HALLUC	0.268	0.251	0.439	0.447	0.563
OTHSYM	0.114	0.089	0.139	0.214	0.174

SAMPLE TETRACHORIC CORRELATIONS		
	HALLUC	OTHSYM
OTHSYM	0.162	

Output from Mplus - Pearson correlations

	Correlations				
	HANGOVER	NAUSEA	SWEATS	SEIZURES	CRAWINGS
HANGOVER	1.000				
NAUSEA	0.392	1.000			
SWEATS	0.273	0.322	1.000		
SEIZURES	0.030	0.100	0.084	1.000	
CRAWINGS	0.280	0.275	0.343	0.066	1.000
PASSOUT	0.275	0.275	0.208	0.087	0.228
BLACKOUT	0.353	0.311	0.228	0.076	0.250
SHAKES	0.265	0.317	0.438	0.138	0.315
SLEEPING	0.253	0.259	0.358	0.051	0.342
DTS	0.067	0.117	0.156	0.133	0.101
HALLUC	0.096	0.156	0.205	0.171	0.173
OTHSYM	0.021	0.059	0.043	0.009	0.054

Slide 11

	Correlations				
	PASSOUT	BLACKOUT	SHAKES	SLEEPING	DTS
PASSOUT	1.000				
BLACKOUT	0.434	1.000			
SHAKES	0.275	0.308	1.000		
SLEEPING	0.180	0.216	0.321	1.000	
DTS	0.104	0.112	0.212	0.103	1.000
HALLUC	0.116	0.106	0.199	0.183	0.228
OTHSYM	0.043	0.035	0.054	0.083	0.042

	Correlations	
	HALLUC	OTHSYM
HALLUC	1.000	
OTHSYM	0.049	1.000

Output from Mplus - treating data as categorical

RESULTS FOR EXPLORATORY FACTOR ANALYSIS

EIGENVALUES FOR SAMPLE CORRELATION MATRIX

	1	2	3	4	5
1	5.186	1.366	1.063	0.932	0.644

EIGENVALUES FOR SAMPLE CORRELATION MATRIX

	6	7	8	9	10
1	0.610	0.499	0.469	0.369	0.327

EIGENVALUES FOR SAMPLE CORRELATION MATRIX

	11	12
1	0.299	0.236

Slide 12

Output from Mplus - treating data as categorical

```

ROOT MEAN SQUARE RESIDUAL IS          0.0894
ESTIMATED FACTOR LOADINGS
1
-----
HANGOVER      0.617
NAUSEA        0.693
SWEATS        0.732
SEIZURES      0.431
CRAVINGS      0.657
PASSOUT       0.610
BLACKOUT      0.639
SHAKES        0.800
SLEEPING      0.632
DTS           0.628
HALLUC        0.598
OTHYSM        0.199

ROOT MEAN SQUARE RESIDUAL IS          0.0562
PROMAX ROTATED LOADINGS
1          2
-----
HANGOVER      0.856      -0.130
NAUSEA        0.604        0.176
SWEATS        0.311        0.502
SEIZURES      -0.086        0.573
CRAVINGS      0.375        0.353
PASSOUT       0.671        0.025
BLACKOUT      0.739      -0.003
SHAKES        0.282        0.608
SLEEPING      0.289        0.412
DTS           -0.064        0.785
HALLUC        -0.109        0.798
OTHYSM        0.025        0.196

Promax factor correlations
1          2
1          1.000
2          0.589  1.000

```

Slide 13

Output from Mplus - treating data as categorical

```

ROOT MEAN SQUARE RESIDUAL IS          0.0343

PROMAX ROTATED LOADINGS
1          2          3
-----
HANGOVER      0.553      -0.232      0.405
NAUSEA        0.401        0.047      0.373
SWEATS        -0.003        0.137      0.715
SEIZURES      0.091        0.655     -0.108
CRAVINGS      0.104        0.053      0.595
PASSOUT       0.754        0.114     -0.059
BLACKOUT      0.840        0.087     -0.069
SHAKES        0.180        0.396      0.405
SLEEPING     -0.056        0.024      0.748
DTS           0.022        0.679      0.159
HALLUC        -0.121        0.556      0.347
OTHYSM        -0.037        0.082      0.187

PROMAX FACTOR CORRELATIONS
1          2          3
-----
1          1.000
2          0.371      1.000
3          0.576      0.480      1.000

```

Slide 14

Response variable approach

For a given response vector \mathbf{x} , we can write its distribution as...

