Latent Class Regression Analysis

Jeroen K. Vermunt

Department of Methodology and Statistics,
Tilburg University

http://spitswww.uvt.nl/~vermunt/

j.k.vermunt@uvt.nl

Workshop at the 24th Biennial Conference of SMABS,
17-22 of July 2004 at Jena University

Introduction: Latent Class or Mixture Models

What is a (finite) mixture model?

- parameters in the statistical model of interest differ across unobserved subgroups (or latent classes)
- a latent class (LC) model is, in fact, the same thing as a mixture model

Four main application types:

- clustering (model based / probabilistic)
- scaling (discretized IRT)
- random-effects modeling (mixture regression / NPML)
- density estimation

Latent GOLD software takes these application types as starting point in four modules:

- LC cluster, LC factor, LC regression, LC choice

Main differences between these modules arise from the data structure:

- LC cluster/factor: multiple indicators
- LC regression/choice: single dependent variable that is repeatedly observed

Other differences:

- LC factor: multiple ordinal latent variables
- LC choice: special regression model for choice/ranking/rating data

LC regression model

- in fact, the most general LC model (LC cluster model can usually be reformulated as a LC regression model)
- limitation: single scale type for “indicators”

Latent Class or Mixture Regression Analysis

Basic idea: parameters of a regression model differ across unobserved classes

Two-level data structure: single dependent variable is observed various times (replications / repeated measures / nested observations)

Regression model from Generalized Linear Modeling (GLM) family

- Logistic regression (binary, multinomial, ordinal, binomial count)
- Poisson regression
- Linear regression

Predictors, whose values may change across replications, affect the dependent variable
Covariates may be used to predict class membership

Some references: Agresti (2002); Aitkin (1999); Skrondal and Rabe-Hesketh (2004); Vermunt and Van Dijk (2001); Wedel and DeSarbo (1994)

More advanced models:

- Truncated, censored, and zero-inflated dependent variables
- Models with random effects
- Multilevel extension

Some application types of LC regression

- extensions of standard LC cluster analysis
  - discretized IRT (Rasch, rating scale, linear logistic test model, etc.)
  - multiple group LC (item bias)
- nonparametric random-effects modeling (multilevel, overdispersion)
- clustering based on a regression model (conjoint studies / clinical trials / growth analysis)

Other software for LC regression analysis

- Mplus and LEM (1-level multiple-indicator model with restrictions)
- GLIMMIX & GLLAMM (2-level model)

(SEM versus multilevel discussion)

Today's program

I. From a standard LC cluster model to a LC binary logistic regression model

II. Dealing with counts: mixture Poisson density estimation and mixture Poisson regression

III. Classical multilevel application: linear model with nonparametric random effects

IV. Repeated measures: LC growth modeling

V. Event history analysis: Cox regression model for correlated events

VI. Modeling choice data

VII. Further topics: censoring and truncation, complex sampling, and multilevel LC analysis

I. From a Standard LC Cluster Model to a LC Binary Logistic Regression Model

Data from the SOCON survey (Heinen, 1996)

Five dichotomous "women's liberation" items (0=agree, 1=disagree):

1) Women's liberation sets women against mean
2) It's better for a wife not to have a job because that always poses problems in the household, especially if there are children
3) The most natural situation occurs when the man is the breadwinner and the woman runs the household and takes care of the children
4) It isn't really as important for a girl to get a good education as it is for a boy
5) A women is better suited to raise small children than a man

Covariate: Gender (0=male; 1=female)

Data structure (first 5 cases)

ID	Female	Item1	Item2	Item3	Item4	Item5
1	0	1	1	0	1	1
2	1	0	0	1	1	0
3	1	0	0	0	1	1
4	0	1	1	1	0	1
5	1	0	0	1	0	1

Notation

Y_{ij}: response of case i on item j; J: # of items
X_i: latent class variable; t: particular class; T: # of classes
Z_{ik}: kth covariate

Realizations: lower case; Vectors: bold case

Basic formula LC cluster model (assumptions: discrete mixture and local independence)

$$P(\mathbf{Y}_i = \mathbf{y}) = \sum_{t=1}^{T} P(X_i = t) \prod_{j=1}^{J} P(Y_{ij} = y_j \mid X_i = t)$$

Binary logistic form for probabilities (item-specific coefficients)

$$P(Y_{ij} = 1 \mid X_i = t) = \frac{\exp(\alpha_j + \beta_{jt})}{1 + \exp(\alpha_j + \beta_{jt})}$$

$$\text{logit}\,[P(Y_{ij} = 1 \mid X_i = t)] = \alpha_j + \beta_{jt}$$

$$\text{with } \sum_{t=1}^{T} \beta_{jt} = 0 \text{ or } \beta_{j1} = 0$$

LC cluster model with covariates:

$$P(\mathbf{Y}_i = \mathbf{y} \mid \mathbf{z}_i) = \sum_{t=1}^{T} P(X_i = t \mid \mathbf{z}_i) \prod_{j=1}^{J} P(Y_{ij} = y_j \mid X_i = t, \mathbf{z}_i)$$

Covariate "Female" affects X_i (concomitant variable model):

$$P(X_i = t \mid z_{i1}) = \frac{\exp(\gamma_{0t} + \gamma_{1t} z_{i1})}{\sum_{t'=1}^{T} \exp(\gamma_{0t'} + \gamma_{1t'} z_{i1})}$$

Covariate "Female" affects Y_{ij} (model with item bias):

$$P(Y_{ij} = 1 \mid X_i = t, z_{i1}) = \frac{\exp(\alpha_{0j} + \alpha_{1j} z_{i1} + \beta_{jt})}{1 + \exp(\alpha_{0j} + \alpha_{1j} z_{i1} + \beta_{jt})}$$

Fit measures for the estimated models (minor differences causes by priors)

	LC Cluster				LC Regression			
	BIC(LL)	L^2	df	p	BIC(LL)	L^2	df	p
A1. 1-class	6674.5	709.0	26	0.00	6674.5	709.0	26	0.00
A2. 2-class	6075.7	67.9	20	0.00	6075.7	67.9	20	0.00
A3. 3-class	**6063.9**	13.9	14	0.46	**6063.4**	13.5	14	0.49
A4. A3 + Rasch					6045.5	51.8	22	0.00
A5. 2-class mixture Rasch					6036.8	22.0	19	0.28
B1 1-class + sex	6674.5	816.9	57	0.00	6674.5	816.9	57	0.00
B2 2-class + sex	6058.4	151.5	50	0.00	6058.4	151.5	50	0.00
B3 3-class + sex	6039.6	83.5	43	0.00	6039.3	83.2	43	0.00
B4. B3+sex*(item3+item5)	**6020.7**	50.5	41	0.15	**6020.3**	50.1	41	0.16
B5. B4+class*sex*(item3+item5)					6046.2	47.8	37	0.11
B6. B4 + Rasch					6005.8	91.9	49	0.00
B7. 2-class mixture Rasch + sex					6005.1	91.2	49	0.00

