

Latent Variable Modeling - PUBH 5482

9:45 - 11:00 Tuesday and Thursday Fall 2004

Melanie M. Wall
Division of Biostatistics
School of Public Health
University of Minnesota

Office: A426 Mayo
email: melanie@biostat.umn.edu

Slide 1

What is a latent variable?

- A variable that is not observable or is not directly measurable.
- A variable that is measured with error or can only be measured with error.
- A latent variable can be used to represent a 'True' variable measured with error OR a Hypothetical construct.

Examples: liberalism, quality of life, self-esteem, social economic status, unhealthy dieting, math ability, parenting skill, satisfaction, social support, sexual maturity, speech difficulties, asthma severity, self-restraint problems, etc.

- Whether the latent variable is a real thing or a summarization of a complex concept the statistical model will be the same.

Slide 2

Latent variables as mathematical convenience

Latent variables are also used in different statistical modeling techniques as a mathematical convenience where they often are not of primary interest:

- Unobserved heterogeneity (e.g. frailties in survival analysis, random effects in longitudinal data or clustered data)
- Missing data
- Counterfactuals or 'potential outcomes'

Slide 3

Need for measuring latent variables in Health Sciences

Traditionally latent variables and methods for measuring them have been dealt with in the realm of psychology, sociology and education.

- Need in clinical research
- Need in assessment/decisions in health services
- Need in behavioral public health

Slide 4

Slide 5

Need in clinical research

“In the past 20 years or so, the situation in clinical research has become more complex. The effects of new drugs or surgical procedures on *quantity* of life is likely to be marginal. Conversely, there is increased awareness of the impact of health and health care on the *quality* of human life. Therapeutic efforts in many disciplines of medicine- psychiatry, respirology, rheumatology, oncology-other health professions-nursing, physiotherapy, occupational therapy- are directed equally if not primarily to the improvement of quality, not quantity of life. If the efforts of these disciplines are to be placed on a sound scientific basis, methods must be devised to measure what was previously thought to be unmeasurable, and assess in a reproducible and valid fashion those subjective states which cannot be converted into the position of a needle on a dial.”

From: Streiner, D.L. and Norman, G.R. (2001) *Health Measurement scales: A practical guide to their development and use* 3rd ed. Oxford Medical Publications.

Examples: quality of life, stress, pain worsening, compliance with treatment, perimenopause, stages of Alzheimers, diagnostic for myocardial infarction

Slide 6

Need in assessment/decisions in health services

Example: “Your interest is in developing programs of interventions to aid individuals who are considering undergoing genetic testing for inheritable cancer. In particular, you are interested in identifying decision-making concerns among women who are at familial risk for breast cancer. By identifying these concerns, you will be better able to devise counseling programs that are specific to this vulnerable group. Need a standardized scale that will provide you and your colleagues with a reliable, valid, and easy-to-use assessment of the genetic testing and cancer outcome concerns of this population.”

From: Pett, M.A., Lackey, N.R., Sullivan, J.J. (2003) *Making sense of factor analysis: The use of factor analysis for instrument development in health care research* Sage Publications.

Examples: proneness to falls, risk for developing perineal dermatitis

Need in behavioral public health

So many health issues are directly related to our behavior. Much theoretical work done in sociology building theories to describe different aspects of personal and social/familial networks that influence our behaviors. Many of these phenomena are not things easily measured.

Examples:

peer relationship strains	
family connectedness	
intimidation/bully-ing	dieting and other weight control behavior
self efficacy for screening practices	physical activity
taste preferences	alcohol consumption
body satisfaction	violence in the workplace
motivation of alcoholism treatment	
post traumatic stress disorder	

Slide 7

Latent Variable Modeling Course

Two general areas

- Measurement of latent variables - Statistical models are used to describe the way that observed variables are related to the latent variables.
- Path Analysis - Statistical models are used to evaluate the **presumed** causal relations (direct and indirect) among several variables (possibly latent).

