Jacques A. Hagenaars

Department of Methodology/WORC

Faculty of Social Sciences
P.O. Box 90153
5000 LE Tilburg,
The Netherlands

tel: + 31 13 4662086
fax: + 31 13 4663002

email: jacques.a.hagenaars@kub.nl

Those who can, DO

Those who cannot, TEACH

Those who have nothing to teach,

become METHODOLOGISTS

Paul Lazarsfeld

INTRODUCTION

- What are categorical data

- **'natural' categories**
- **'forced' categories**

- Why loglinear analysis?

Standard(regression-)model:
linearity; no interaction; normality; interval level; 'impossible' est. values; no categorical dependent variable

- Pearson: Lisrel/EQS; optimal scaling

vs

Yule: d %, information approaches, loglinear modeling

CONTENTS

Required:

Hagenaars *Categorical Longitudinal Data*, Sage. 1990, ch. 2,3

Hagenaars *Loglinear Models with Latent Variables*, Sage. 1993

Introductory:

Knoke/Burke *Loglinear Modeling*, Sage

McCutcheon *Latent Class Analysis*, Sage

More comprehensive:

Agresti *Categorical Data Analysis*, Wiley. 1990

Program/Theory: Vermunt: LEM

LOGLINEAR MODELS

BASIC CONCEPTS/NOTATION

- **MLE (WLS/GSK) , (product) multinomial sampling**

- $\boldsymbol{p}, \pi, \hat{\pi}$

  $\boldsymbol{f}, \boldsymbol{F}, \hat{\boldsymbol{F}}$ $(= \boldsymbol{Np}, N\pi, N\hat{\pi})$

- f_{ijk}^{ABC}; $\quad f_{i+k}^{ABC} = \sum_{j} f_{ijk}^{ABC}$ **(marginal tables)**

- **odds : marginal odds**
  - **: conditional odds**
  - **: partial odds ($\bar{X}_{geom}$)**
  - **: odds ratio**
  - **: higher order odds ratio**
  - **: 'symmetric' and**
  - **: 'asymmetric'**
  - **: ‘invariant' odds ratios**
  - **: logits**
  - **: polytomies**

	Vote			
	yes	no	Tot.	*%yes*
young	195	141	336	*58%*
old	311	103	414	*75%*
	506	244	750	*67%*

marginal odds:

young/old: $\frac{336}{750} / \frac{414}{750} = .448/.552 = 336/414 = .812$

yes/no : 506/244 = 2.074

no/yes : 244/506 = 1/2.074 = .482

conditional odds:

among young: yes/no: 195/141 = 1.383

among old : yes/no : 311/103 = 3.019

partial odds: yes/no : $\sqrt[2]{(195/141)(311/103)} = 2.043$
(geometric mean)

odds ratio : $\frac{195}{141} / \frac{311}{103} = .458 = \alpha$

$ln 2.074 = .729 \leftrightarrow ln .482 = -.729$

- anchor points:

odds :	**0**	**1**	**+∞**
log odds/logit :	**-∞**	**0**	**+∞**

$$- \alpha = \frac{195}{141} / \frac{311}{103} = \frac{195}{311} / \frac{141}{103} = \frac{(195)(103)}{(141)(311)} = 0.458$$

- 'invariance' property

- extensions:

- **polytomous variables (ordered/unordered)**
- **> 2 variables: higher order odds ratios**

- parameters of loglinear models are odds (ratios)

- Think multiplicatively!!!

SATURATED LOGLINEAR MODEL

- $F_{ijk}^{ABC} = \eta \; \tau_i^A \tau_j^B \tau_k^C \; \tau_{ij}^{AB} \tau_{ik}^{AC} \tau_{jk}^{BC} \; \tau_{ijk}^{ABC}$

$$\ln F_{ijk}^{ABC} = G_{ijk}^{ABC} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

- **identifiability:**

dummy coding:

$$\tau_1^A = \tau_{1j}^{AB} = \tau_{i1}^{AB} = 1$$
$$\lambda_1^A = \lambda_{1j}^{AB} = \lambda_{i1}^{AB} = 0$$

effect coding:

$$\prod_i \tau_i^A = \prod_i \tau_{ij}^{AB} = \prod_j \tau_{ij}^{AB} = 1$$
$$\sum_i \lambda_i^A = \sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = 0$$

- $\hat{F} = f \qquad \hat{\pi} = p$

		B	
		1.yes	2.no
A	1.young	195	141
	2.old	311	103

$$F_{ij}^{AB} = \eta \ \tau_i^A \ \tau_j^B \ \tau_{ij}^{AB}$$

Effect coding:

AB

1 1 $\quad 195 = \eta \tau_1^A \tau_1^B \tau_{11}^{AB}$

1 2 $\quad 141 = \eta \tau_1^A (1/\tau_1^B)(1/\tau_{11}^{AB})$

2 1 $\quad 311 = \eta (1/\tau_1^A) \tau_1^B (1/\tau_{11}^{AB})$

2 2 $\quad 103 = \eta (1/\tau_1^A)(1/\tau_1^B) \tau_{11}^{AB}$

$$\eta = \sqrt[4]{(195)(141)(311)(103)} = 172.271$$

$$\hat{\tau}_1^A = \sqrt[4]{\frac{195 * 141}{311 * 103}} = .963 \quad \hat{\tau}_1^A = \frac{\sqrt[2]{195 * 141}}{\eta}$$

$$\hat{\tau}_1^B = \sqrt[4]{\frac{195 * 311}{141 * 103}} = 1.430 \quad \hat{\tau}_1^B = \frac{\sqrt[2]{195 * 311}}{\eta}$$

$$\hat{\tau}_{11}^{AB} = \sqrt[4]{\frac{195 * 103}{141 * 311}} = .823 \quad \hat{\tau}_{11}^{AB} = \frac{195}{\eta \hat{\tau}_1^A \hat{\tau}_1^B}$$

$\hat{\lambda}_{11}^{AB} = -.195$ $\qquad \alpha = .458$

		B	
		1.yes	**2.no**
A	**1.young**	**195**	**141**
	2.old	**311**	**103**

Dummy coding:

A B

1 1 $195 = \eta \ \tau_1^A \ \tau_1^B \ \tau_{11}^{AB}$

1 2 $141 = \eta \ \tau_1^A \ (1) \ (1)$

2 1 $311 = \eta \ (1) \ \tau_1^B \ (1)$

2 2 $103 = \eta \ (1) \ (1) \ (1)$

$$\eta = 103$$

$$\hat{\tau}_1^A = 141/103 = 1.369 \qquad \hat{\tau}_1^A = \frac{141}{\eta}$$

$$\hat{\tau}_1^B = 311/103 = 3.019 \qquad \hat{\tau}_1^B = \frac{311}{\eta}$$

$$\hat{\tau}_{11}^{AB} = \frac{195}{141} / \frac{311}{103} = \frac{195}{311} / \frac{141}{103} = \frac{195 * 103}{311 * 141} = .458$$

$$\hat{\tau}_{11}^{AB} = \frac{195}{\eta \hat{\tau}_1^A \hat{\tau}_1^B}$$

UNSATURATED MODELS

- hierarchical vs. nonhierarchical models

- in hierarchical models:

sufficient statistics are the observed marginals to be reproduced;

for model { AB, C}: sufficient statistics:

A
| **C**
B

$$\hat{F}^{ABC}_{ij+} = f^{ABC}_{ij+}$$

$$\hat{F}^{ABC}_{++k} = f^{ABC}_{++k}$$

- closed form expressions *vs.* iterative procedures (Iterative Proportional Fitting, Newton/Raphson)

- test statistics:

$$X^2 = \sum_i \sum_j [(f^{AB}_{ij} - \hat{F}^{AB}_{ij})^2 / \hat{F}^{AB}_{ij}]$$

$$L^2 \ (G^2) = 2 \sum_i \sum_j [f^{AB}_{ij} \ln(f^{AB}_{ij} / (\hat{F}^{AB}_{ij})]$$

df (degrees of freedom): number of independent knowns minus number of parameters to be (independently) estimated