Now same model in LC regression form

Data structure (first two cases)

ID	ItemNr	Female	Response	Dum1	Dum2	Dum3	Dum4	Dum5	Dum3Fem	Dum5Fem
1	1	0	1	1	0	0	0	0	0	0
1	2	0	1	0	1	0	0	0	0	0
1	3	0	0	0	0	1	0	0	0	0
1	4	0	1	0	0	0	1	0	0	0
1	5	0	1	0	0	0	0	1	0	0
2	1	1	0	1	0	0	0	0	0	0
2	2	1	0	0	1	0	0	0	0	0
2	3	1	1	0	0	1	0	0	1	0
2	4	1	1	0	0	0	1	0	0	0
2	5	1	0	0	0	0	0	1	0	1

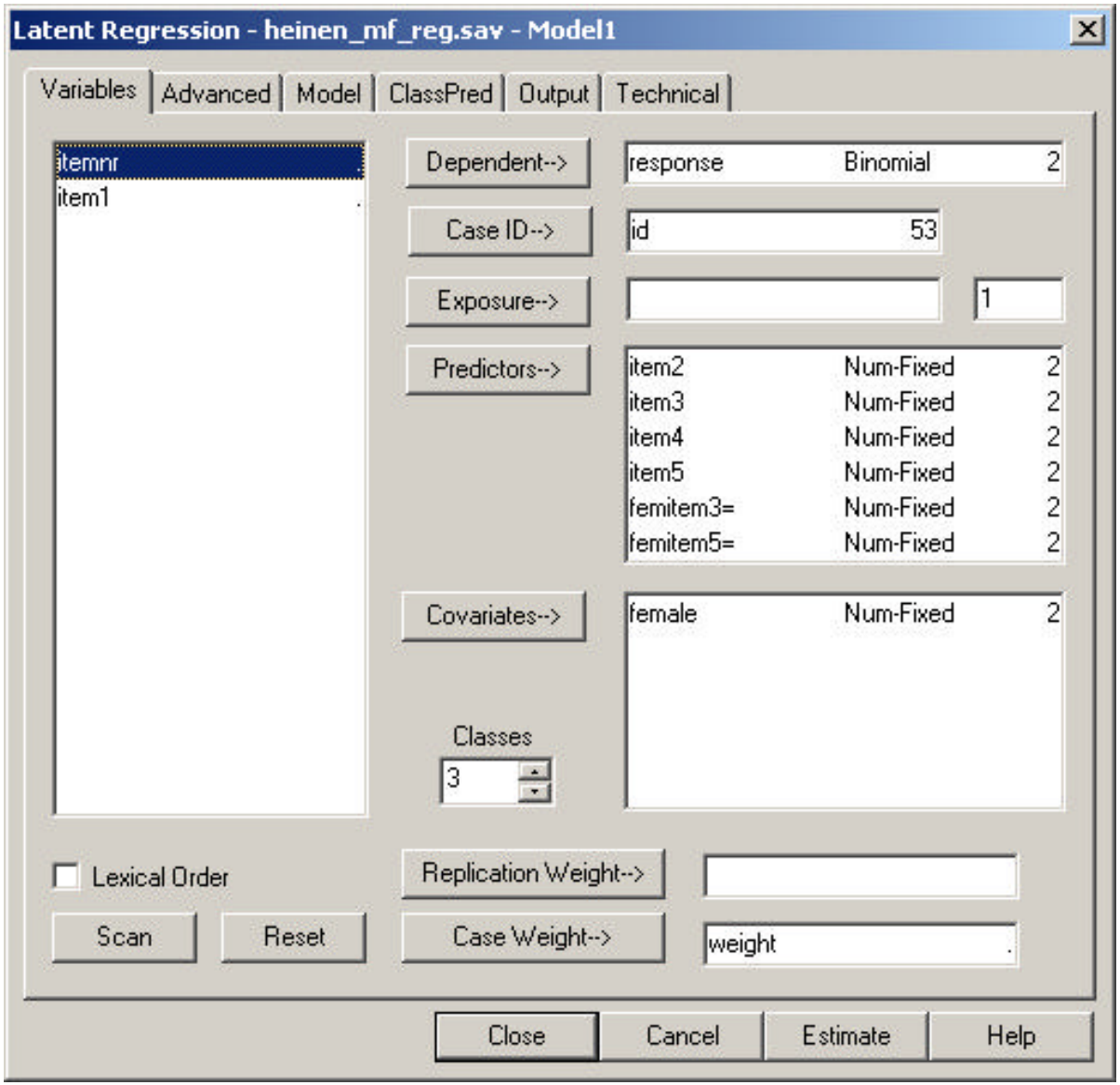
Item dummies serve as predictors z_{ijk} whose values may vary across the J_i replications of case *i*. General model form:

$$P(\mathbf{Y}_i = \mathbf{y} \mid \mathbf{z}_i) = \sum_{t=1}^{T} P(X_i = t \mid \mathbf{z}_i) \prod_{j=1}^{J_i} P(Y_{ij} = y_j \mid X_i = t, \mathbf{z}_{ij})$$

$$P(Y_{ij} = 1 \mid X_i = t, \mathbf{z}_{ij}) = E(Y_{ij} \mid X_i = t, \mathbf{z}_{ij}) = \frac{\exp(\alpha_t + \sum_{k=1}^{K} \beta_{kt} z_{ijk})}{1 + \exp(\alpha_t + \sum_{k=1}^{K} \beta_{kt} z_{ijk})}$$

Note:

- local independence
- regression coefficients do NOT vary between replications
- regression coefficients (may) vary between classes


Latent Regression - heinen_mf_reg.sav - Model1
Variables
Advanced
Model
ClassPred
Output
Technical
itemnr
item1
Dependent-->
response Binomial 2
Case ID-->
id 53
Exposure-->
1
Predictors-->
item2 Num-Fixed 2
item3 Num-Fixed 2
item4 Num-Fixed 2
item5 Num-Fixed 2
femitem3= Num-Fixed 2
femitem5= Num-Fixed 2
Covariates-->
female Num-Fixed 2
Classes
3
Lexical Order
Replication Weight-->
Scan
Reset
Case Weight-->
weight
Close
Cancel
Estimate
Help