Slide 8

Measurement Models

NAMES OF MODELS: exploratory and confirmatory factor analysis, latent trait models or item response theory models or Rasch models, latent class models or latent mixture models or hidden Markov models

- Measures the latent variables
- reduces the dimensionality of the data
- find patterns of correlations among several observed variables that are measuring the same thing
- observed variables are just a reflection of some underlying phenomena (i.e. latent variable)
- goal is to lose as little information as possible when reducing the dimensionality
- goal is to quantify how well each observed variables actually measure the latent variable

Slide 9

Measurement Models

Like observed variables, latent variables can be (hypothesized to be) continuous or categorical and if they are categorical they can be ordinal (ordered) or nominal (unordered). Depending upon what is assumed about the distribution of the latent variable and upon what kind of observed variables are used to measure them (i.e. continuous or categorical), the method for estimating the measurement model will change.

	latent continuous	latent discrete
observed continuous	factor analysis	latent mixture model
observed discrete	latent trait model	latent class model

Slide 10

Path analysis

- allows researcher to translate idea about how causes are related to effects into a model
- total effect of one variable on another can be broken down into direct and indirect effects
- mediation and moderation
- Can be used with or without latent variables, that is variables of interest can be observed directly (i.e. no need for a measurement model).
- When there are no latent variables or when latent variables are treated as if they can be observed, it is often called path analysis, when there are latent variables and the measurement error in them is taken into account statistically by incorporating a measurement model, it is often called structural equation modeling.
- The general ideas are introduced using models without measurement error models included, then the measurement models will be added in later.

Slide 11

Measurement of latent variables

When the latent variable is continuous, we will refer to it as a “latent factor”

When the latent variable is categorical, we will refer to it as a “latent class”

Slide 12

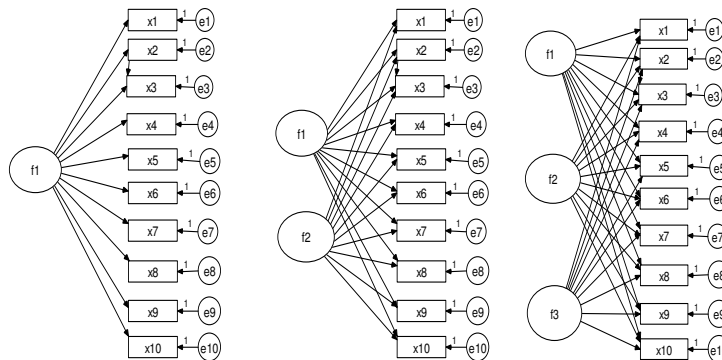
Factor analysis for continuous observed variables - “EFA and CFA”

Credit is given to Spearman (1904) as the originator of factor analysis

Exploratory factor analysis (EFA) general purposes:

- To determine how many underlying factors are necessary to explain most of the correlations and variance in the data.
- To determine the relationship via **rotation** between each of these underlying factors with each of the observed variables in a meaningful way so that the factors can be interpreted and named.
- To weed out observed variables that do not tend to measure well the underlying factors shared by the other variables.
- To propose blocks of variables that may be subsequently be used to create a simple sum scale.
- To propose a CFA model

Slide 13



Slide 14

In EFA every element in Λ is estimated and it is assumed that Ψ is diagonal. Also, it is common to assume that $Var(\mathbf{f}) = \Phi = \mathbf{I}$, i.e. the factors are uncorrelated with variance 1 (but this is not a necessary assumption, it is dropped when examining oblique rotations).

EFA for developing scales from questionnaires

Example:

Concerns about decisions to get genetic testing (from Pett, Lackey, Sullivan 2003)

Slide 15

Slide 16

THESE ARE SOME ISSUES THAT PEOPLE HAVE INDICATED ARE CONCERNS WHEN MAKING THEIR DECISION ABOUT GENETIC TESTING	IS THIS A CONCERN FOR YOU?				
	1 Not at all	2 Slightly	3 Moderately	4 Quite a bit	5 Extremely
1. I might increase my sense of personal control over the condition after I receive the testing results.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. I worry about being faced with an uncertain diagnosis.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. I want to know what I should do to manage my risk for cancer.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. I hope that knowing the results will help reduce my uncertainty about the future.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. I fear being faced with the ambiguity of the results.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. I worry about being faced with a diagnosis I don't know what to do about.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. I hope to be able to make better health and lifestyle choices as a result.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. I am worried about being able to maintain health and life insurance coverage.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. I am worried about how my family will react to the testing information.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. I need information about the types of cancers I am at risk for.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. I am concerned about marital/family problems that might occur from obtaining the test results.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12. I might be helped to make important future life decisions by knowing I carry the gene, e.g., getting married, having children.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13. I need information about participation in future screening activities.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14. I want to know how a positive test will affect my children and other family members.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15. I am worried about my future life.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16. I want information about the difference between the diagnosis of having the gene and getting cancer.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17. I have financial concerns related to the screening.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18. I want to know the financial and social implications of being identified as a carrier.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
19. I worry about being targeted as a carrier.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
20. I want to know my survival prospects.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Slide 17