- traditional chi-square statistics:

$$L^2\ (G^2) \qquad X^2$$

- dissimilarity index D: $D = \frac{\sum |f - \hat{F}|}{2N}$

- descriptive fit statistics:

$$e = \frac{L^2}{N}\ \left(e = \frac{X^2}{N}\right) \qquad w = \sqrt{e} \qquad F = \frac{L^2}{df}$$

$$R^{/} = \frac{L_r^2 - L_u^2}{L_r^2} = \frac{e_r - e_u}{e_r} \qquad \hat{\delta} = \frac{F_r - F_u}{F_r}$$

- information criteria

AIC = L² - 2(df) **BIC = L² - (lnN)(df)**

IPF:

Model {AB, AC. BC}

AB: $$\hat{F}_{ijk}^{ABC}(1) = \frac{\hat{F}_{ijk}^{ABC}(0)}{\hat{F}_{ij+}^{ABC}(0)} f_{ij+}^{ABC}$$

$$(\hat{F}_{ijk}^{ABC}(1) = \hat{\pi}_{k\,ij}^{C|AB}(0) f_{ij+}^{ABC})$$

AC: $$\hat{F}_{ijk}^{ABC}(2) = \frac{\hat{F}_{ijk}^{ABC}(1)}{\hat{F}_{i+k}^{ABC}(1)} f_{i+k}^{ABC}$$

$$\hat{F}_{ijk}^{ABC}(2) = \hat{\pi}_{j\,ik}^{B|AC}(1) f_{i+k}^{ABC}$$

BC: $$\hat{F}_{ijk}^{ABC}(3) = \frac{\hat{F}_{ijk}^{ABC}(2)}{\hat{F}_{+jk}^{ABC}(2)} f_{+jk}^{ABC}$$

$$\hat{F}_{ijk}^{ABC}(3) = \hat{\pi}_{i\,jk}^{A|BC}(2) f_{+jk}^{ABC}$$

- Newton/Raphson; Design Matrix

- testing and estimation when empty cells

Formulas + meaning:

$$\eta = \sqrt[IJK]{\prod_i^I \prod_j^J \prod_k^K F_{ijk}} \qquad \tau_i^A = \frac{\sqrt[JK]{\prod_j^J \prod_k^K F_{ijk}}}{\eta}$$

$$\tau_{ij}^{AB} = \frac{\sqrt[K]{\prod_k^K F_{ijk}}}{\eta \tau_i^A \tau_j^B} \qquad \tau_{ijk}^{ABC} = \frac{F_{ijk}}{\eta \tau_i^A \tau_j^B \tau_k^C \tau_{ij}^{AB} \tau_{ik}^{AC} \tau_{jk}^{BC}}$$

$$\theta = \frac{1}{IJK} \sum_i^I \sum_j^J \sum_k^K G_{ijk} \qquad \lambda_i^A = \frac{1}{JK} \sum_j^J \sum_k^K G_{ijk} - \theta$$

$$\lambda_{ij}^{AB} = \frac{1}{K} \sum_k^K G_{ijk} - \theta - \lambda_i^A - \lambda_j^B$$

$$\lambda_{ijk}^{ABC} = G_{ijk} - \theta - \lambda_i^A - \lambda_j^B - \lambda_k^C - \lambda_{ij}^{AB} - \lambda_{ik}^{AC} - \lambda_{jk}^{BC}$$

- ‘special’ case: dichotomies

- interaction: different conditional effects: interaction A-B-C (with 3-cat. C):

$$\tau^{AB}_{ij} = \sqrt[3]{\tau^{AB|C}_{ij\ 1}\,\tau^{AB|C}_{ij\ 2}\,\tau^{AB|C}_{ij\ 3}}$$

$$\lambda^{AB}_{ij} = \frac{1}{3}(\lambda^{AB|C}_{ij\ 1} + {}^{AB|C}_{ij\ 2} + {}^{AB|C}_{ij\ 3})$$

$$\tau^{ABC}_{ijk} = \tau^{AB|C}_{ij\ k} / \tau^{AB}_{ij}$$

$$\lambda^{ABC}_{ijk} = \lambda^{AB/C}_{ij\ k} - \lambda^{AB}_{ij}$$

$$\tau^{AB|C}_{ij\ k} = (\tau^{AB}_{ij})(\tau^{ABC}_{ijk})$$

$$\lambda^{AB|C}_{ij\ k} = \lambda^{AB}_{ij} + \lambda^{ABC}_{ijk}$$

- s.e. (2x2x2) =

$$\hat{s}^2_\lambda = \frac{1}{8}\sum_i \sum_j \sum_k \frac{1}{f_{ijk}}$$

$$H_0\!:\ \lambda = 0 \qquad z = \frac{\hat{\lambda}}{\hat{s}_\lambda}$$

CONDITIONAL TESTING

model 1: {A, B, C}; model 2: {AB, C};

model 3: {AB, AC, BC}

$$L^2_{r/u} = L^2_r - L^2_u \qquad df_{r/u} = df_r - df_u$$

$$L^2_1 = (L^2_1 - L^2_2) + (L^2_2 - L^2_3) + L^2_3$$

$$= L^2_{1/2} + L^2_{2/3} + L^2_3$$

- forward selection:

residuals $(f - \hat{F})$ (standardized or adjusted)

- backward selection:

$z = \hat{\lambda} / \hat{s}_\lambda$ **(en Wald statistic)**

- large samples / overfitting
small samples / underfitting

EXPLORATORY MODEL SELECTION

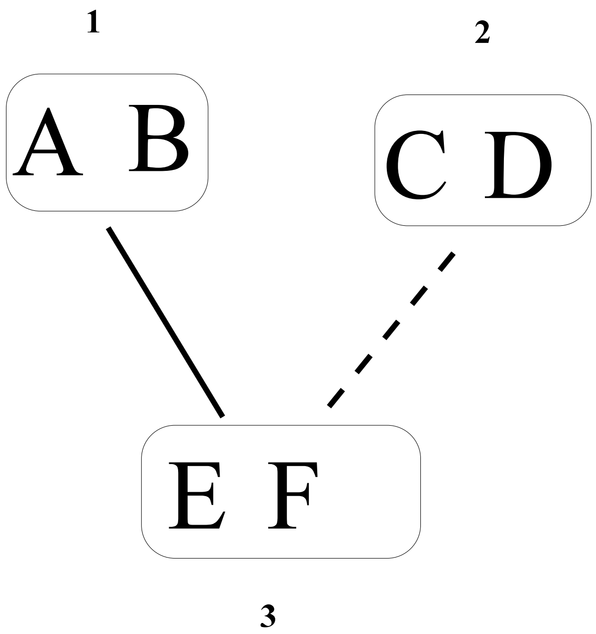
principles

- **use theory**
- **large samples/power considerations**
- **try all meaningful directions**
- **suspend judgment**
- **inspect parameter estimates**
- **random division of sample (cross-validation)**

procedure

- **select base-line model**
- **add effects one at a time**
- **add extra effects**
- **delete effects one at a time**
- **accept model if adding effects does not result in significant extra effects and does not change the existing effects**

COLLAPSIBILITY



group 1: variables to be collapsed (A,B)
group 2: variables independent of the first group (C,D)
group 3: the remaining variables (E, F)

the values of all parameters which have superscripts containing one or more variables of the second group will not change when the table is collapsed over one or more variables of the first group

e.g., models {ABEF,CDEF} or
{AB,AE,BF,EF,CD}

and parameters: $\lambda_{kl}^{CD}, \lambda_{lm}^{DE}, \lambda_{kmn}^{CEF}$ *vs.* λ_{mn}^{EF}