Class sizes and item probabilities for males and females (model B4)

	Male			Female		
	Class3	Class1	Class2	Class3	Class1	Class2
Class size	0.13	0.54	0.33	0.10	0.36	0.54
Item1	0.12	0.44	0.78	0.12	0.44	0.78
Item2	0.07	0.71	0.96	0.07	0.71	0.96
Item3	0.00	0.29	0.94	0.00	0.14	0.86
Item4	0.52	0.94	0.99	0.52	0.94	0.99
Item5	0.01	0.20	0.69	0.02	0.35	0.82

Classification is based on the posterior class membership probabilities

$$P(X_i = t \mid \mathbf{Y}_i = \mathbf{y}, \mathbf{z}_i) = \frac{P(X_i = t \mid \mathbf{z}_i) \prod_{j=1}^{J} P(Y_{ij} = y_j \mid X_i = t, \mathbf{z}_i)}{\sum_{t=1}^{T} P(X_i = t \mid \mathbf{z}_i) \prod_{j=1}^{J} P(Y_{ij} = y_j \mid X_i = t, \mathbf{z}_i)}$$

Classification table: membership probabilities versus modal allocation (model B4)

	Modal			
Probability	Class3	Class1	Class2	Total
Class3	110.4	18.1	0.0	128.6
Class1	48.5	381.4	78.2	508.1
Class2	0.1	44.4	452.8	497.4
Total	159.0	444.0	531.0	1134.0

Classification errors=0.17; Reduction of errors (Lambda)=0.70

Posterior mean predicted values are weighted averages of the class-specific predicted values, with the posterior membership probabilities as weights:

$$E_i(Y_{ij} \mid \mathbf{z}_i) = \sum_{t=1}^{T} P(X_i = t \mid \mathbf{Y}_i = \mathbf{y}, \mathbf{z}_i) E(Y_{ij} \mid X_i = t, \mathbf{z}_i)$$

These predicted values are used in R^2 formulas, e.g., R^2 based on Squared Error equals

$$R^2 = 1 - \frac{\sum_{i=1}^{I} \sum_{j=1}^{J_i} [Y_{ij} - E_i(Y_{ij} \mid \mathbf{z}_i)]^2}{\sum_{i=1}^{I} \sum_{j=1}^{J_i} [Y_{ij} - E(Y_{ij})]^2} = 1 - \frac{0.1100 \cdot 5670}{0.2321 \cdot 5670} = 0.5262$$

Note: 5970 = total number of replications

Posterior mean estimates of the coefficients for case *i* are obtained as weighted averages of the class-specific coefficients, with the posterior membership probabilities as weights

$$E_i(\beta_k) = \sum_{t=1}^{T} P(X_i = t \mid \mathbf{Y}_i = \mathbf{y}, \mathbf{z}_i)\beta_{kt}$$

$$\sigma_i(\beta_k) = \sqrt{\sum_{t=1}^{T} P(X_i = t \mid \mathbf{Y}_i = \mathbf{y}, \mathbf{z}_i)[\beta_{kt} - E_i(\beta_k)]^2}$$

Predicted values can also be obtained by filling in these individual estimates in the regression equation (=HB prediction)

$$E_i(Y_{ij} \mid \mathbf{z}_{ij}) = \frac{\exp[E_i(\alpha) + \sum_{k=1}^{K} E_i(\beta_k) z_{ijk}]}{1 + \exp[E_i(\alpha) + \sum_{k=1}^{K} E_i(\beta_k) z_{ijk}]}$$

Variants & Extensions

- Rasch model: item effects class independent
- Equality restrictions on difficulties (adding up item dummies)
- Linear logistic test model: class-independent slopes, difficulties function of design factors
- Item bias: gender*item interactions
- Multiple dimensions via equality restrictions across classes
- Mixture Rasch model: continuous trait

II. Dealing with Counts: Mixture Poisson Density Estimation and Mixture Poisson Regression

Dental health trial on prevention of tooth decay (797 Brazilian children; see Skrondal & Rabe-Hesketh, 2004)

Dependent variable: # of decayed, missing or filled teeth (DMFT)

Explanatory variables

- Treatment: 1 = no treatment; 2 = oral health education; 3 = school diet enriched with ricebran; 4 = mouthrinse with 0.2% NaF solution; 5= oral hygiene; 6 = all four treatments
- Ethnic group: 1= brown; 2 = white; 3 = black
- Gender: 1 = male; 2 = female

Note: Single replication per case

Big problem in modeling counts: overdispersion caused by unobserved heterogeneity

Modeling options for counts (no predictors for the moment)

Poisson:

$$E(Y_i) = \exp(\alpha)$$

Finite mixture Poisson:

$$E(Y_i \mid X_i = t) = \exp(\alpha_t)$$

Zero-inflated Poisson (T=2):

$$E(Y_i \mid X_i = 1) = \exp(\alpha_1);\ E(Y_i \mid X_i = 2) = 0$$

Poisson-Normal and Poisson-Gamma:

$$E(Y_i \mid u_i) = \exp(\alpha + u_i) \text{ with } u_i \sim N(0, \tau^2) \text{ and } \exp(u_i) \sim \text{Gamma}(1/\tau^2)$$

Frequencies and fit measures

# DMFT	observed	1-class	2-class	3-class	ZIN	Normal	Gamma
0	231	124.76	231.33	231.33	231.00	185.89	201.47
1	163	231.36	161.54	161.54	138.68	221.93	211.10
2	140	214.53	143.84	143.84	164.08	165.82	156.51
3	116	132.61	118.03	118.03	129.42	101.80	100.05
4	70	61.48	76.23	76.23	76.56	56.90	58.85
5	55	22.80	39.61	39.61	36.23	30.43	32.81
6	22	7.05	17.17	17.17	14.29	15.99	17.62
LL		-1500.6	-1421.0	-1421.0	-1425.7	-1453.5	-1444.5
BIC		3007.9	2862.0	2875.4	2864.8	2920.4	2902.4
Npar		1	3	5	2	2	2
L^2		185.0	25.8	25.8	35.2	90.9	72.9

Note: 2-class model is NPML estimator

LC Poisson regression models with predictors (and normal random effects):

$$E(Y_i \mid X_i = t, \mathbf{z}_i) = \exp(\alpha_t + \sum_{k=1}^{K} \beta_k z_{ik} + u_i)$$