Table 6.4 Content Areas and Factors on Which the 20 CGTS Items Loaded

Item	Content Area	Factor			
		1	2	3	4
C1	Personal control
C3	How to manage risk
C7	Make better lifestyle choices
C10	Need cancer information
C12	Helped to make future life decisions
C16	Information re diagnosis
C13	Need screening information
C20	Information re survival prospects
C6	Worry about the diagnosis
C2	Uncertain diagnosis
C4	Reduce uncertainty
C5	Fear ambiguity
C15	Worried about future life
C8	Health, life insurance
C18	Financial and social implications
C19	Being targeted as a carrier
C17	Financial concerns
C9	Family reactions
C11	Marital, family problems
C14	Effect of positive test on family members

④ C20 fits best with these items.

① C6 fits best with these items.

③ C17 fits best with these items.

② C14 fits best with these items.

EFA for developing scales from questionnaires

Examples:

Adolescent body satisfaction (from Neumark-Sztainer, et.al 2003)

Coping with Cystic Fibrosis questionnaire (from Patterson, et.al 2004)

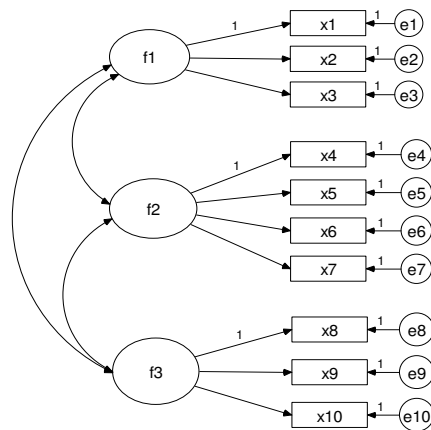
Slide 18

Factor analysis for continuous observed variables - “EFA and CFA”

Confirmatory factor analysis (CFA) general purposes:

- To define a measurement model for the relationship between multivariate observations and underlying factors
- To test the statistical significance of factor loadings and correlations. Note this testing cannot currently be done in the EFA model. Thus one may be interested in testing whether rotated factor loadings from an EFA that look “close to zero” are, in fact, significantly different from zero or not.
- To test whether the measurement model for one group is the same as the measurement model for some other group
- As a precursor to a Structural equation model

Slide 19



Slide 20

In CFA usually several elements in Λ are fixed to zero and it is possible to consider correlated ϵ which means that Ψ is not necessarily diagonal. Furthermore, it is usually assumed that the factors are correlated so that no restriction is placed on Φ .

Factor analysis for ordered categorical observed variables - “Latent Trait Models”

Comes from Education testing, (latent variable are labeled as traits), Item Response Theory (IRT), large literature related to IRT

Answer (0,1) to a series of p questions, thus there are 2^p possible response patterns (dichotomous data). Answer (1,2,... c) to a series of p questions, thus there are c^p possible response patterns (polytomous data).

Slide 21

Questions to answer:

1. How much of the differences in these responses can be explained by supposing all items depend on one or more continuous latent variables?
2. How many underlying variables are there?
3. Which observed variables help discriminate individuals the best?
4. What is the best way to combine the observed variables in order to create a scale or score for each individual?

Factor analysis for ordered categorical observed variables

A web site devoted to Item Response Theory (another name for Latent trait models) is:

<http://www.education.umd.edu/Depts/EDMS/tutorials/frontpage.html>

EXAMPLE: DEMOCRACY AS A LATENT VARIABLE

Slide 22

Democracy as a Latent Variable*

Shawn Treier
Stanford University
satreier@stanford.edu

Simon Jackman
Stanford University
jackman@stanford.edu

July 17, 2003

Slide 23

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 24

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

Slide 25

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 26

	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 27

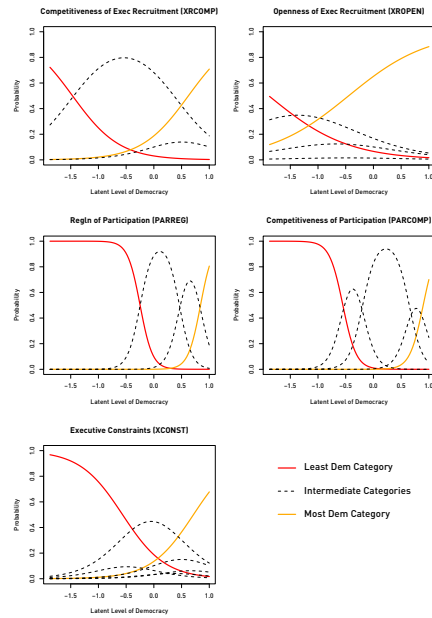


Figure 1: Item Characteristic Curves, Five Polity IV indicators

Slide 28

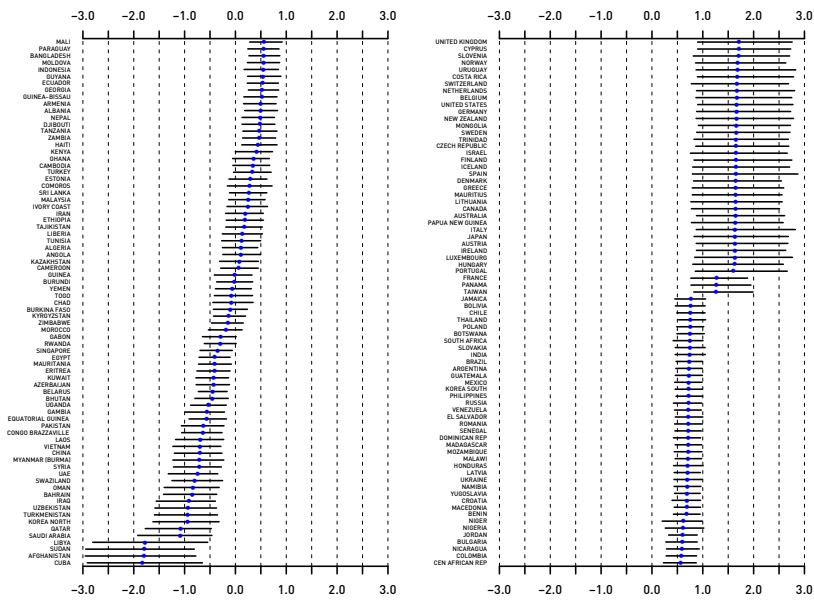


Figure 4: IRT Measures for 2000. Countries are ordered by their posterior means. Error bars indicate 95% highest posterior density regions.

Slide 29

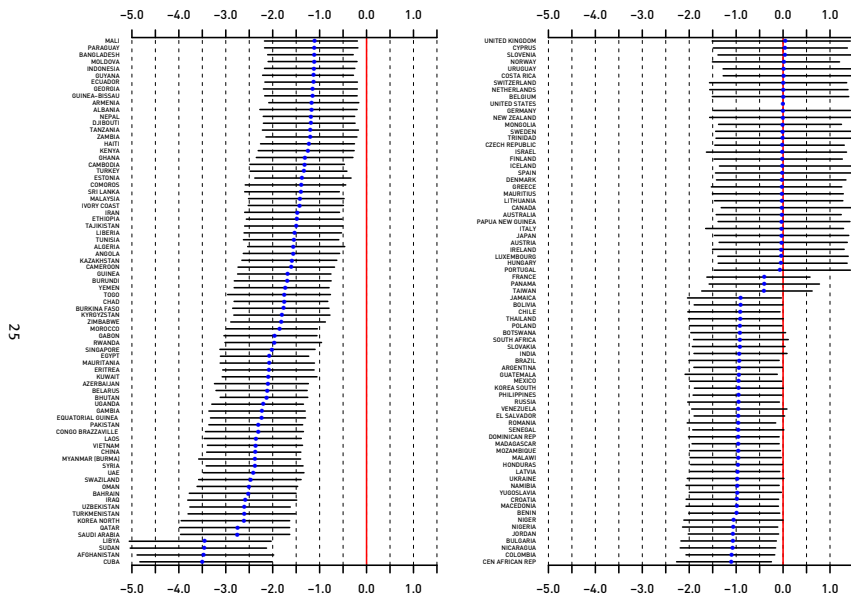


Figure 5: Difference from U.S. Posterior mean of difference between country measure and the score for the U.S., ordered by posterior means. Error bars are 95% highest posterior density regions.

