`catsem`

A Stata ado for categorical data analysis with latent variables

Hans-Jürgen Andreß, Maximilian Hörl & Alexander Schmidt-Catran

University of Cologne
hja@wiso.uni-koeln.de

26.6.2015

Definition: categorical variables

- variables with just a few exhaustive and mutually exclusive categories
- nominal, ordinal, metric scale
- abound in social science (survey) research

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by "general linear reality" (Abbott 1988)
- "mostly harmless econometrics" (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by “general linear reality” (Abbott 1988)
- “mostly harmless econometrics” (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by “general linear reality” (Abbott 1988)
- “mostly harmless econometrics” (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by "general linear reality" (Abbott 1988)
- "mostly harmless econometrics" (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by "general linear reality" (Abbott 1988)
- "mostly harmless econometrics" (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by "general linear reality" (Abbott 1988)
- "mostly harmless econometrics" (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by "general linear reality" (Abbott 1988)
- "mostly harmless econometrics" (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by "general linear reality" (Abbott 1988)
- "mostly harmless econometrics" (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by "general linear reality" (Abbott 1988)
- "mostly harmless econometrics" (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by “general linear reality” (Abbott 1988)
- “mostly harmless econometrics” (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by "general linear reality" (Abbott 1988)
- "mostly harmless econometrics" (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

Why do we want something like `catsem`?

Reason 1: Theoretical

- social sciences dominated by "general linear reality" (Abbott 1988)
- "mostly harmless econometrics" (Angrist und Pischke 2008)
- non-linear models have become increasingly popular
- however, latent variables almost always treated as continuous
- see, e.g., Stata with `sem` and `gsem`
- but it is easy to find counter examples
  - social class (Marx), authority (Dahrendorf), deprivation (Townsend)
  - typologies
  - heterogenous samples (movers & stayers, attitudes & non-attitudes, unobserved heterogeneity)
  - typological methods: cluster analysis, sequence analysis

Reason 2: Practical

- Second edition of German textbook on categorical data analysis (Andreß et al. 1997)

1. SEM without latent variables
   - Example 1: Data set on vote turnout
   - Path diagram
   - `catsem` command for the example
2. Measurement models including latent variables
   - Example 2: Data set on welfare state attitudes in the Netherlands
   - One latent variable
   - Two latent variables
   - Example 3: Data set on welfare state attitudes in two countries
   - Measurement equivalence
3. SEM with latent variables
   - Example 4: Data on party preferences and welfare state attitudes
   - Best fitting model for Example 4
4. Description of `catsem` ado

SEM without latent variables

Example 1: Data set on vote turnout

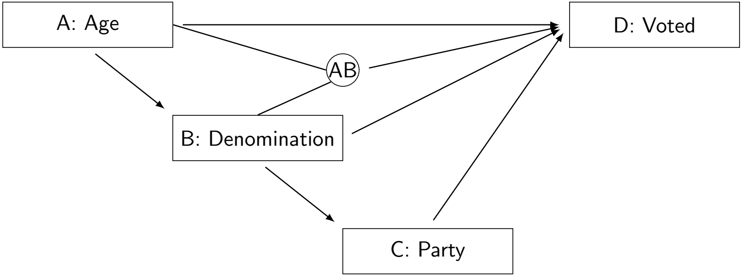
- Participated in election
  1. yes
  2. no
- Party preference
  1. SPD
  2. FDP
  3. CDU/CSU
- Member of a religious denomination
  1. yes
  2. no
- Age
  1. young
  2. old

Multivariate contingency table

A. Age	B. Denomination	C. Party preference	D. Voted	
			1. yes	2. no
1. young	1. with	1. SPD	38	13
		2. FDP	7	3
		3. CDU/CSU	60	20
	2. without	1. SPD	37	41
		2. FDP	35	25
		3. CDU/CSU	25	34
2. old	1. with	1. SPD	81	11
		2. FDP	20	1
		3. CDU/CSU	127	23
	2. without	1. SPD	31	34
		2. FDP	24	16
		3. CDU/CSU	19	25

Source: simulated data, see Andreß et al. (1997, Tabelle 1.2).