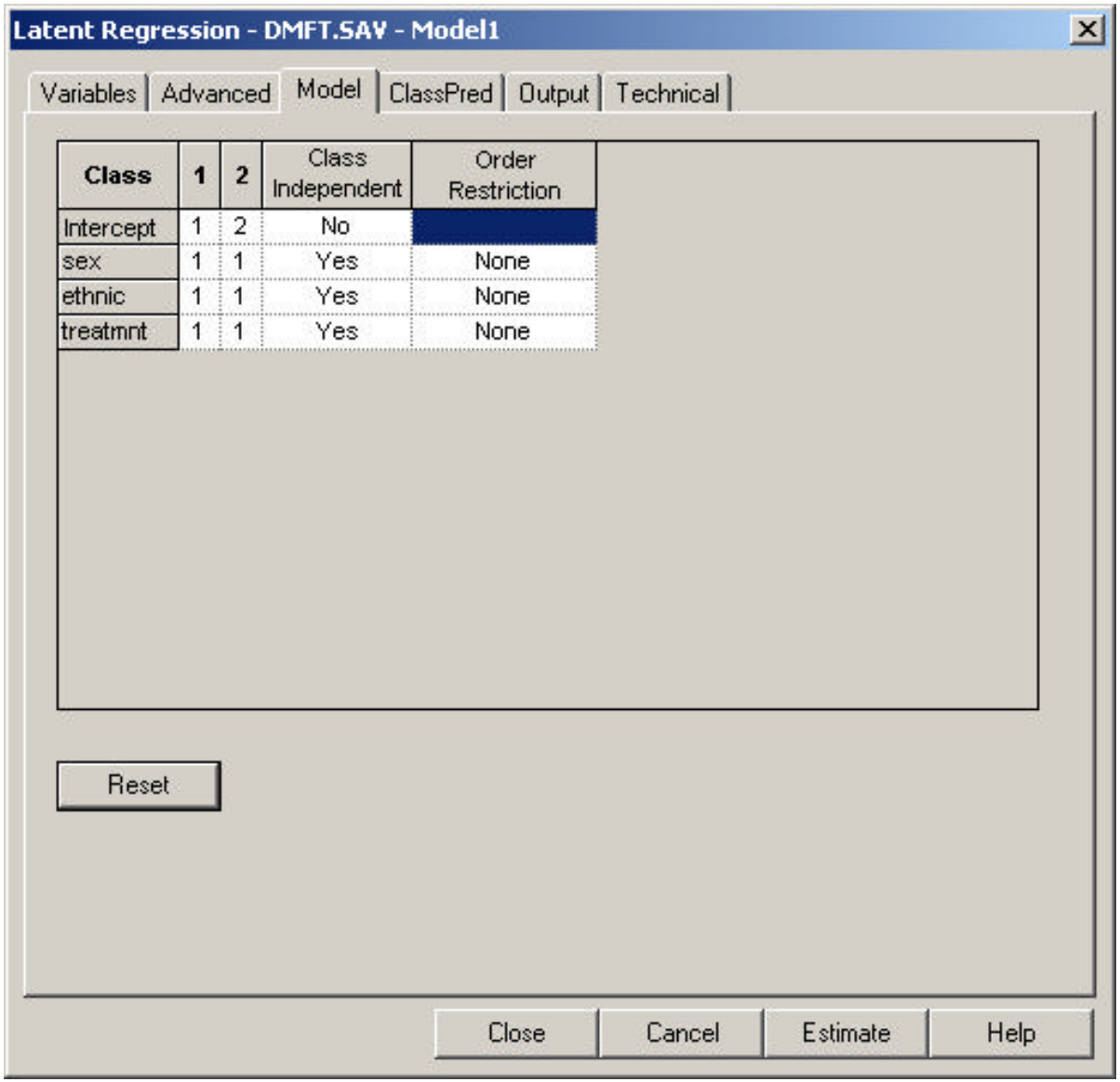
Fit measures and regression coefficients

	1-class	2-class	normal
LL	-1469	-1406	-1433
BIC	2998	2886	2932
Intercept			
class1	0.55 (0.03)	0.86 (0.05)	0.41 (0.05)
class2		-1.28 (0.36)	
mean ($\bar{\alpha}$)	**0.55**	**0.23**	**0.41**
std.dev. (τ)	**0.00**	**0.98**	**0.54** (0.04)

$$\bar{\alpha} = \sum_{t=1}^{T} \alpha_t P(X = t); \quad \tau = \sqrt{\sum_{t=1}^{T} [\alpha_t - \bar{\alpha}]^2 P(X = t)}$$

Table Continued

	1-class	2-class	normal
Sex			
female	-0.07 (0.03)	-0.05 (0.03)	-0.07 (0.03)
male	0.07 (0.03)	0.05 (0.03)	0.07 (0.03)
Ethnic			
brown	0.01 (0.04)	0.01 (0.05)	0.02 (0.05)
white	0.11 (0.04)	0.10 (0.05)	0.12 (0.05)
black	-0.13 (0.05)	-0.11 (0.06)	-0.14 (0.07)
Treatment			
control	0.26 (0.05)	0.22 (0.06)	0.27 (0.07)
educ	0.03 (0.06)	-0.02 (0.07)	0.04 (0.07)
enrich	0.17 (0.06)	0.14 (0.06)	0.18 (0.07)
rinse	-0.09 (0.06)	-0.04 (0.07)	-0.10 (0.08)
hygiene	-0.04 (0.06)	-0.01 (0.07)	-0.05 (0.08)
all	-0.33 (0.07)	-0.28 (0.08)	-0.34 (0.09)


Latent Regression - DMFT.SAV - Model1
Variables
Advanced
Model
ClassPred
Output
Technical
Class
1
2
Class Independent
Order Restriction
Intercept 1 2 No
sex 1 1 Yes None
ethnic 1 1 Yes None
treatmnt 1 1 Yes None
Reset
Close
Cancel
Estimate
Help

III. Classical Multilevel Application: Linear Model with Nonparametric Random Effects

Example taken for Snijders and Bosker's (1999) book on multilevel analysis

2287 pupils nested within 131 schools

Dependent variable: performance on a language test (continuous variable)