LOGIT MODELS

Notation/Basic Model

$$\pi_{k\,ij}^{C|AB},\ \Omega_{k/k'\,ij}^{C\ |AB} \qquad (\pi_{ijk}^{AB\bar{C}},\ \Omega_{ij\,k/k'}^{AB\bar{C}})$$

$$\Omega_{k/k'\,ij}^{C\ |AB} = \frac{\pi_{k\,ij}^{C|AB}}{\pi_{k'\,ij}^{C\ |AB}} = \frac{\pi_{ijk}^{ABC}/\pi_{ij}^{AB}}{\pi_{ijk'}^{ABC}/\pi_{ij}^{AB}} = \frac{\pi_{ijk}^{ABC}}{\pi_{ijk'}^{ABC}} = \frac{F_{ijk}^{ABC}}{F_{ijk'}^{ABC}}$$

multpl.: $\Omega,\ \hat{\Omega},\ \omega$ *logl.* $\Phi\ \hat{\Phi}\ \varphi$

(cf $\pi,\ \hat{\pi},\ p;\ F,\ \hat{F},\ f$)

$$\Omega_{k/k'\,ij}^{C\ |A\,B} = \gamma_{k/k'}^{C}\,\gamma_{k/k'\,i}^{C\ |A}\,\gamma_{k/k'\,j}^{C\ |B}\,\gamma_{k/k'\,ij}^{C\ |AB}$$

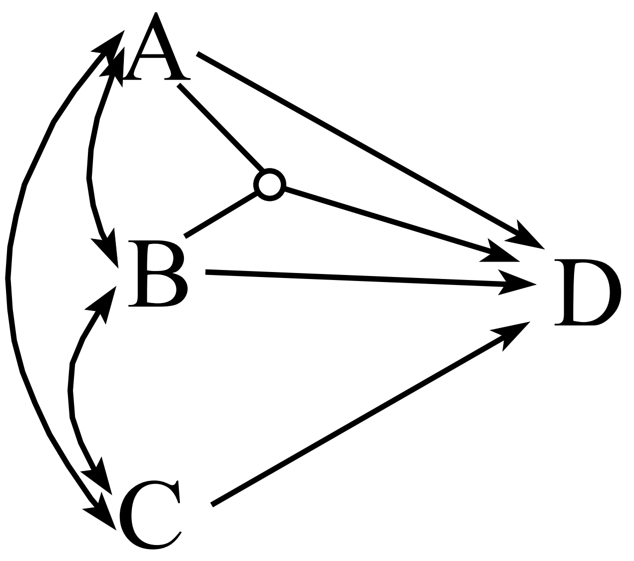
$$= \frac{F_{ijk}^{ABC}}{F_{ijk'}^{ABC}} = \frac{\eta\tau_i^A\tau_j^B\tau_{ij}^{AB}\ \ \tau_k^C\tau_{ik}^{AC}\tau_{jk}^{BC}\tau_{ijk}^{ABC}}{\eta\tau_i^A\tau_j^B\tau_{ij}^{AB}\ \ \tau_{k'}^C\tau_{ik'}^{AC}\tau_{jk'}^{BC}\tau_{ijk'}^{ABC}}$$

$$= \frac{\tau_k^C}{\tau_{k'}^C}.\frac{\tau_{ik}^{AC}}{\tau_{ik'}^{AC}}.\frac{\tau_{jk}^{BC}}{\tau_{jk'}^{BC}}.\frac{\tau_{ijk}^{ABC}}{\tau_{ijk'}^{ABC}}$$

$$\Phi_{k/k'\,ij}^{C\ |AB} = \beta_{k/k'}^{C} + \beta_{k/k'\,i}^{C\ |A} + \beta_{k/k'\,j}^{C\ |B} + \beta_{k/k'\,ij}^{C\ |AB}$$

$$\beta_{k/k'}^{C} = (\lambda_k^C - \lambda_{k'}^C),\ \ \beta_{k/k'\,i}^{C\ |A} = (\lambda_{ik}^{AC} - \lambda_{ik'}^{AC}), etc.$$

Logit model as a 'modified multiple regression model'



A logit model is identical to an ordinary loglinear model that includes all loglinear parameters corresponding to the logit-effect parameters plus all parameters needed for reproducing the joint distribution of the independent variables.

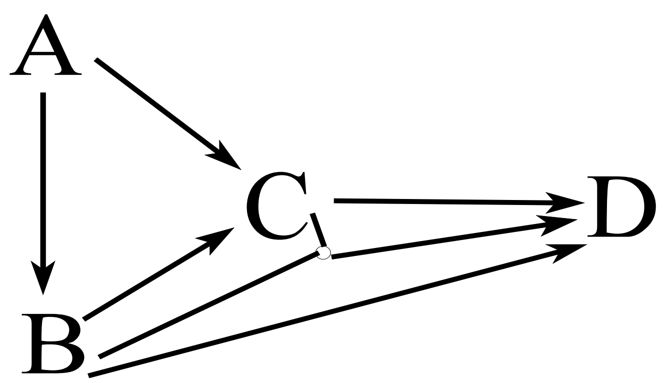
In Figure: Model {ABC,ABD,CD}

MODIFIED PATH MODEL

- **Causal Modeling: Structural Equation Model**

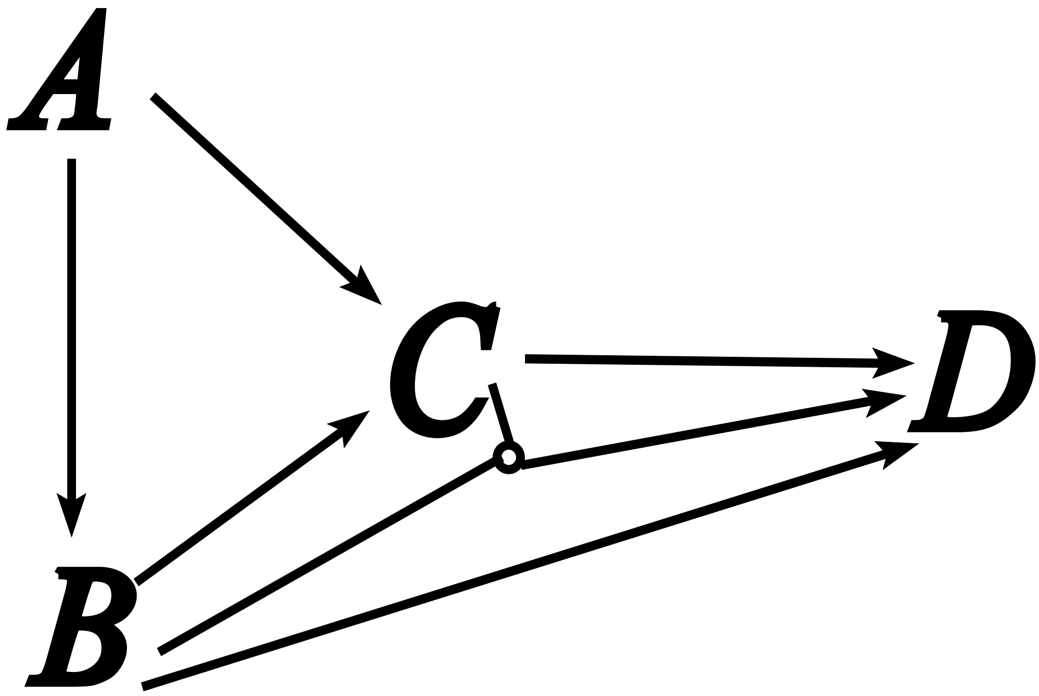
- **System of (Logit) Equations; Directed Graph; Directed Loglinear Model (DLM)**

- **Recursive System; Acyclical Directed Graph**



A-Age; B-Education; C-Occupation; D-Income

$$\pi_{ijkl}^{ABCD} = \pi_{i}^{A}\,\pi_{ji}^{B|A}\,\pi_{kij}^{C|AB}\,\pi_{lijk}^{D|ABC}$$