Latent Class Analysis Measuring categorical latent variables

Credit usually given to Paul Lazarsfeld as being the originator of LCA, Foundation book is Lazarsfeld, P.F. and Henry, N.W. (1968) Latent Structure Analysis. Houghton Mifflin.

Latent class analysis is a statistical method for finding subtypes of related cases from multivariate categorical data.

Questions to answer:

1. How many underlying classes are there?
2. What is the prevalence in each of the latent classes?
3. What is the relationship between the observed responses and the latent classes
4. What is the probability that a particular individual will be in a particular class?

Good web site including a FAQ about Latent class models is found at <http://ourworld.compuserve.com/homepages/jsuebersax/index.htm>

Another web site containing materials related to the Sage book, *Latent Class Scaling Analysis*, in the Series: Quantitative Methods in the Social Sciences (126) is <http://www.education.umd.edu/EDMS/Latent/Dayton.html>

Slide 30

Latent Class Analysis

Measuring categorical latent variables

EXAMPLE: Measuring unhealthy weight control behavior. Hypothetically a categorical latent variable.

Have you done the following in the last year in order to lose weight or maintain your weight: (yes, no)

To control weight	marginal	2-class		3-class			4-class			
	1	1	2	1	2	3	1	2	3	4
fasted	17.9	38.8	2.8	58.5	32.6	2.6	24.9	71.4	29.2	2.6
ate little	44.1	92.5	9.0	94.2	89.9	7.9	74.1	100.0	87.5	6.9
diet pills	6.3	13.6	1.1	40.4	6.5	1.2	49.4	31.0	6.1	0.8
vomit	6.3	15.0	0.1	45.0	7.1	0.1	33.2	43.6	5.6	0.1
laxatives	1.6	3.5	0.2	17.1	0.1	0.2	20.7	11.7	0.0	0.2
diuretics	1.4	3.3	0.1	16.1	0.1	0.1	29.4	7.9	0.0	0.1
food substitutes	9.3	19.2	2.1	41.6	13.1	2.1	54.4	34.5	12.0	1.7
skipped meals	44.4	89.7	11.1	85.7	89.7	9.5	43.5	100.0	87.4	8.5
smoked more cigs	9.3	18.6	2.5	39.1	13.1	2.5	9.8	47.6	11.2	2.4
% in each class	100	42.0	58.0	8.4	35.2	56.4	2.6	8.0	34.7	54.7

Estimated θ_{jk} (probability of saying yes to the variable j given that the individual is in latent class k) under latent class models with different K

Slide 31

Latent Mixture Models

categorical latent variables distinguishing longitudinal profiles

Recently there has been a lot of statistical methods work on more complex models for longitudinal data.

Slide 32

Path Analysis

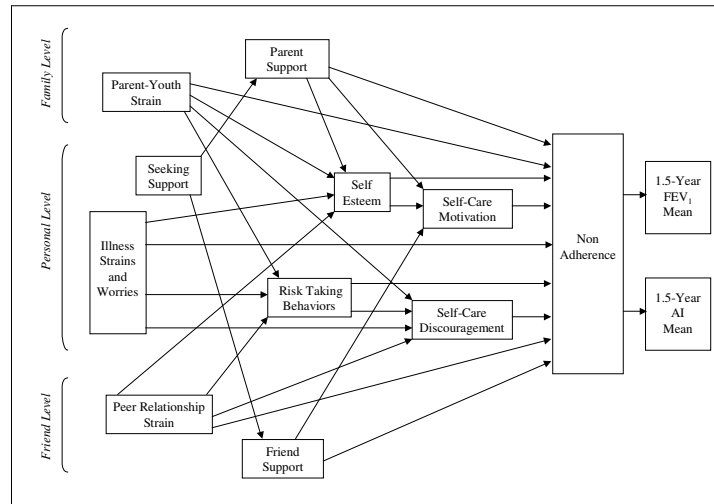
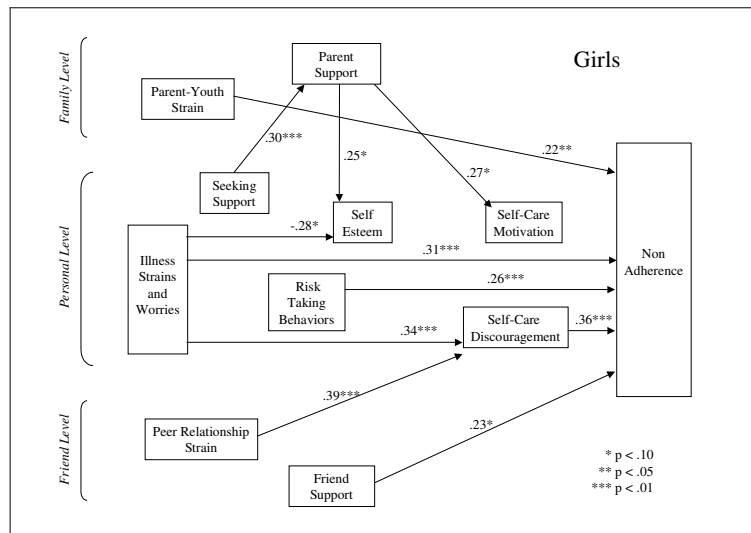