Individual-level predictors: IQ, SES

School-level predictors: average IQ, average SES, Groupsize, Combigroup

Parameter estimates and fit measures

	1-class	3-class	random intercept
intercept	41.44 (0.19)	41.06	41.41 (0.36)
iq	2.20 (0.08)	2.17 (0.07)	2.20 (0.07)
ses	0.17 (0.02)	0.17 (0.02)	0.17 (0.02)
sch_iq	1.39 (0.21)	1.54 (0.25)	1.40 (0.35)
sch_ses	-0.11 (0.03)	-0.04 (0.04)	-0.09 (0.05)
combi	-1.86 (0.44)	-1.30 (0.90)	-1.91 (0.80)
gs	-0.03 (0.03)	-0.06 (0.06)	-0.02 (0.05)
τ	0.00	2.79	2.68 (0.23)
σ^2	47.00 (1.39)	41.42	39.92 (1.22)
LL	-7647.4	-7540.7	-7551.6
BIC	15333.9	15149.7	15146.2
Npar	8	14	9

IV. Repeated Measures: LC Growth Modeling

Data on 237 children (13 years of age) from 5 waves (1976-1980) of the National Youth Survey (see, e.g., Vermunt & Hagenaars 2004)

Ordinal dependent variable "Marijuana use in the past year" (0 = never, 1 = no more than once a month, 2 = more than once a month).

Predictors: time (z_{ij1}) and sex (z_{i2}); Covariate: sex (z_{i2});

We use a random-effect or growth modeling approach. Alternatives are transitional (see below) and marginal models.

Questions:
Can we identify latent classes with different growth patterns?
Does it make sense to include a random intercept in the model?

Most general form of the relevant LC adjacent-category ordinal logit model:

$$\log \frac{P(Y_{ij} = c+1 \mid X_i = t, \mathbf{z}_{ij})}{P(Y_{ij} = c \mid X_i = t, \mathbf{z}_{ij})} = (\alpha_{(c+1)t} - \alpha_{ct}) + \beta_{1t} z_{ij1} + \beta_2 z_{i2} + u_i$$

Three special cases:

A. only intercept is class specific, no random intercept

$\beta_{1t} = \beta_1; \gamma = 0$

B. class-specific intercept and slope, no random intercept

$\gamma = 0$

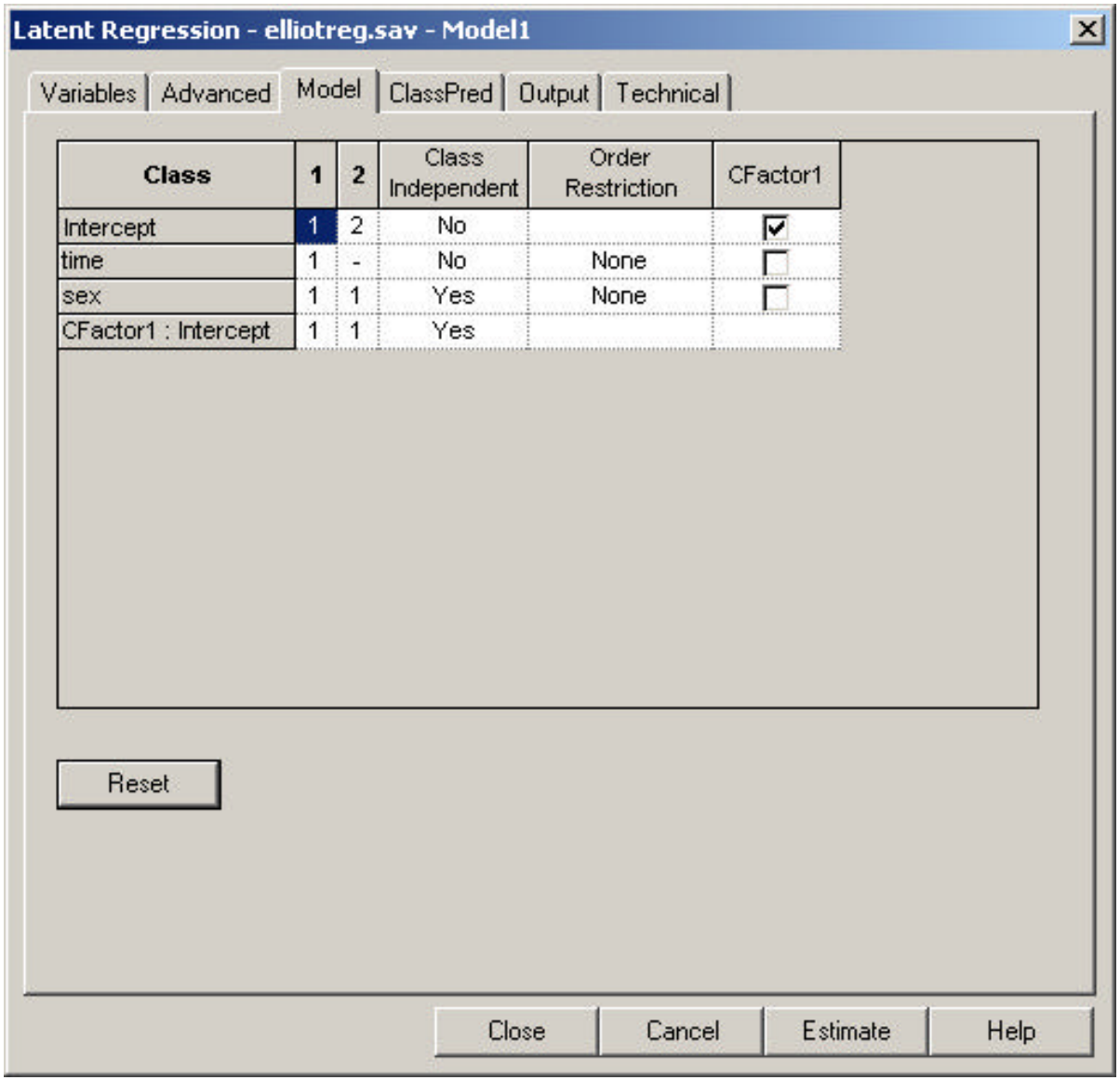
C. class-specific intercept and slope and random intercept

Fit measures for the estimated models

	A. Class-specific Intercept			B. A + Class-specific Slope			C. B + Random Intercept		
	BIC	L^2	df	BIC	L^2	df	BIC	L^2	df
1-Class	1697.7	536.7	480	1697.7	536.7	480	1402.0	235.5	479
2-Class	1438.0	260.5	477	1442.8	259.9	476	**1395.2**	206.8	475
3-Class	**1408.7**	214.9	474	**1417.9**	213.0	472	1408.9	198.6	471
4-Class	1420.9	210.6	471	1426.0	199.3	468	1427.8	195.6	467
5-Class	1436.1	209.4	468	1436.4	187.9	464	1441.5	187.5	463