A
B
C
D

$$\pi_{k\,ij}^{C|AB} = \frac{\pi_{ijk}^{ABC}}{\pi_{ij+}^{ABC}}$$

$$= \frac{\eta\tau_i^A\tau_j^B\tau_{ij}^{AB}\ \tau_k^C\tau_{ik}^{AC}\tau_{jk}^{BC}\tau_{ijk}^{ABC}}{\sum_k \eta\tau_i^A\tau_j^B\tau_{ij}^{AB}\ \tau_k^C\tau_{ik}^{AC}\tau_{jk}^{BC}\tau_{ijk}^{ABC}}$$

$$= \frac{\eta\tau_i^A\tau_j^B\tau_{ij}^{AB}\ \tau_k^C\tau_{ik}^{AC}\tau_{jk}^{BC}\tau_{ijk}^{ABC}}{\eta\tau_i^A\tau_j^B\tau_{ij}^{AB}\ \sum_k \tau_k^C\tau_{ik}^{AC}\tau_{jk}^{BC}\tau_{ijk}^{ABC}}$$

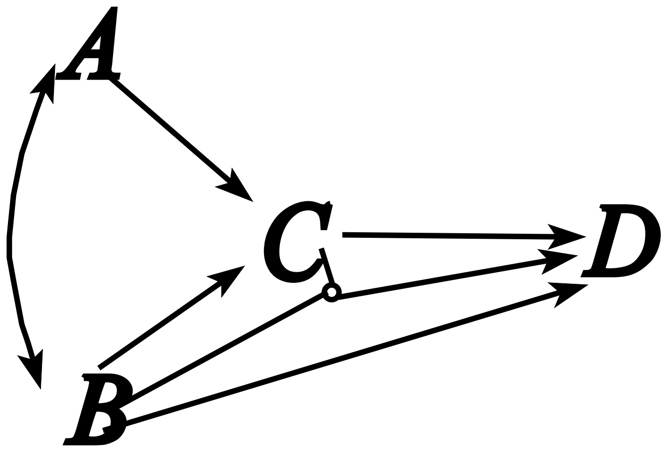
$$= \frac{\tau_k^C\tau_{ik}^{AC}\tau_{jk}^{BC}\tau_{ijk}^{ABC}}{\sum_k \tau_k^C\tau_{ik}^{AC}\tau_{jk}^{BC}\tau_{ijk}^{ABC}}$$

$$\pi_{ijkl}^{ABCD} = \pi_i^A\,\pi_{j\,i}^{B|A}\,\pi_{k\,ij}^{C|AB}\,\pi_{l\,ijk}^{D|ABC}$$

Probabilities	Loglinear Model	Table
1. π_i^A	**{A}**	**A**
2. $\pi_{j\,i}^{B\|A}$	**{AB}**	**AB**
3. $\pi_{k\,ij}^{C\|AB}$	**{AB, AC, BC}**	**ABC**
4. $\pi_{l\,ijk}^{D\|ABC}$	**{ABC, BCD}**	**ABCD**

$$L^* = L_1^2 + L_2^2 + L_3^2 + L_4^2 \qquad df^* = df_1 + df_2 + df_3 + df_4$$

also L^* *by comparing* $\hat{F}^*(=N\hat{\pi}_{ijkl}^{ABCD})$ *with* f_{ijkl}^{ABCD}



$$\pi_{ijkl}^{ABCD} = \pi_{ij}^{AB} \pi_{k\ ij}^{C|AB} \pi_{l\ ijk}^{D|ABC}$$

A, B : exogenous variables; C, D: endogenous

- Simplification of the Graph because of (conditional) independence relations (Markov properties; Graph Theory: ignore variables that have no direct effect):

$$\pi_{ijkl}^{ABCD} = \pi_{ij}^{AB} \pi_{k\ ij}^{C|AB} \pi_{l\ jk}^{D|BC}$$

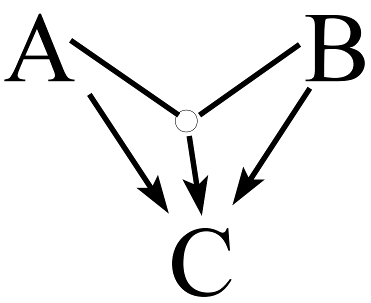
- further 'simplification' using appropriate loglinear models (e.g. 'no three-variable interaction' cannot be expressed in terms of (conditional) independence):

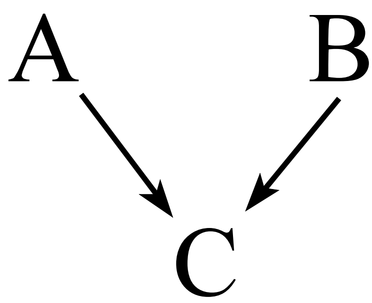
Probabilities	**Loglinear Model**	**Table**
1. π_{ij}^{AB}	**{AB}**	**AB**
2. $\pi_{k\,ij}^{C\|AB}$	**{AB, AC, BC}**	**ABC**
3. $\pi_{l\,jk}^{D\|BC}$	**{BCD}**	**BCD**

L^* as usual through comparing $\hat{F}^*$ and f; but note: after 'graphical simplification' L^* is NOT equal to sum of L^2's (not to $(L_1^2 + L_2^2 + L_3^2)$**)**

If an amoral graph is present, the causal model cannot be tested/estimated by means of one nondirected graph/loglinear model

a) b)





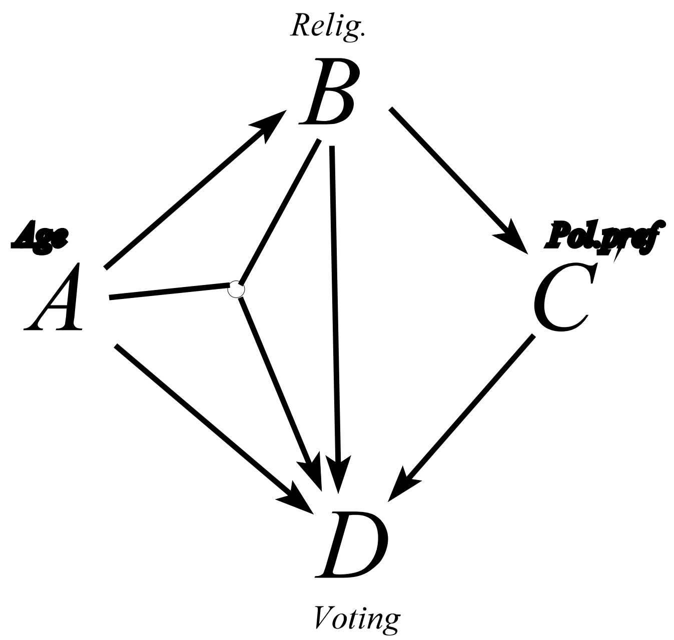
– **Model a) cannot be estimated/tested by means of {ABC} and model b) not by means of {AC,BC}**

– **starting point:** $\pi^{ABC}_{ijk} = \pi^{AB}_{ij} \pi^{C|AB}_{kij}$

– $\hat{\pi}^{AB}_{ij}$ **by means of model {A, B} (for a and b)**

– **and** $\hat{\pi}^{C|AB}_{kij}$ **by means of {ABC} for a) and {AB, AC, BC} for b)**

– ***cf.*** $\pi^{ABC}_{ijk} = \pi^{A}_{i} \pi^{B}_{j} \pi^{C|AB}_{kij}$



$$\pi^{ABCD}_{i\,j\,k\,l} = \pi^{A}_{i}\,\pi^{B|A}_{j\,i}\,\pi^{C|AB}_{k\,ij}\,\pi^{D|ABC}_{l\,i\,j\,k}$$

π^{A}_{i}: model {A} for table A

π^{AB}_{ij}: model {AB} for table AB

π^{ABC}_{ijk}: model {AB, BC} for table ABC

π^{ABCD}_{ijkl}: model {ABC, ABD, CD} for table ABCD

or $\pi^{ABCD}_{ijkl} = \pi^{A}_{i}\,\pi^{B|A}_{j\,i}\,\pi^{C|B}_{kj}\,\pi^{D|ABC}_{l\,ijk}$

and models {A}-table A; {AB}-table AB; {BC}-table BC; model {ABC, ABD, CD}-table ABCD

