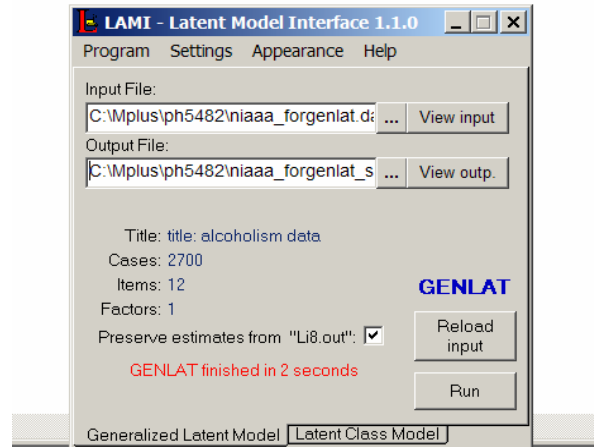


Using the GENLAT software for IRT



Slide 1

Download the LAMI software from <http://multilevel.ioe.ac.uk/team/aimdss.html>

Using the GENLAT software for IRT

Here is the input file into the GENLAT program for the NIAAA symptoms data. Note this analysis is restricted (due to software limitation) to 2700 patients rather than the over 8000 patients in the original NIAAA dataset.

```
title: alcoholism data
2700 12 0 0 0 48
```

```
1 0 2000 1 0.0000001 0
1 0 0 0 0 1 0 0 1 0 0 0
0 0 1 0 1 0 1 1 1 0 0 0
0 0 0 1 0 1 0 1 0 0 1 0
0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 1 1 1 1 0 0 0
1 1 1 0 1 1 1 1 1 0 1 0
...
```

See the handout of the README file for GENLAT to decipher what the different numbers mean before the data

Slide 2

Slide 3

README file for GENLAT

Program GENLAT

Description of the Input and Output files

The program fits a latent trait model for binary, ordinal nominal and metrical observed variables with up to two latent variables. The program computes parameter estimates, standard errors, chi-squared residuals, and scoring methods based on the posterior mean and the component score.

Limitations: Number of factors=2, number of items= 20, sample size=3000.

File names:

Input File: (Lat.inp) (ascii)

Output file: (all files ascii)
LI7.out is the main output file which contains the parameter estimates, standard errors, chi-squared residuals and scoring results.

LI8.out contains only the parameter estimates. The parameter estimates given in li8.out can be used as initial estimates in the program if more iterations are needed to achieve convergence. To do that rename the li8.out file as lat3.inp.

Slide 4

README file for GENLAT

Description of the Input file (Lat.inp):

Line 1:

Title - up to 72 characters

Line 2: information must be separated by at least one space

N NPB NPM NPN NPO NQ

N: Number of response patterns to be read (maximum=3000)

NPB: Number of binary observed variables

NPM: Number of metrical observed variables

NPN: Number of nominal observed variables

NPO: Number of ordinal observed variables

NQ: Number of quadrature points (8,16,36,48) *

Line 3: denote the number of categories for the nominal observed variables. Different items can have different number of categories (maximum number of categories=6).

Line 4: denote the number of categories for the ordinal observed variables. Different items can have different number of categories (maximum number of categories=6).

Line 5: (information must be separated by at least one space)

NFAC INIT ITER OUT PREC SCOR

NFAC: Number of factors (1 or 2)

INIT: 0 if the initial parameter values are set in the program or 1 if the initial parameter estimates are to be read from file lat3.inp.

ITER: Number of iterations (maximum is 2000)

OUT: 1 if a file LI8.OUT is to be created, 0 otherwise

PREC: Precision for maximization (e.g. 0.0000001, convergence tolerance of the EM algorithm)

SCOR: 1 if scoring results to be printed, 0 otherwise.

Line 6 and onwards: raw data.

Output from GENLAT software

```
*** PROGRAM LATENT ***
MAXIMUM LIKELIHOOD ESTIMATION OF A 1 FACTOR MODEL

Filename: title: alcoholism data

MODEL = 1
NUMBER OF OBSERVED VARIABLES = 12
NUMBER OF BINARY VARIABLES = 12
NUMBER OF NORMAL VARIABLES = 0
NUMBER OF NOMINAL VARIABLES = 0
NUMBER OF ORDINAL VARIABLES = 0
NUMBER OF CASES SAMPLED = 2700
PROPORTION OF RESPONSE PATTERNS WITH AT LEAST ONE MISSING OBSERVATION = 0.000
NUMBER OF QUADRATURE POINTS USED = 48
MAXIMUM NUMBER OF ITERATIONS PERMITTED = 2000
CONVERGENCE TOLERANCE FOR THE RELATIVE LIKELIHOOD VALUE = 0.00000010

***Initial Estimates of Item Parameters***

ITEM      ALPHAB(0,I)  ALPHAB(1,I)

1          0.000    1.000
2          0.000    1.000
3          0.000    1.000
4          0.000    1.000
5          0.000    1.000
6          0.000    1.000
7          0.000    1.000
8          0.000    1.000
9          0.000    1.000
10         0.000    1.000
11         0.000    1.000
12         0.000    1.000
```