Further modifications in 2-class model C:

- sex as covariate affecting classes: BIC=**1398.5**; L^2=204.6; df=474
- no time effect in class 2: BIC=**1393.3**; L^2=210.3;df=476
- no time effect in class 2 and "nominal" slope: BIC=**1390.0**; L^2=190.6; df=473; $p_{bootstrap}$=.09 ((final model)


Latent Regression - elliotreg.sav - Model1
Variables
Advanced
Model
ClassPred
Output
Technical
Class
1
2
Class Independent
Order Restriction
CFactor1
Intercept
1
2
No
time
1
-
No
None
sex
1
1
Yes
None
CFactor1 : Intercept
1
1
Yes
Reset
Close
Cancel
Estimate
Help

Development of class-specific means (final model)

	Male		Female	
Time	Class1	Class2	Class1	Class2
1976	0.16	0.09	0.06	0.02
1977	0.50	0.09	0.23	0.02
1978	0.86	0.09	0.47	0.02
1979	1.03	0.09	0.61	0.02
1980	1.21	0.09	0.77	0.02

Very simple pattern emerges:

- class 2: low use class at first time point that remains low
- class 1: higher use class at first time point that increases consumption a lot
- females use less than males, but show same development pattern

Latent classes combined with a random intercept may yield a much simpler solution

V. Event History Analysis: Cox Regression Model for Correlated Events

Survey among 145 young adults on the effect of "youth centrism" on the timing of first experiences with relationships (see Vermunt, 2002)

We have information on the age of first time "sleeping with someone", "having a steady friend", "being very much in love", and "going out"

A Cox model is used to cluster respondents with similar rates of experiencing these four correlated events

A Cox model is equivalent to a Poisson regression model in the form of a piecewise exponential survival model. This requires that the data is in the form of episode records whose end points correspond with the times at which events occur (see Vermunt, 1997)

The LC Cox regression model for multiple events:

$$h_{ij\ell} = h_{j\ell} \exp(\sum_{k=1}^{K} \beta_{kt} z_{ij\ell k})$$

Rate of occurrence of event j in time interval ℓ is a function of a baseline hazard rate and time- and event-specific covariates.

Formulated as a LC Poisson regression model for the expected number of events:

$$E(Y_{ij\ell} \mid X_i = t, \mathbf{z}_{ij\ell}) = R_{ij\ell} \exp(\alpha_{j\ell} + \sum_{k=1}^{K} \beta_{kt} z_{ij\ell k})$$

where $R_{ij\ell}$ denotes an exposure time or risk period.

Fit measures for the estimated Cox regression models

Model	LL	BIC(LL)	Npar
1-Class	-1650.0	3563.5	53
2-Class	-1605.1	3498.5	58
3-Class	-1593.8	3500.8	63
4-Class	-1589.9	3517.7	68
1-CFactor random effects	-1609.8	3502.9	57
3-Class + covariates	-1587.4	3517.7	69

Parameters of Model for Dependent

	Class1	Class2	Class3	Wald	p-value
sleeping	-0.39	-2.39	-3.80	152.77	0.00
friend	-0.69	-2.25	-4.01	148.54	0.00
inlove	-1.50	-1.87	-2.88	256.49	0.00
goingout	-1.00	-1.74	-2.05	81.99	0.00

Parameters of Model for Classes

	Class1	Class2	Class3	Wald	p-value
Intercept	0.53	0.26	-0.79	7.62	0.02
youthcen	0.14	0.17	-0.31	0.48	0.78
boy	-0.65	-0.42	1.08	8.94	0.01
loweduc	0.13	0.45	-0.58	1.64	0.44

Classification table (3-class model with covariates)

	Modal			
Probability	Class1	Class2	Class3	Total
Class1	56.4	3.7	0.0	60.1
Class2	8.4	47.0	2.4	57.8
Class3	0.2	4.3	21.6	26.1
Total	65.0	55.0	24.0	144.0

Classification errors = 0.13
Reduction of errors (Lambda) = 0.77

VI. Modeling Choice Data

Application: choice-based conjoint study in the development of a new coffee maker. Experiment in which 185 persons select a product out of each of 8 sets of 3 alternatives (see Skrondal and Rabe-Hesketh, 2004).

We use a LC conditional logit (LC discrete choice) model. The effects of alternative-specific predictors (called attributes) may vary across latent classes.

Product attributes:

- brand (Philips, Braun, Moulinex)
- capacity (6 cups, 10 cups, 15 cups)
- price (F39, F69, F99)
- filter (yes, no)
- thermos flask (yes, no)

Purpose: segmentation (and simulation) based on utilities assigned to attributes

Standard LC multinomial logit model (parameter values depend on *c*)

$$P(Y_{ij} = c \mid X_i = t, \mathbf{z}_{ij}) = \frac{\exp(\alpha_{ct} + \sum_{k=1}^{K} \beta_{ckt} z_{ijk})}{\sum_{c'=1}^{C} \exp(\alpha_{c't} + \sum_{k=1}^{K} \beta_{c'kt} z_{ijk})}$$

LC conditional logit model (attribute values depend on *c*)

$$P(Y_{ij} = c \mid X_i = t, \mathbf{z}_{ij}) = \frac{\exp(\sum_{k=1}^{K} \beta_{kt} z_{ijck})}{\sum_{c'=1}^{C} \exp(\sum_{k=1}^{K} \beta_{kt} z_{ijc'k})}$$

The latter requires a special data organization to be able to indicate the attribute values of all alternatives in a choice set

- one-file format (case-set-alternative records)
- three-file format (responses, sets, and alternatives files)

Alternatives file

AltID	Brand	Capacity	Price	Filter	Thermos
1	philips	10 cups	F69	yes	no
2	braun	6 cups	F69	no	no
3	braun	10 cups	F39	no	no
4	philips	6 cups	F39	yes	yes
5	philips	10 cups	F69	no	yes
6	braun	6 cups	F69	yes	no
7	philips	15 cups	F99	no	no
8	braun	15 cups	F69	no	yes
…					
…					
47	philips	6 cups	F69	yes	no
48	philips	6 cups	F69	yes	no

Sets file (16 different sets)