ORDERED DATA

Nominal level variables

– **no restrictions on relationships**

Ordinal level variables

– **no restrictions on relationships or**

– **(weak) monotonicity; inequality restrictions**

Interval/ratio level variables

– **no restrictions or**

– **(weak)monotonicity; inequality restrictions or**

– **(curvi-)linearity**

Treating Ordinal Data

Variable A ordered in table AB; saturated model:

$$\Omega^{B|A}_{j/j'\ i} = \gamma^{B}_{j/j'}\gamma^{B|A}_{j/j\ i'}$$

- **all variables treated as *nominal* (including adjacent categories odds, continuation-ratio odds, and cumulative odds)**

- **'A' treated as truly *ordinal*: inequality restrictions:**

$$\gamma^{B\ |A}_{j/j'\ 1} \geq \gamma^{B\ |A}_{j/j'\ 2} \geq \gamma^{B\ |A}_{j/j'\ 3} \geq \ldots . \geq \gamma^{B\ |A}_{j/j'\ I}$$

- **'A' *interval*:** $\beta^{B\ |A}_{j/j'\ i} = (\beta_{B_{j/j'}|A}) A_i$

$$\gamma^{B\ |A}_{j/j'\ i} = (\gamma_{B_{j/j'}|A})^{A_i}$$

- **quasi-ordinal/quasi-interval: as 'interval' but estimating A_i (association-2/log-bilinear models)**

Association Models

- **possibilities for (partially) ordered table AB, applying ordinal, association 1, or association 2 models:**

  - **both variables treated as nominal level variables (and hope they come out ordered); cumulative odds etc.**

  - **one variable ordered; ordered row variable: Column Association models; ordered column variable: Row Association models**

  - **both variables ordered: Row+Column Association models; Row*Column Association Models**

- **interval level variable(s); linearly increasing logits:**

$$\lambda_{ij}^{AB} = \lambda_{A_i}(B_j)$$

$$\lambda_{ij}^{AB} = \lambda_{AB}(A_i)(B_j)$$

Design Matrices

Saturated Model; Effect Coding

$$
\begin{array}{ll|ccccccccc|l}
G^{AB}_{ij} & & \theta & \lambda^A_1 & \lambda^A_2 & \lambda^B_1 & \lambda^B_2 & \lambda^{AB}_{11} & \lambda^{AB}_{12} & \lambda^{AB}_{21} & \lambda^{AB}_{22} & \textit{param.} \\
11 & & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & \theta \\
12 & & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & \lambda^A_1 \\
13 & & 1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & \lambda^A_2 \\
21 & & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & \lambda^B_1 \\
22 & = & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & \bullet\ \lambda^B_2 \\
23 & & 1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 & \lambda^{AB}_{11} \\
31 & & 1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 & \lambda^{AB}_{12} \\
32 & & 1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 & \lambda^{AB}_{21} \\
33 & & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & \lambda^{AB}_{22}
\end{array}
$$

$$\underset{\sim}{G} = \underset{\sim}{X} \; \underset{\sim}{\lambda}$$

– **Linear Restrictions**

- **deleting effects**
- **setting effects equal to each other**
- **linear relationships**

effect coding:

$$\lambda_{ij}^{AB} = (\lambda_{A_i})B_j \qquad \tau_{ij}^{AB} = (\tau_{A_i})^{B_j}$$

$$\sum_i \lambda_{ij}^{AB} = \sum_j \lambda_{ij}^{AB} = (\sum_i \lambda_{A_i})B_j = \lambda_{A_i}\sum_j B_j = 0$$

A: nominal, three categories
B: interval

scores

A	D_{A1}	D_{A2}
1	1	0
2	0	1
3	-1	-1

B
-1
0
1

G_{ij}^{AB}	λ_{A_1}	λ_{A_2}
11	-1	0
12	0	0
13	1	0
21	0	-1
22	0	0
23	0	1
31	1	1
32	0	0
33	-1	-1

$-\lambda_1^{*A}$	0	$+\lambda_1^{*A}$
$-\lambda_2^{*A}$	0	$+\lambda_2^{*A}$
$+\lambda_1^{*A}$ $+\lambda_2^{*A}$	0	$-\lambda_2^{*A}$ $-\lambda_1^{*A}$

B. Left P.	***A. Education (low/high)***				
	1	***2***	***3***	***4***	***5***
yes	**69**	**67**	**44**	**26**	**5**
no	**131**	**133**	**106**	**124**	**45**
%yes	**34.5**	**33.5**	**29.3**	**17.3**	**10.0**
$\Omega^{B\mid A}_{1/2\ i}$	**.527**	**.504**	**.415**	**.210**	**.111**
$\Phi^{B\mid A}_{1/2\ i}$	**-.641**	**-.686**	**-.879**	**-1.562**	**-2.197**
$\gamma^{B\mid A}_{1/2\ i}$	**1.736**	**1.661**	**1.369**	**.691**	**.367**
$\beta^{B\mid A}_{1/2\ i}$	**.552**	**.507**	**.314**	**-.369**	**-1.004**

$\gamma^{B}_{1/2}$ **= .306** $\beta^{B}_{1/2}$**= -1.193**

– **Linear Relationship?**

$$\beta^{B\mid A}_{1/2\ i} = (\beta_{B_{1/2}})A_i \qquad (A_i = -2, -1, 0, 1, 2)$$

$$L^2 = 4{,}78 \quad df = 3 \quad p = 0{,}19 \ (X^2 = 4{,}72)$$

$\boldsymbol{\gamma^{B}_{1/2} = .328}$ $\boldsymbol{\gamma_{B_{1/2}} = .734}$ $\boldsymbol{\beta^{B}_{1/2} = -1.116}$ $\boldsymbol{\beta_{B_{1/2}} = -.3095}$

$\gamma^{B\mid A}_{1/2\ i}$	1.857	1.363	1	0,734	0,538
$\beta^{B\mid A}_{1/2\ i}$	**0,619**	**0,309**	**0**	**-0,309**	**-0,619**

CATEGORICAL LATENT VARIABLE MODELS

Latent: not (directly) observed

- **measurement model; true scores**
- **Independent Classification Errors (ICE)/'random' misclassifications; unreliability, invalidity**
- **Systematic Error/'correlated' misclassifications; unreliability, invalidity**
- **(underlying) typologies/clusters**
- **Unobserved Heterogeneity**
- **(partial) Nonresponse/Missing Data**
- **random coefficient models/multi-level analysis**

consequences of ICErrors

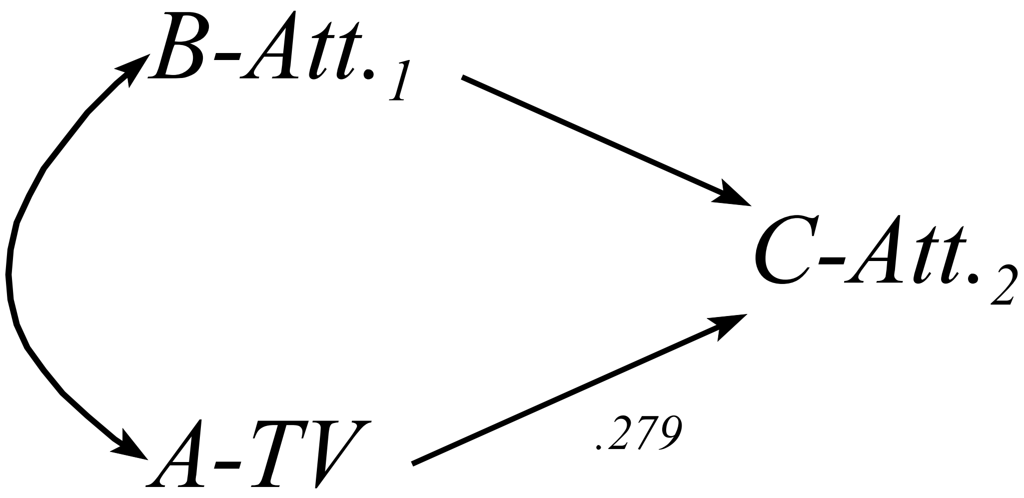
- **biased marginals (means and variances)**
- **biased (size and pattern of) association/turnover; biased 'causal' conclusions**