Slide 5

Output from GENLAT software

```
***Maximum Likelihood Estimates Of Item Parameters and Standard Errors***
Binary Items
ITEM I  ALPHAB(0,I)  S.E.    ALPHAB(1,I)  S.E    P(X=1/Z=0)
1       0.2888     0.0561   1.4980     0.0831   0.5717
2      -1.0110     0.0690   1.6765     0.1004   0.2668
3      -1.2924     0.0812   1.9502     0.1187   0.2155
4      -4.1078     0.2007   1.2011     0.1677   0.0162
5       0.0523     0.0546   1.4714     0.0837   0.5131
6      -1.6149     0.0762   1.3928     0.0939   0.1659
7      -0.5375     0.0579   1.4238     0.0857   0.3688
8      -1.4220     0.0924   2.2897     0.1395   0.1943
9      -0.1141     0.0521   1.2904     0.0749   0.4715
10     -5.9081     0.5119   2.3389     0.3307   0.0027
11     -3.3555     0.1542   1.5774     0.1386   0.0337
12     -2.9779     0.0933   0.3066     0.1087   0.0484

***Standardized Loadings***

Binary Items  STALPHAB
1             0.8317
2             0.8588
3             0.8898
4             0.7685
5             0.8271
6             0.8123
7             0.8183
8             0.9164
9             0.7904
10            0.9195
11            0.8446
12            0.2931
```

Slide 6

Response variable approach - more than two category outcomes

Recall each observed variable x_j can have m_j responses (they do not have to be the same, for example x_1 can be binary ($m_1 = 2$) while x_2 has 3 category response ($m_2 = 3$))

Define $\gamma_{j(s)}(\mathbf{f}) = Pr(x_j \leq s)$ That is γ represents the cumulative probability, the probability of x_j being less than s . Thus $1 - \gamma_{j(s)}(\mathbf{f})$ is just the probability of x_j being greater than s .

Cumulative logistic modeling (or ordered logistic regression or in the most commonly considered special case - proportional odds modeling) is the following

$$\log \frac{1 - \gamma_{j(s)}(\mathbf{f})}{\gamma_{j(s)}(\mathbf{f})} = \alpha_{j(s)} + \alpha_j \mathbf{f}$$

Note that there is a different intercept $\alpha_{j(s)}$ for each category s , but that the slope α_j is the same for all categories. In fact this constant slope is the assumption of the “proportional odds model”.

Slide 7

Response variable approach - more than two category outcomes

$$\log \frac{1 - \gamma_{j(s)}(\mathbf{f})}{\gamma_{j(s)}(\mathbf{f})} = \alpha_{j(s)} + \alpha_j \mathbf{f}$$

We now have defined the model by supposing that the binary logit model holds for all possible divisions of the m_j categories into two groups.

The model is again fit by Maximum Likelihood.

Usually the cumulative odds are backtransformed into category response probabilities where they are more easily interpreted.

See the polity data example from first week of class.

Slide 8

Goodness-of-fit for models with Categorical observed data

The most commonly used fit statistics for multinomially distributed data (i.e. categorical data) are the likelihood ratio test (for a multinomial likelihood) G^2 or the “Deviance” and the Pearson chi-squared goodness of fit test statistic X^2 .

Both of these statistics consider how close the “predicted” number of responses in each of the possible response profiles matches the observed number of responses in each of the response profiles.

Slide 9

$$G^2 = 2 \sum_{r=1}^{\prod_{j=1}^p m_j} O(r) \log \frac{O(r)}{E(r)}$$

$$X^2 = \sum_{r=1}^{\prod_{j=1}^p m_j} \frac{(O(r) - E(r))^2}{E(r)}$$

Both are compared to a χ^2 distribution with $\prod_{j=1}^p m_j - \sum_{j=1}^p m_j - p - pq - 1$

Rule of thumb that expected cell counts are greater than 5 is often not true for these types of models. Common to consider fit of particular marginalized distribution - one way, two way, three way tables.