Path diagram



Directed acyclical graph (DAG)

Step 1: Causal order and distributional assumption

Variable	Predetermined	Subtable	Causal status
A: Age	–	–	exogenous
B: Denomination	age	AB	endogenous
C: Party	denomination, age	ABC	endogenous
D: Voted	party, denomination, age	ABCD	endogenous

Data distributed multinomially with

$$Pr(A=i, B=j, C=k, D=\ell) = \pi_{ijk\ell}^{ABCD} = \pi_i^{A} \times \pi_{j|i}^{B|A} \times \pi_{k|ij}^{C|AB} \times \pi_{\ell|ijk}^{D|ABC}$$

$$F_{ijk\ell}^{ABCD} = N \times \pi_{ijk\ell}^{ABCD} = N \times \pi_i^{A} \times \pi_{j|i}^{B|A} \times \pi_{k|ij}^{C|AB} \times \pi_{\ell|ijk}^{D|ABC}$$

Step 2: Hypothesized relationships

$$F_{ijk\ell}^{ABCD} = N \times \pi_i^A \times \pi_{j|i}^{B|A} \times \pi_{k|ij}^{C|AB} \times \pi_{\ell|ijk}^{D|ABC}$$

Link		Linear predictor	Log-linear model
$logit(\pi_{j\|i}^{B\|A})$	$=$	$\beta_{j\|i}^{B\|A}$	$\{AB\}$
$logit(\pi_{k\|j}^{C\|AB})$	$=$	$\beta_{k\|j}^{C\|B}$	$\{BC, AB\}$
$logit(\pi_{\ell\|ijk}^{D\|ABC})$	$=$	$\beta_{\ell\|i}^{D\|A} + \beta_{\ell\|j}^{D\|B} + \beta_{\ell\|ij}^{D\|AB} + \beta_{\ell\|k}^{D\|C}$	$\{ABD, CD, ABC\}$

Step 3: catsem command for the Example 1

Link		Linear predictor	Log-linear model
$logit(\pi_{j\|i}^{B\|A})$	$=$	$\beta_{j\|i}^{B\|A}$	$\{AB\}$
$logit(\pi_{k\|j}^{C\|AB})$	$=$	$\beta_{k\|j}^{C\|B}$	$\{BC, AB\}$
$logit(\pi_{\ell\|ijk}^{D\|ABC})$	$=$	$\beta_{\ell\|i}^{D\|A} + \beta_{\ell\|j}^{D\|B} + \beta_{\ell\|ij}^{D\|AB} + \beta_{\ell\|k}^{D\|C}$	$\{ABD, CD, ABC\}$

- ```
  catsem ///
  (i.age -> i.denomination) ///
  (i.denomination | i.age -> i.party) ///
  (i.age##i.denomination i.party -> i.voted) ///
  , lemdir("C:\lemwin")
  ```
- Stata: do, output
  ```

Measurement models including latent variables

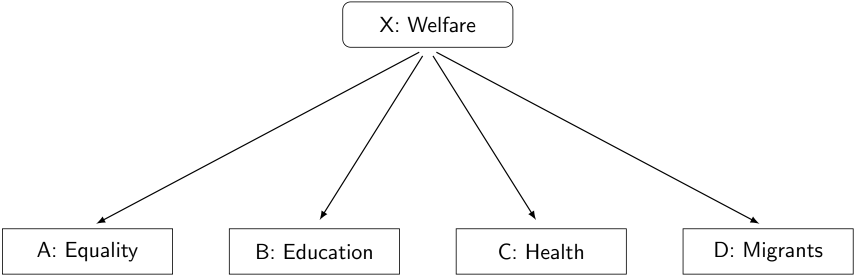
Example 2: Data set on welfare state attitudes in the Netherlands

A. Gender equality	B. Education	C. Health	D. Migrants	
			1. yes	2. no
1. yes	1. yes	1. yes	59	56
		2. no	14	36
	2. no	1. yes	7	15
		2. no	4	23
2. no	1. yes	1. yes	75	161
		2. no	22	115
	2. no	1. yes	8	68
		2. no	22	123

Source: Political Action Study (1973-76), see Andreß et al. (1997, Tabelle 1.4).

Path diagram with one latent variable

- Welfare state: encompassing vs. residual



catsem command for Example 2 (one latent variable)

- ```
  catsem ///
  (i.welfare -> i.equality i.education i.health
  i.migrants) ///
  , lemdir("C:\lemwin") latent(welfare(2)) seed(1234567)
  ```
- Stata: do, output
  ```