ILLUSTRATION

		A. Watching tv-program		
B - Att.1	*C - Att.2*	*1. Yes*	*2. No*	*Tot.*
1. Fav.	*1. Fav.*	**81**	**326**	**407**
1. Fav.	*2. Unf.*	**10**	**89**	**99**
2. Unf.	*1. Fav.*	**10**	**89**	**99**
2. Unf.	*2. Unf.*	**11**	**484**	**495**
	Total	**112**	**988**	**1100**

Watchers only:

	C. Att. After-Posttest		
B. Att. Before -Pretest	*1. Fav.*	*2. Unfav.*	*Tot.*
1. Fav.	**89.0%**	**11.0%**	**100%**
2. Unf.	**47.6%**	**52.4%**	**100%**
Total	**81.3%**	**18.7%**	**100**



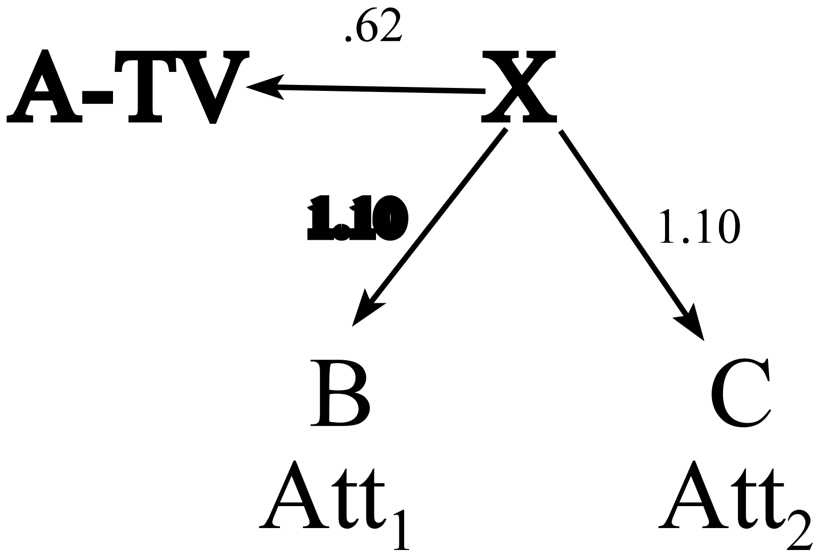
{ABC} : $\hat{\lambda}^{ABC}_{111} = -.101, \hat{\sigma}_{\hat{\lambda}} = .072$

{AB, BC} : $L^2 = 17.11, df = 2, p = .00$

{AB, AC, BC} : $L^2 = 1.89\ df = 1\ p = .17$
$\hat{\lambda}^{AC}_{11} = .279,\ \hat{\sigma}_{\hat{\lambda}} = .075$

TV	TRUE *ATT.*	MEAS. t_1	MEAS. t_2
W - 112	F - 101 (90%)	F - 90.9	F - 81.81 U - 9.09
		U - 10.1	F - 9.09 U - 1.01
	U - 90 (10%)	F - 1.1	F - .11 U - .99
		U - 9.9	F - .99 U - 8.91
$\bar{W}$. - 988	F - 395 (40%)	F - 355.5	F - 319.95 U - 35.55
		U - 39.5	F - 35.55 U - 3.95
	U - 593 (60%)	F - 59.3	F - 5.93 U - 53.37
		U - 533.7	F - 53.37 U - 480.33

For all measurements: probability of misclassification .10 (probability of correct classification .90)



Effects: estimated λ-parameters

Loglinear Model with Latent Variables {AX, BX, CX}

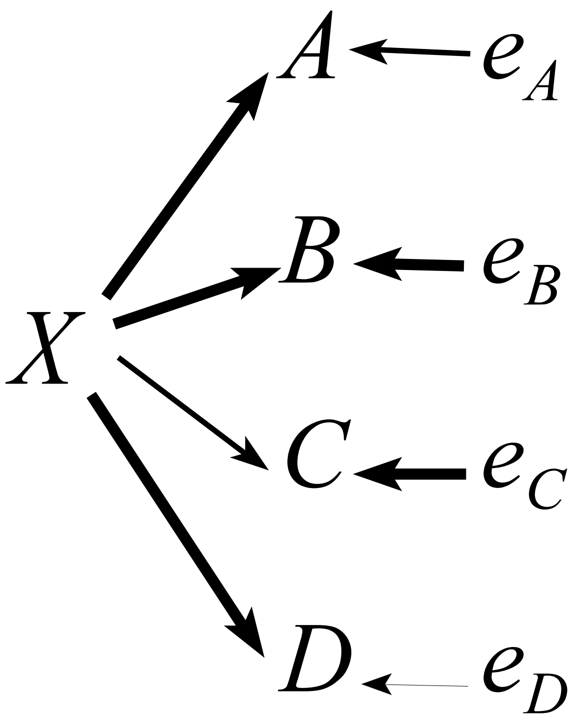
$$F^{ABCX}_{i\,j\,k\,t} = \eta \tau^{A}_{i} \tau^{B}_{j} \tau^{C}_{k} \tau^{X}_{t} \tau^{AX}_{i\,t} \tau^{BX}_{j\,t} \tau^{CX}_{k\,t}$$

Latent Class Model

$$\pi^{ABCX}_{i\,j\,k\,l} = \pi^{X}_{t} \pi^{A|X}_{i\;t} \pi^{B|X}_{j\;t} \pi^{C|X}_{k\;t} \qquad \left(\pi^{ABC}_{i\,j\,k} = \sum_{t} \pi^{ABCX}_{i\,j\,k\,t}\right)$$

	$\hat{\pi}^{X}_{t}$	$\hat{\pi}^{A\|X}_{i\;t}$		$\pi^{B\|X}_{j\;t}$		$\hat{\pi}^{C\|X}_{k\;t}$	
		1	***2***	***1***	***2***	***1***	***2***
1.	**.45**	**.20**	**.80**	**.90**	**.10**	**.90**	**.10**
2.	**.55**	**.02**	**.98**	**.10**	**.90**	**.10**	**.90**

LATENT STRUCTURE MODELS



- **latent (X) + manifest variables (A-D)**
- **local (conditional) independence**
- **probabilistic relation between latent and manifest variables; error terms**
-

		latent		
		nom.	***ord.***	***int.***
	nom.	**1**	**2**	**3**
manifest	***ord.***	**4**	**5**	**6**
	int.	**7**	**8**	**9**

1-standard latent class analysis; 3-latent trait; 5-OLCA; 7-latent profile; 9-factor analysis, latent trait.
LCA: in all cells, but all variables categorical.