Factor score estimation

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

Once the model is fit, we have $p(\mathbf{x}|\mathbf{f})$ we can then use Bayes Theorem to get

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

and thus $E(\mathbf{f}|\mathbf{x})$ can be computed for each observed \mathbf{x} and a predicted value for \mathbf{f} is obtained

Slide 10

In the case of binary data with the 2 parameter logistic, a component score can be obtained more simply by taking

$$\hat{f} = \sum_{j=1}^p \hat{\alpha}_{1j} x_j$$

That is create a weighted sum score using the discrimination parameters as the weights

Democracy as a Latent Variable*

Shawn Treier
Stanford University
satreier@stanford.edu

Simon Jackman
Stanford University
jackman@stanford.edu

July 17, 2003

Slide 11

Abstract

Measurement is critical to the social scientific enterprise. Many key concepts in social-scientific theories are not observed directly, and researchers rely on assumptions (tacitly or explicitly, via formal measurement models) to operationalize these concepts in empirical work. In this paper we apply formal, statistical measurement models to the Polity IV data, a set of country-level indicators of democracy. In so doing, we make explicit the hitherto implicit assumptions underlying scales built using the Polity indicators. We apply two models: one in which democracy is operationalized a latent continuous variable, and another in which democracy is operationalized a latent class. We show how to better exploit the information in the Polity data set so as to produce a more reliable scale measure (or classification) of democracy. Our modeling approaches also let us assess the "noise" (measurement error) in our resulting measure of democracy. We show that this measurement error is considerable, and has substantive consequences when using a measure of democracy as an independent variable in cross-national statistical analysis. Our analysis suggests that skepticism as to the precision of the Polity democracy scale is well-founded, and that many researchers have been overly sanguine about the properties of the Polity democracy scale in applied statistical work.

1 Latent Variables Abound in Political Science

Social and political theories often refer to constructs that can not be observed directly. Examples include public opinion, socio-economic status, social capital, ideology, or democracy. Instead of observing these quantities, researchers may have *indicators* of these concepts,

*Prepared for delivery at the 2003 Annual Meeting of the Society for Political Methodology, University of Minnesota, Minneapolis, July 17-19, 2003. Earlier versions of this work were presented at the 2003 Annual Meeting of Midwestern Political Science Association and at Stanford University. We thank Jon Bendor, Alberto Diaz, Jim Fearon, Steve Krasner, David Laitin, Andrew Martin, Doug Rivers, and Mike Tomz for useful comments and references. Errors and omissions remain our own responsibility.

Polity IV Data

"Many different collections of indicators of democracy have been employed at one time or another in studies of international relations and comparative politics. We base our empirical analysis on the Polity collection from the Polity IV Project (Marshall and Jaggers, 2002)...The observed data are indicators related to executive recruitment, directiveness and responsiveness, constraints on the executive, and political participation. The Polity scores use five expert-coded categorical indicators, all capable of being ordered: they are

Slide 12

1. Competitiveness of executive recruitment
2. Openness of executive recruitment
3. Executive Constraints/Decision Rules
4. Regulation of Participation
5. Competitiveness of Participation "

Slide 13

Marginal Distributions					
	XRCOMP	XROPEN	PARREG	PARCOMP	XCONST
1	.11	.11	.41	.28	.33
2	.54	.18	.28	.15	.05
3	.06	.08	.10	.30	.25
4	.29	.01	.18	.06	.02
5		.61		.17	.06
6					.02
7					.27
NA			.04	.04	
Mean	2.5	3.8	2.0	2.7	3.6
Std Dev	1.0	1.6	1.1	1.4	2.4

Pearson Product Moment Correlation Matrix					
	XRCOMP	XROPEN	PARREG	PARCOMP	XCONST
XROPEN	.67				
PARREG	.71	.39			
PARCOMP	.68	.36	.95		
XCONST	.75	.48	.72	.72	
Eigenvalues of Correlation Matrix					
	3.61	.81	.32	.19	.05

Table 2: Summary Statistics, Correlation Matrix, and Eigenvalues, Five Indicators from Polity IV