Figure 1. Conceptual model showing relationship between strains, resources, non-adherence feelings/behaviors and health outcomes for youth with CF

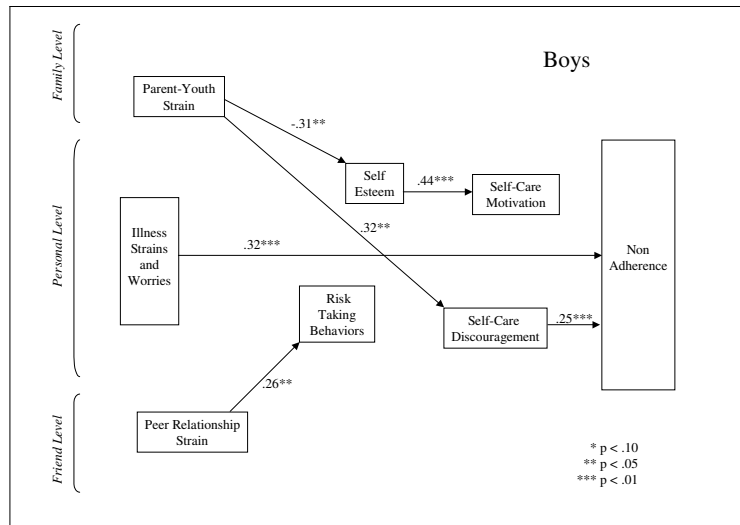
Slide 33

Path Analysis



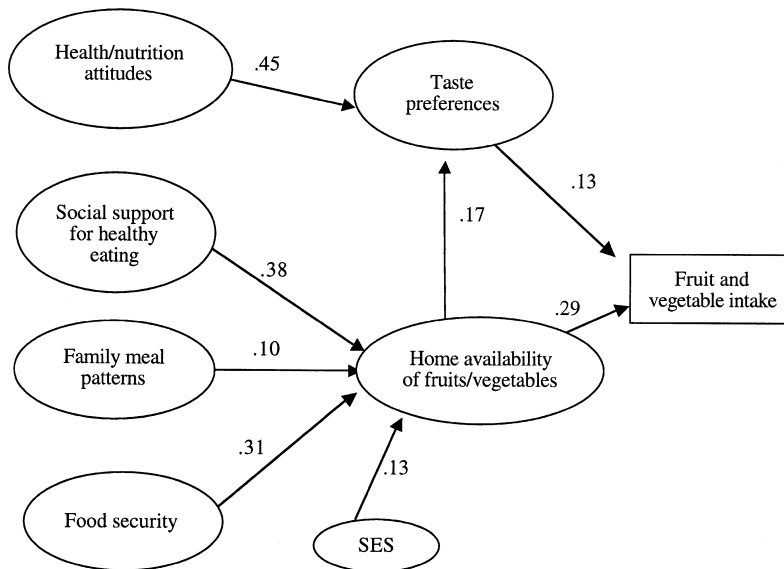
Slide 34

Path Analysis



Slide 35

Structural equation model



Slide 36