LATENT CLASS ANALYSIS; LOGLINEAR MODEL WITH LATENT VARIABLES

$$\pi^{ABCDX}_{ijklt} = \pi^{X}_{t}\pi^{ABCD|X}_{ijkl\ t}$$
$$= \pi^{X}_{t}\pi^{A|X}_{i\ t}\pi^{B|X}_{j\ t}\pi^{C|X}_{k\ t}\pi^{D|X}_{l\ t}$$

{AX, BX, CX, DX}:

$$F^{ABCDX}_{ijklt} = \eta\tau^{A}_{i}\tau^{B}_{j}\tau^{C}_{k}\tau^{D}_{l}\tau^{X}_{t}\tau^{AX}_{i\ t}\tau^{BX}_{j\ t}\tau^{CX}_{k\ t}\tau^{DX}_{l\ t}$$

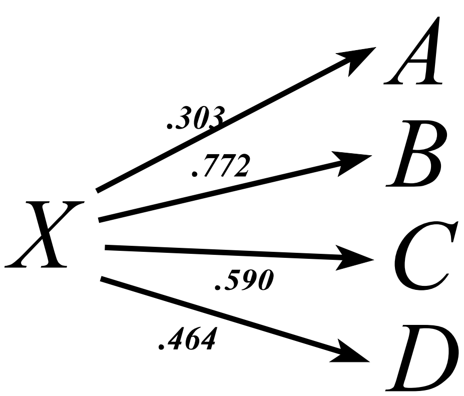
manifest table: $\pi^{ABCD}_{ijkl} = \sum_{t=1}^{T} \pi^{ABCDX}_{ijklt}$

number of latent variables AND for each latent variable: number of categories (latent classes)

EXAMPLE

Essential Government Responsibility (yes/no)

A-Equal Rights Men/Women
B-Good Education
C-Good Medical Care
D-Equal Rights Guest Workers



{XA,XB,XC,XD}

two latent classes

L^2 = 13.99 df = 6 p = .03
X^2 = 13.97

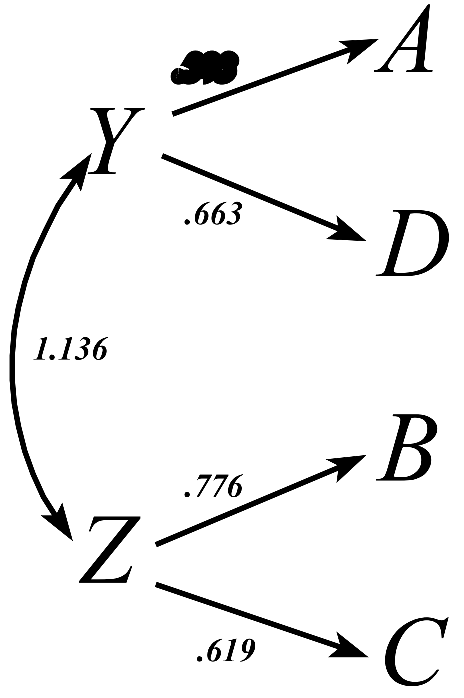
X	$\hat{\pi}^{X}_{t}$	$\hat{\pi}^{A\|X}_{i\ t}$		$\hat{\pi}^{B\|X}_{j\ t}$		$\hat{\pi}^{C\|X}_{k\ t}$		$\hat{\pi}^{D\|X}_{l\ t}$	
		1	*2*	*1*	*2*	*1*	*2*	*1*	*2*
1	.41	.40	.60	.95	.05	.85	.15	.47	.53
2	.59	.17	.83	.47	.53	.35	.65	.12	.88

{YZ, YA, YD, ZB, ZC}

two dichotomous latent variables

L^2 = 5,76 *df* = 4
p = .22 X^2 = 5.75

X-YZ	$\hat{\pi}^{YZ}_{rs}$	$\hat{\pi}^{A\|X}_{i\ t}$		$\hat{\pi}^{B\|YZ}_{j\ rs}$		$\hat{\pi}^{C\|X}_{k\ t}$		$\hat{\pi}^{D\|X}_{l\ t}$	
		1	*2*	*1*	*2*	*1*	*2*	*1*	*2*
1-11	.259	.507	.493	.948	.052	.852	.148	.654	.346
2-12	.009	.507	.493	.449	.551	.327	.673	.654	.346
3-21	.176	.176	.824	.948	.052	.852	.148	.118	.882
4-22	.556	.176	.824	.449	.551	.327	.673	.118	.882



- **MLE(stimation): (product)multinomial sampling distribution**

- **Newton/Raphson and/or EM-algorithm**

- **EM-algorithm: (the Estimation/Maximization algorithm for missing data):**

  - **start with random estimates for all parameters and compute $\hat{F}(0)$**

  - **E(expectation)-step: find the conditional expected values of the 'missing data' given the observed data and current estimated parameters (find the 'sufficient statistics' or the complete table):**

$$\hat{f}^{ABCDX}_{i\,j\,k\,l\,t} = \frac{\hat{F}^{ABCDX}_{i\,j\,k\,l\,t}}{\hat{F}^{ABCD}_{i\,j\,k\,l}} f^{ABCD}_{i\,j\,k\,l} = \hat{\pi}^{ABCD\overline{X}}_{i\,j\,k\,l\,t}\ f^{ABCD}_{i\,j\,k\,l}$$

$$\textit{note}: \ (\hat{f}^{ABCDX}_{i\,j\,k\,l\,+} = f^{ABCD}_{i\,j\,k\,l})$$

- **M(aximization)-step: maximize the expected loglikelihood in the ‘normal’ way as if the complete table was observed (using IPF, N/R, etc.)**

- **repeat E and M step till convergence**

– **local and global maxima**

– **terminal/boundary estimates**

– **identifiability: necessary and sufficient conditions**

– **testing: standard χ^2-statistics (L^2, X^2) and ‘nonstandard’ test statistics.**
Be careful: $L^2_{r/u}$ cannot be used to determine the number of latent classes

– **modifying models (if a model does not fit):**
  - **inspect residuals (e.g., direct effects among indicators)**
  - **inspect 'loadings'**
  - **more latent classes**
  - **more latent variables**

RESTRICTED MODELS

- **restrictions on the (conditional response) probabilities or on the loglinear parameters**

- **setting parameters equal to a constant, e.g.,**

$$\pi_{i\ t}^{A|X} = 1 \qquad = 0 \qquad = .5$$

$$\pi_{t}^{X} = c \qquad \lambda_{i\,t}^{AX} = c$$

- **setting parameters equal to each other, e.g.,**

$$\pi_{1\,1}^{A|X} = \pi_{2\,2}^{A|X} \qquad \pi_{i\,t}^{A|X} = \pi_{i\,t}^{B|X} \qquad \lambda_{i\,t}^{AX} = \lambda_{i\,t}^{BX}$$

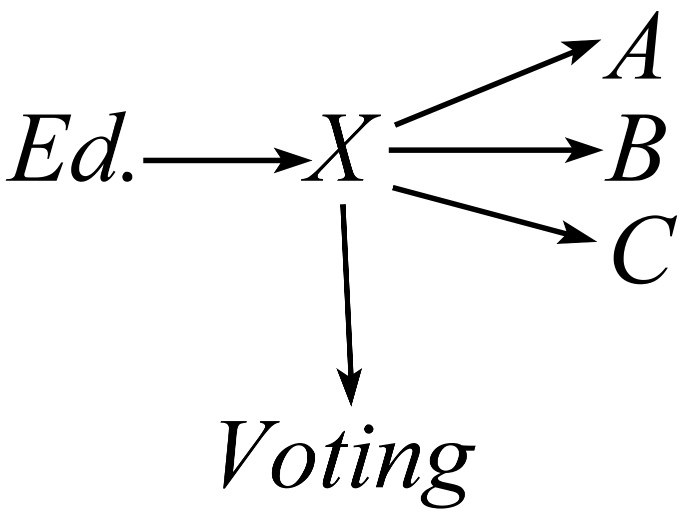
- **ordered data: 'linear' and ordinal restrictions, e.g.,**

$$\lambda_{i\,t}^{AX} = (\lambda_{A_i})X_t \qquad \pi_{i\ 1}^{A|X} \geq \pi_{i\ 2}^{A|X} \geq ,\ldots, \pi_{i\ T}^{A|X}$$

- **nonsaturated models for latent variables, e.g.,**

$$\{\mathrm{Y,Z}\}\!: \ \pi_{r\,s}^{YZ} = \pi_{r}^{Y}\pi_{s}^{Z}$$

- **extra restrictions can make identified models unidentifiable**



'estimating' latent scores

– **compute the classification probability:**

$$\hat{\pi}^{X|ABCD}_{t\ i\ j\ k\ l} = \hat{\pi}^{ABCDX}_{i\ j\ k\ l\ t} \,/\, \hat{\pi}^{ABCDX}_{i\ j\ k\ l\ +}$$