Slide 14

	Discrimination		Thresholds		
	Parameter				
Competitiveness of Executive Recruitment (XRCOMP)	2.36 [2.29, 2.42]	T ₁₁	-3.46	[-3.54, -3.38]	
		T ₁₂	0.90	[0.85, 0.94]	
		T ₁₃	1.46	[1.40, 1.51]	
Openness of Executive Recruitment (XROPEN)	1.40 [1.35, 1.45]	T ₂₁	-2.65	[-2.72, -2.59]	
		T ₂₂	-1.19	[-1.24, -1.15]	
		T ₂₃	-0.69	[-0.74, -0.65]	
		T ₂₄	-0.63	[-0.68, -0.59]	
Regulation of Participation (PARREG)	8.98 [8.76, 9.20]	T ₃₁	-2.26	[-2.36, -2.17]	
		T ₃₂	4.10	[3.97, 4.23]	
		T ₃₃	7.50	[7.40, 7.59]	
Competitiveness of Participation (PARCOMP)	8.28 [8.15, 8.42]	T ₄₁	-4.59	[-4.64, -4.52]	
		T ₄₂	-1.64	[-1.73, -1.55]	
		T ₄₃	5.31	[5.20, 5.42]	
		T ₄₄	7.39	[7.32, 7.46]	
Executive Constraints (XCONST)	2.60 [2.54, 2.67]	T ₅₁	-1.48	[-1.53, -1.43]	
		T ₅₂	-1.10	[-1.15, -1.06]	
		T ₅₃	0.82	[0.77, 0.87]	
		T ₅₄	0.99	[0.93, 1.03]	
		T ₅₅	1.60	[1.55, 1.65]	
		T ₅₆	1.84	[1.79, 1.90]	

Table 3: Discrimination Parameters and Thresholds. Posterior Means, with 95% Highest Posterior Density Intervals in brackets.

Slide 15

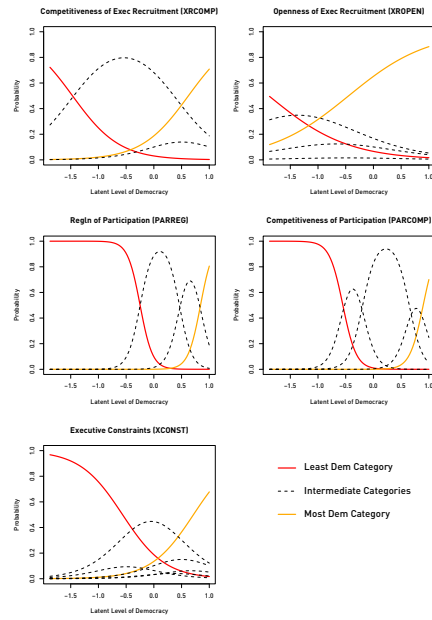


Figure 1: Item Characteristic Curves, Five Polity IV indicators

18

Slide 16

22

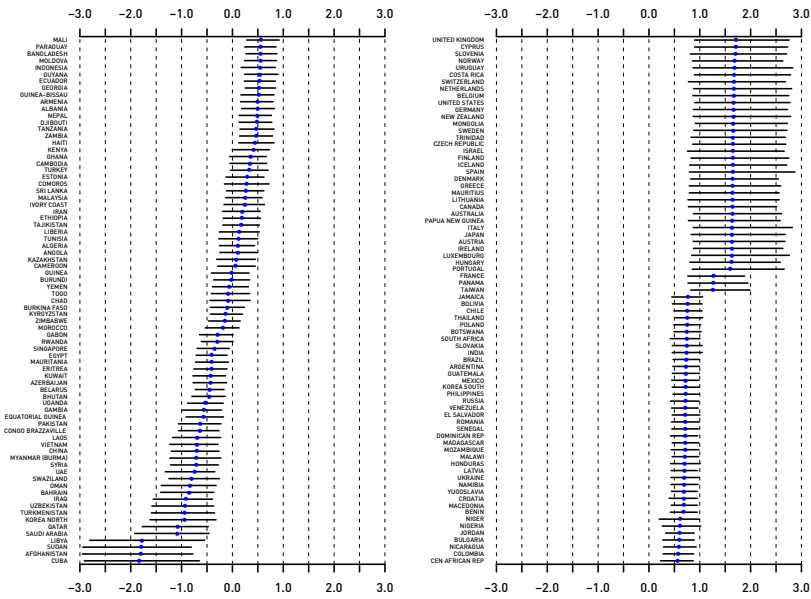


Figure 4: IRT Measures for 2000. Countries are ordered by their posterior means. Error bars indicate 95% highest posterior density regions.