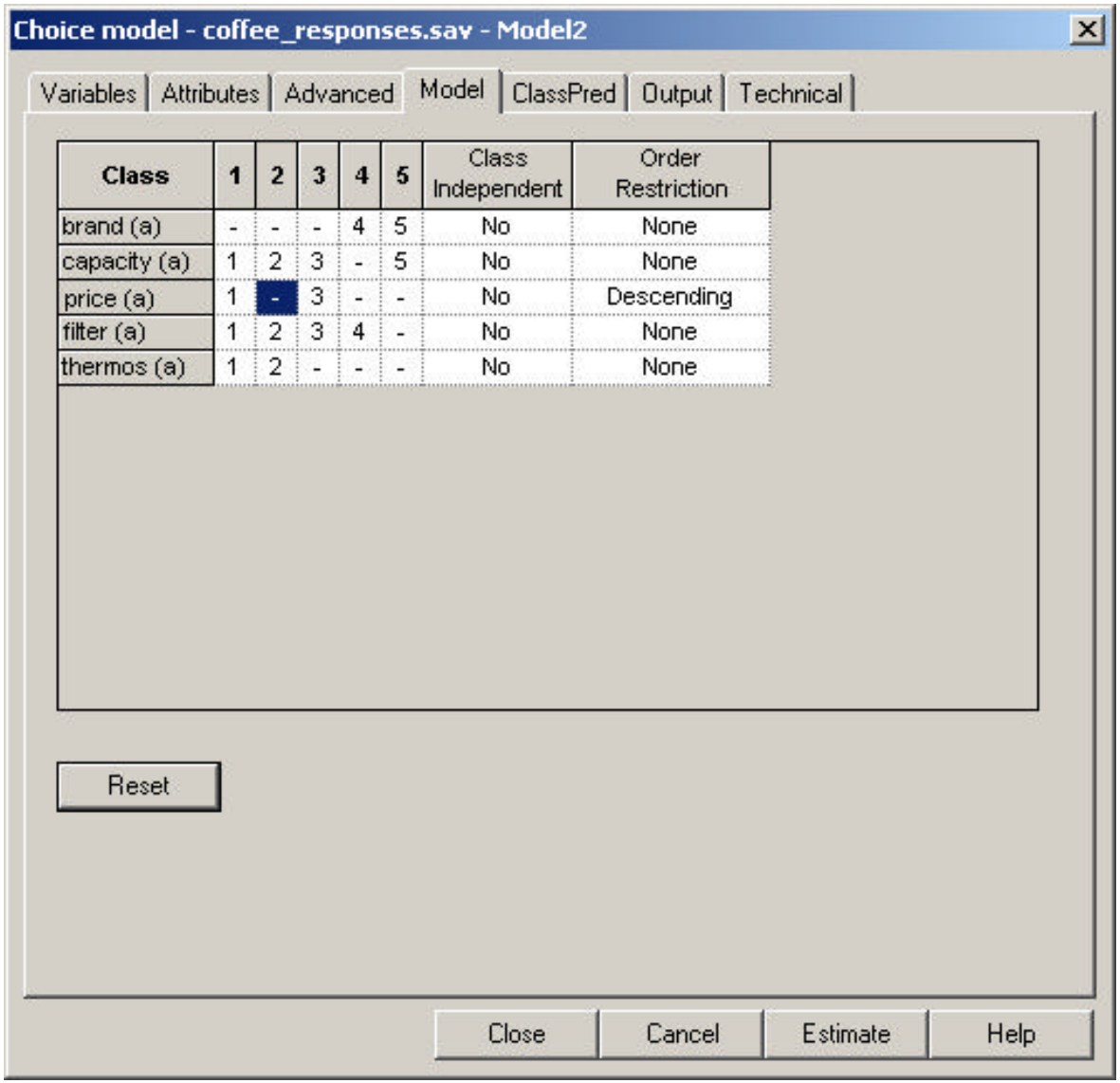
SetID	Alt1	Alt2	Alt3
1	1	17	33
2	2	18	34
3	3	19	35
4	4	20	36
5	5	21	37
6	6	22	38
7	7	23	39
8	8	24	40
…			
…			
15	15	31	47
16	16	32	48

Responses file (185*8=1480 records)

CaseID	SetNR	Choice
1	1	3
1	2	3
1	3	3
…		
…		
26	7	3
26	8	3
27	1	2
27	2	3
27	3	1
…		
…		
185	14	2
185	15	1
185	16	3

Fit measures of the estimated choice models

Model	LL	BIC(LL)	Npar	Class.Err.	$R^2(0)$	R^2
1-class	-1298.7	2639.2	8	0.00	0.23	0.16
2-Class	-1110.6	2310.0	17	0.03	0.41	0.36
3-Class	-1079.7	2295.2	26	0.07	0.48	0.43
4-Class	-1050.4	2283.5	35	0.09	0.51	0.47
5-Class	-1013.6	2257.0	44	0.06	0.56	0.52
6-Class	-1004.9	2286.6	53	0.08	0.56	0.53
5-Class restricted	-1028.1	2192.0	26	0.08	0.55	0.51


Choice model - coffee_responses.sav - Model2
Variables
Attributes
Advanced
Model
ClassPred
Output
Technical
Class
1
2
3
4
5
Class Independent
Order Restriction
brand (a) - - - 4 5 No None
capacity (a) 1 2 3 - 5 No None
price (a) 1 - 3 - - No Descending
filter (a) 1 2 3 4 - No None
thermos (a) 1 2 - - - No None
Reset
Close
Cancel
Estimate
Help

Parameter estimates for 5-class restricted model

Attribute	Value	Class1	Class2	Class3	Class4	Class5
Brand	philips				1.12	1.89
	braun				-0.97	-6.68
	moulinex				-0.14	4.79
Capacity	6 cups	-3.06	-0.65	-3.33		-1.20
	10 cups	1.36	0.67	0.62		1.56
	15 cups	1.70	-0.02	2.70		-0.35
Price	F39	1.16		2.35		
	F69	0.95		0.75		
	F99	-2.10		-3.10		
Filter	yes	0.69	0.27	2.78	0.43	
	no	-0.69	-0.27	-2.78	-0.43	
Fhermos	yes	0.38	0.61			
	No	-0.38	-0.61			

Variants & extensions:

Choice format
- full / partial rankings
- best-worst
- pick K out of N

Continuous random effects

Predictors (person or set characteristics) in addition to attributes

Covariates (person characteristics)