– **assign to conditional modal class X' = t*, given observed pattern**

– **assignment errors**

$$\epsilon^{ABCD}_{i\,j\,k\,l} = 1 - \pi^{X|ABCD}_{t^*\,i\ j\ k\ l}$$

$$E = \sum_{i,j,k,l} \pi^{ABCD}_{ijkl}\, \epsilon^{ABCD}_{ijkl}$$

$$\lambda_{X.ABCD} = \frac{(1-\pi_{t^*}) - E}{(1-\pi_{t^*})}$$

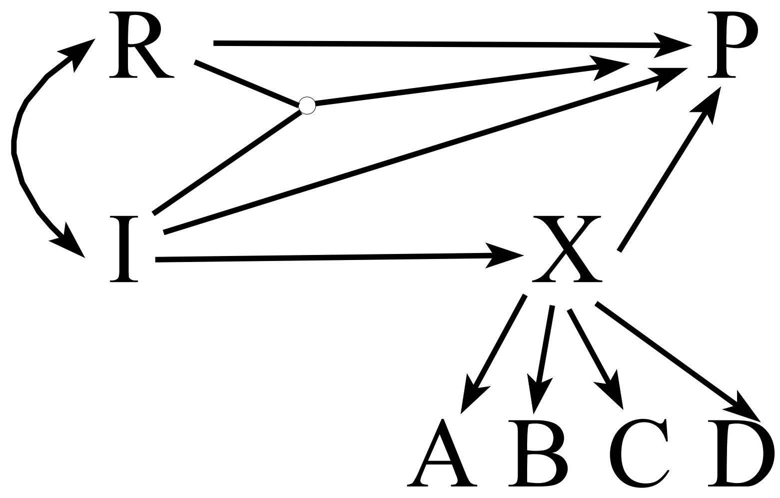
- **if variables are interval level: assign all respondents to conditional expected value (average) of X**

- **'Estimated' latent scores X_i' are not individually identified and are biased (as are factor scores)**

- **Solution: make external variables 'internal', that is fit model {EX,VX,AX,BX,CX} (cf. factor analysis and LISREL)**

loglinear models with latent variables; modified path models with latent variables/modified lisrel models/directed loglinear models with latent variables

DIRECTED LOGLINEAR MODELS WITH LATENT VARIABLES



$$\pi^{RIPABCDX}_{mnoijklt} = (\pi^{RI}_{mn}\pi^{X|RI}_{t\,mn}\pi^{P|RIX}_{o\,mnt}).(\pi^{A|T}_{i\,t}\pi^{B|T}_{j\,t}\pi^{C|T}_{k\,t}\pi^{D|T}_{l\,t})$$

$$= (\pi^{RI}_{mn}\pi^{X|I}_{t\,n}\pi^{P|RIX}_{o\,mnt}).(\pi^{A|T}_{i\,t}\pi^{B|T}_{j\,t}\pi^{C|T}_{k\,t}\pi^{D|T}_{l\,t})$$

Note: structural (causal) part plus measurement part

E-step: $$\hat{f}^{RIPABCDX}_{mnoijklt} = f^{RIPABCD}_{mnoijkl}\,\hat{\pi}^{X|RIPABCD}_{tmnoijkl}$$

M-step:

Table	Model
RI	{RI}
IX	{IX}
RIXP	{RIX, RIP, XP}
AX	{AX}
BX	{BX}
CX	{CX}
DX	{DX}

MARGINAL MODELS

- **(loglinear) models for linear combinations (sums) of cell frequencies**

- **marginal homogeneity:**

O_1 (X_e)
----------------- ***vs.*** **O_1 X_e O_2**
(X_e) O_2

- **comparing associations etc.**

O_1 X_e O_2

O_3 X_c O_4

NONRESPONSE IN CATEGORICAL DATA

- response mechanisms: ignorable and nonignorable; MAR and MCR

- Weighting

- nonresponse as 'just' another category: for nominal data

- listwise deletion
  - loss of cases/power
  - biased results, unless MCAR

- pairwise deletion
  - (less) loss of power
  - MCAR

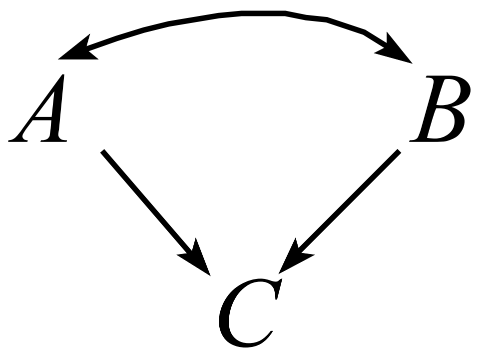
- imputation
  - (conditional) mean imputation, if MAR
  - proportional, if MAR
  - modal (biased)
  - biased 'variances': multiple imputation

- **Modeling nonresponse: computing the expected sufficient statistics**

  - **indicator variables: R_A= 0: missing on A, R_A = 1: responding/not missing on A; R_B= 0: missing on B, R_B = 1: responding/not missing on B, etc.**

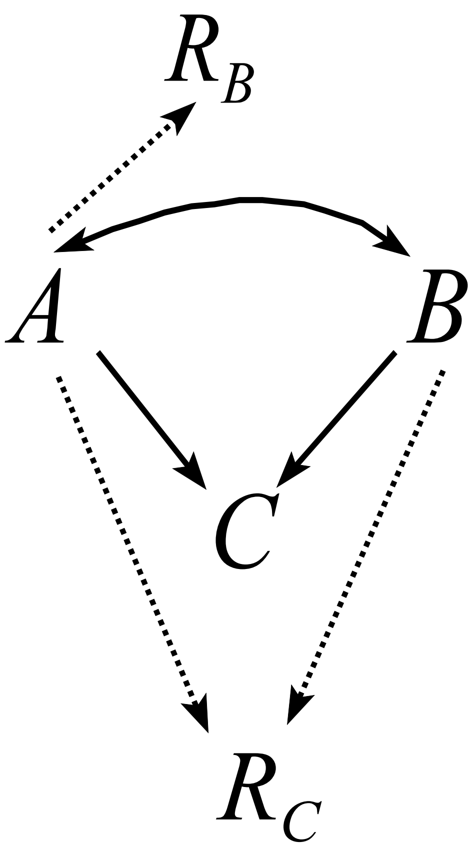
  - **MCAR, MAR ignorable, nonignorable response mechanisms**

  - **selection models (R's are caused by A, B) pattern-mixture models (R's cause A, B, etc.)**



- **complete case analysis: f^{ABC}_{ijk} model { AB, AC, BC}**

–   missing data analysis



Response Groups:

ABC: $f^{ABCR_BR_C}_{i\,j\,k\,1\,1}$ $(f^{ABC}_{i\,j\,k})$

AC : $f^{ABCR_BR_C}_{i\,j\,k\,0\,1}$ $(f^{AC}_{i\,k})$

AB : $f^{ABCR_BR_C}_{i\,j\,k\,1\,0}$ $(f^{AB}_{i\,j})$

A : $f^{ABCR_BR_C}_{i\,j\,k\,0\,0}$ (f^{A}_{i})

EM- algorithm

E-step:

$$\hat{f}^{ABCR_BR_C}_{i\,j\,k\,1\,1} = f^{ABC}_{i\,j\,k}$$

$$\hat{f}^{ABCR_BR_C}_{i\,j\,k\,0\,1} = f^{AC}_{i\,k}\,\hat{\pi}^{B|ACR_BR_C}_{j\;\;i\,k\,0\,1}$$

$$\hat{f}^{ABCR_BR_C}_{i\,j\,k\,1\,0} = f^{AB}_{i\,j}\,\hat{\pi}^{C|ABR_BR_C}_{k\;\;i\,j\,1\,0}$$

$$\hat{f}^{ABCR_BR_C}_{i\,j\,k\,0\,0} = f^{A}_{i}\,\hat{\pi}^{BC|AR_BR_C}_{j\,k\,i\,0\,0}$$

M-step: modified path model for table $\mathbf{ABCR_AR_B}$