Slide 17

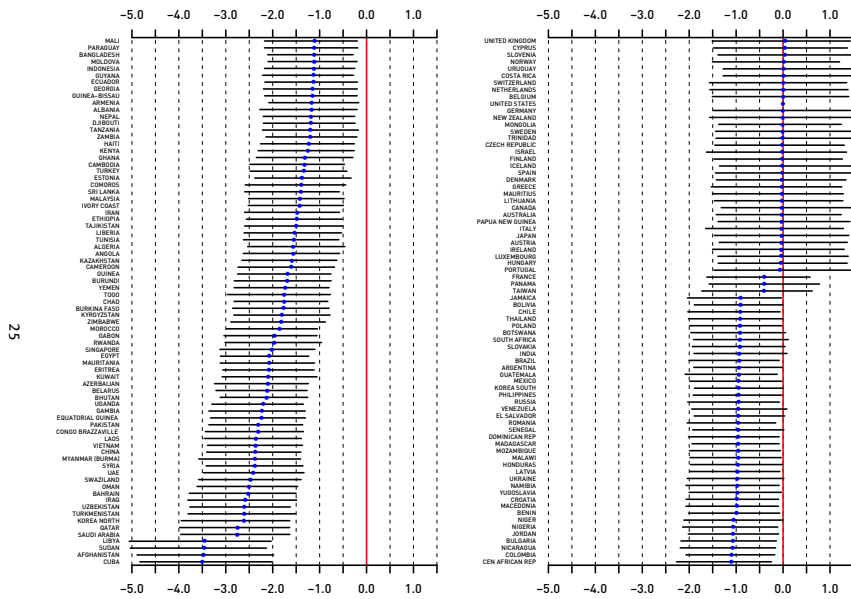


Figure 5: Difference from U.S. Posterior mean of difference between country measure and the score for the U.S., ordered by posterior means. Error bars are 95% highest posterior density regions.

Slide 18

Comparing IRT to CTT

- Fan, X. (1998). "Item response theory and classical test theory: An empirical comparison of their item/person statistics." *Educational and Psychological Measurement*, 58(3), 357-381.
- "Does complex analysis (IRT) pay any dividends in achievement testing?" John O. Anderson University of Victoria available at <http://www.educ.uvic.ca/epl/faculty/anderson/Paper2.doc>
- Stage C (2003) "Classical test theory or item response theory: The Swedish Experience", at Umea Universitet available in Nov 2004 at the following web address:
<http://www.umu.se/edmeas/publikationer/pdf/em%20no%2042.pdf>

A basic summary of these paper is that the results of the two methods are quite similar

Panter AT and Reeve BB (2002) Assessing tobacco beliefs among youth using item response theory models. *Drug and Alcohol Dependence*, 68 (Supplement), S21-S39.

Questions about the paper:

In section 1.2 the authors draw comparisons between CTT (classical test theory, i.e. $x = f + e$) and IRT. Describe one of the differences they give trying not to simply copy what is in the paper.

- One of the stated differences is that in CTT the model parameters depend on the sample of respondents whereas in IRT the model parameters (somehow) do not. I am not convinced by this. See references on previous slide.
- Another difference is the measure of precision. In CTT only a single estimate of reliability is assumed and estimated across all values of the underlying factor, whereas in IRT the reliability is a function of the underlying latent variable.
- Another difference is that IRT can handle observed variables of mixed types (binary, polytomous with different number of categories), whereas CTT assumes all the observed variables are continuous.
- Another stated difference is that CTT uses a simple sum score and thus gives equal weight to each question whereas IRT takes in to account the fact that some questions measure the underlying latent variable better and differentially weights. While it is common to use a simple sum score in CTT it also possible to calculate a factor score which is a weighted sum of the items (weighted so that items with larger loadings contribute more to the scale).

Slide 19

Questions about the paper (continued):

On page S25 the authors talk about a three-parameter logistic model. Describe how the shape of this model differs from the two-parameter IRT model described in class.

Slide 20

Slide 21

Questions about the paper (continued):

What are the four assumptions of the IRT model (Described in Section 1.4)?

The four assumptions are:

1. Unidimensionality - There is one underlying latent variable that must also be continuous and ranging from negative to positive infinity
2. Conditional Independence - All association among the observed variables should be explained by the underlying relationship with the latent variable
3. Normality of latent variable
4. Correct model - The chosen model fits the data better than other models (This isn't exactly an assumption, the assumption would be that the specified model is correct)

Slide 22

Questions about the paper (continued):

Describe what the authors found with regard to DIF in the analysis of the smoking questions?

DIF (Differential item functioning) analysis investigates whether an item functions differently for some groups than others despite the two groups maybe not being similar on the underlying construct (i.e. at specific levels of the underlying trait, DIF asks if the response probability for a particular question is the same). In this paper, the researchers found that there were, in fact, questions that functioned differently for various groups. Specifically, the "social comfort" item was found to be slightly more salient for older male smokers than for older female smokers. The "weight" item was less related to the underlying trait (utility of smoking) for the younger non-smoking females than the older non-smoking females.

Additional point(s): One thing to be noticed is that at the beginning of the results the authors have found two underlying factors, but because IRT is geared more toward fitting one underlying factor, the three questions measuring "Harm" are completely dropped out of the subsequent analysis.