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{f}) d\mathbf{f} = \int p(\mathbf{x}|\mathbf{f}) p(\mathbf{f}) d\mathbf{f}$$

We will assume local conditional independence for the elements x_j in \mathbf{x} given \mathbf{f} , so that

$$p(\mathbf{x}|\mathbf{f}) = \prod_{j=1}^p p(x_j|\mathbf{f})$$

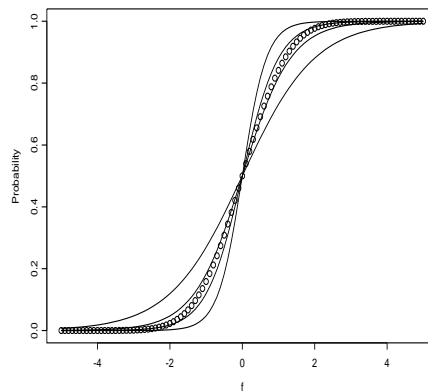
We first consider the binary case where each element in \mathbf{x} is equal to 0 or 1. We assume the underlying factors \mathbf{f} are $N(0, I)$. A natural function/distribution to choose for $p(x_j|\mathbf{f})$ is the logistic function,

$$p(x_j|\mathbf{f}) = \pi_j(\mathbf{f}) = \frac{\exp(\alpha_{0j} + \boldsymbol{\alpha}'_{1j}\mathbf{f})}{1 + \exp(\alpha_{0j} + \boldsymbol{\alpha}'_{1j}\mathbf{f})}$$

where α_{0j} is the intercept or “difficulty” of item j , and the vector $\boldsymbol{\alpha}_{1j}$ represent the slopes relating each of the k factors to item j , these are also called the “discrimination” parameters. When there is only one factor, the $\boldsymbol{\alpha}_{1j}$ becomes a single scalar α_{1j} .

Slide 15

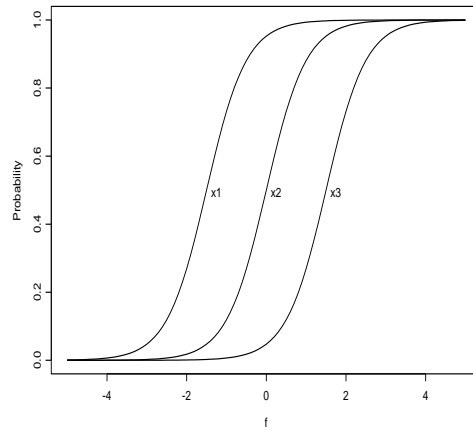
Response variable approach



Other link functions besides the logistic can be used, e.g. the normit link or normal ogive. The dots represent the normit function (i.e. the cumulative distribution of the $N(0,1)$ distribution). The lines represent the logit function with $\alpha_0 = 0$ and $\alpha_1 = 1, 1.5, 2, 3$, respectively.

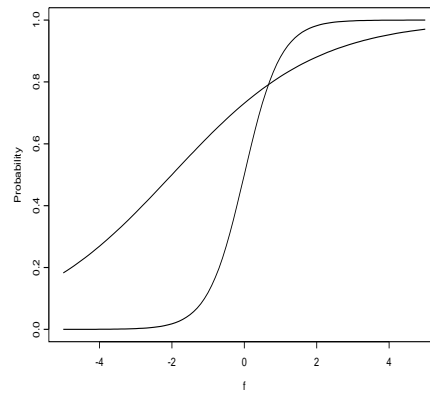
Slide 16

Slide 17



Rasch model. The $\alpha_{11} = \alpha_{12} = \alpha_{13}$ (i.e. the slopes “discriminations” are the same), but the intercepts “difficulties” vary $\alpha_{01} = 3, \alpha_{02} = 0, \alpha_{03} = -3$

Slide 18



Here the two observed variables have different slopes and intercepts.

$\alpha_{01} = 0$ and $\alpha_{11} = 2$

$\alpha_{01} = 1$ and $\alpha_{11} = .5$

Response variable approach - more than two category outcomes

Similar idea, use the cumulative odds model. See the polity data example from first week of class.

Slide 19

Pros and cons of the two approaches

Software:

- The underlying variable approach is implemented in Mplus and LISREL (not AMOS). Very simply (in Mplus), the user specifies which observed variables are categorical and the software implements the method.
- The response variable approach is not available in Mplus, LISREL or AMOS. It can be programmed directly (writing down the likelihood) in Proc NL MIXED or within a Bayesian framework using Winbugs. An add-on to STATA called gllam ("generalized linear latent variable modeling") will do this approach. Specialized softwares exist within developed by researchers in the education field for using this approach.

The following are points taken from "Factor analysis of ordinal variables: A comparison of three Approaches" by Karl Jöreskog and Irini Moustaki *Multivariate Behavioral Research*, 36(3), 347-387, 2001.

- In theory, the response function approach is better than the underlying variable approach, but from a practical point of view their use is limited.
- The Response function approach are computationally heavy. The likelihood is not available in closed form and is approximated by numerical integration requiring several quadrature points. They work reasonably well for 1 and 2 factors. Eg, NL MIXED does not recommend fitting models with more than two "random effects", that is underlying factors. If an accurate solution is not required, it is possible to handle three and four factors by reducing the number of quadrature points along each dimension.
- By contrast, the underlying variable approach (using bivariate moments) is feasible for many variables and many factors. One particular advantage is that computer time does not increase with sample size.

Goals:

- If the goal is a factor analysis with many factors in particular to identify clusters of variable "measuring" the same things, then there is no choice but to use the UV approach.
- If on the other hand the goal is to find/choose a set of items measuring only one factor and then to use these items to specifically score and rank individuals, the response theory approach is appropriate.

Slide 20