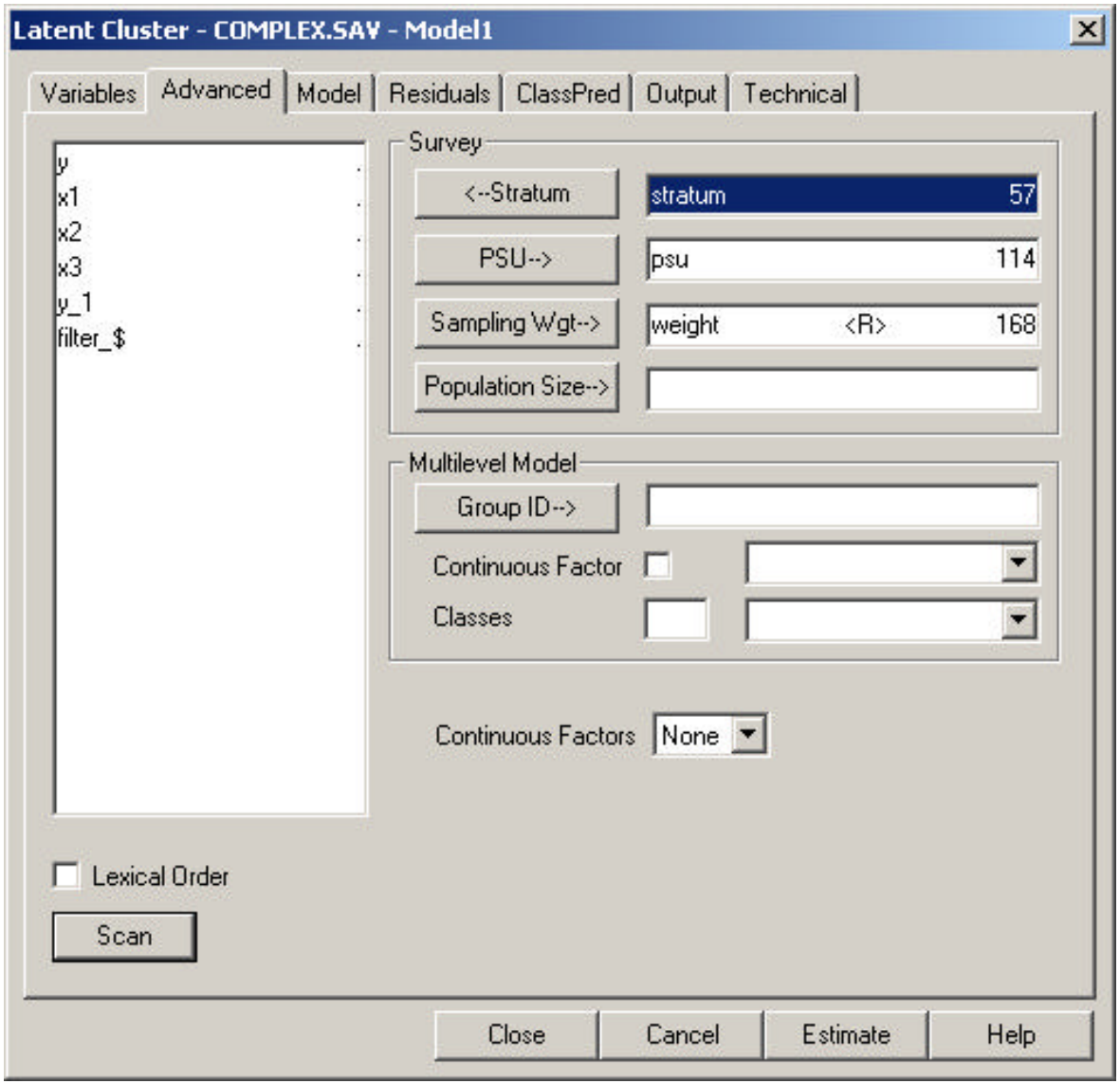
Nonparametric random-effects log-linear model for dependent observations
- application: personal networks

Further topics

Censored and truncated dependent variables
- tobit: nonnegative continuous with lots of zeroes
- truncated Poisson: only cases with at least one event are in the sample
    - criminal records, capture-recapture data, purchases of clients

Complex sampling: pseudo ML with corrected SE's and Wald tests
- stratification, clustering (psu), weighting, finite population corrections

Multilevel extension (see Vermunt, 2003/2004)
- 2-level LC model: latent class distribution varies across groups
- 3-level LC regression model: regression coefficients vary across level-3 units
- random effects:
    - parametric: continuous normal
    - nonparametric: LCs of higher-level units


Latent Cluster - COMPLEX.SAV - Model1
Variables
Advanced
Model
Residuals
ClassPred
Output
Technical
y
x1
x2
x3
y_1
filter_$
Survey
<--Stratum
stratum 57
PSU-->
psu 114
Sampling Wgt-->
weight <R> 168
Population Size-->
Multilevel Model
Group ID-->
Continuous Factor
Classes
Continuous Factors
None
Lexical Order
Scan
Close
Cancel
Estimate
Help

References

Agresti, A. (2002). *Categorical Data Analysis.* New York: Wiley.

Aitkin (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55, 218-234.

Heinen, T. (1996). *Latent Class and Discrete Latent Trait Models: Similarities and differences*. Thousand Oakes: Sage Publications.

Skrondal A., and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. London: Chapman & Hall/ CRC.

Snijders, T., and Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Vermunt, J.K. (1997). *Log-linear Models for Event Histories*. Thousand Oakes: Sage Publications.

Vermunt, J.K. (2002) A general non-parametric approach to unobserved heterogeneity in the analysis of event history data. J. Hagenaars and A. McCutcheon (eds.): *Applied Latent Class Models*, 383-407 Cambridge University Press.

Vermunt, J.K.(2003) Multilevel latent class models. *Sociological Methodology*, 33, 213-239.

Vermunt, J.K. (2004) An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica*, 58, 220- 233.

Vermunt, J.K. & Hagenaars, J.A. (2004). Ordinal longitudinal data analysis. In: R. Hauspie, N. Cameron and L. Molinari (eds.). *Methods in Human Growth Research*, Chapter 15. Cambridge University Press.

Vermunt, J.K. and Van Dijk. L. (2001). A nonparametric random-coefficients approach: the latent class regression model. *Multilevel Modelling Newsletter*, 13, 6-13.

Wedel, M., and DeSarbo, W.S (1994). A review of recent developments in latent class regression models. R.P. Bagozzi (ed.), *Advanced Methods of Marketing Research*, 352-388, Cambridge: Blackwell Publishers.