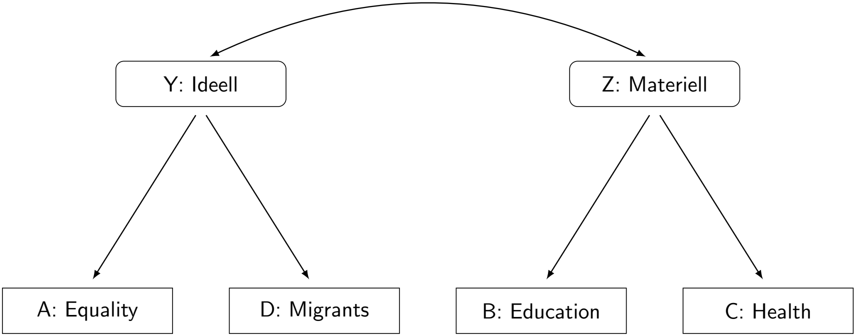
Latent class output

Latent class		A. Equality		B. Education		C. Health		D. Migrants	
X_t	$\hat{\pi}_t^x$	$\hat{\pi}_{i\|t}^{A\|X}$		$\hat{\pi}_{j\|t}^{B\|X}$		$\hat{\pi}_{k\|t}^{C\|X}$		$\hat{\pi}_{\ell\|t}^{D\|X}$	
		1. yes	2. no	1. yes	2. no	1. yes	2. no	1. yes	2. no
1	0,410	0,404	0,596	0,951	0,049	0,851	0,149	0,465	0,535
2	0,590	0,168	0,832	0,468	0,532	0,351	0,649	0,120	0,880

Note: $L^2 = 13.99$, $df = 6$, $p = 0.03$, $X^2 = 13.97$.
Estimated expected proportion of classification errors when using modal assignment: $E = 0.1668$.
Reduction in the proportion of classification errors: $\lambda = 0.5928$.

Path diagram with two latent variables

Welfare state responsible for “ideational” or “material” goods



catsem command for Example 2 (two latent variables)

- ```
  catsem ///
  (i.ideell -> i.equality i.migrants) ///
  (i.materiell -> i.education i.health) ///
  , lemdir("C:\lemwin") ///
  latent(ideell(2) materiell(2)) seed(222)
  ```
- Stata: do, output
  ```

Latent class output

Latent class r, s	$\hat{\pi}_{rs}^{YZ}$	A. Equality $\hat{\pi}_{i\|rs}^{A\|YZ}$ 1. yes	2. no	B. Education $\hat{\pi}_{j\|rs}^{B\|YZ}$ 1. yes	2. no	C. Health $\hat{\pi}_{k\|rs}^{C\|YZ}$ 1. yes	2. no	D. Migrants $\hat{\pi}_{\ell\|rs}^{D\|YZ}$ 1. yes	2. no
1,1	0.556	0.177	0.823	0.448	0.552	0.327	0.674	0.118	0.882
1,2	0.178	0.177	0.823	0.947	0.053	0.852	0.148	0.118	0.882
2,1	0.007	0.509	0.491	0.448	0.552	0.327	0.674	0.656	0.344
2,2	0.258	0.509	0.491	0.947	0.053	0.852	0.148	0.656	0.344

Note: $L^2 = 5.76$, $df = 4$, $p = 0.22$, $X^2 = 5.75$, $E = 0.2374$, $\lambda = 0.4650$

Measurement equivalence

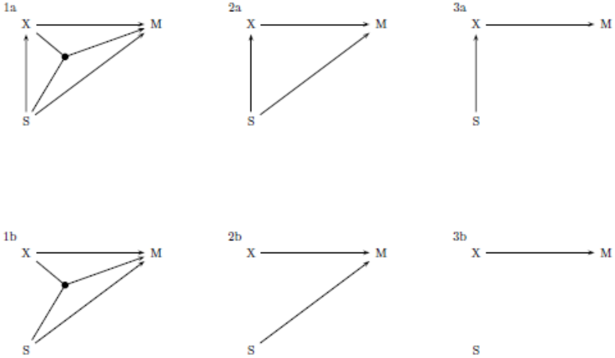
Example 3: Data set on welfare state attitudes in Germany and Switzerland

- S: Country
  1. Switzerland
  2. Germany
- A: Gender equality
  1. yes
  2. no
- B: Education
  1. yes
  2. no
- C: Health
  1. yes
  2. no
- D: Equal rights for migrants
  1. yes
  2. no

Source: Political Action Study (1973-76), see Andreß et al. (1997, Tabelle 4.3).

Types of measurement models

1. (completely) heterogenous (heterogenous slopes)
2. partially homogenous (heterogenous intercepts)
3. homogenous



Notes: S = group variable, X = latent variable(s), M = manifest variables.

Testing measurement invariance for Example 3

Type	Log-linear model	L^2	*df*	*p*
1a	$\{SYZ\}\{SYA, SYD, SZB, SZC, SYZ\}$	13.67	8	0.09
1b	$\{SY, Z\}\{SYA, SYD, SZB, SZC, SYZ\}$	17.23	11	0.10
2a	$\{SYZ\}\{YA, YD, ZB, ZC, SA, SB, SC, SD, SYZ\}$	18.56	12	0.10
2b	$\{S, YZ\}\{YA, YD, ZB, ZC, SA, SB, SC, SD, SYZ\}$	22.40	15	0.10
3a	$\{SYZ\}\{YA, YD, ZB, ZC, SYZ\}$	35.30	16	0.004
3b	$\{S, YZ\}\{YA, YD, ZB, ZC, SYZ\}$	76.02	19	0.000

Conditional Likelihood-Ratio-Tests

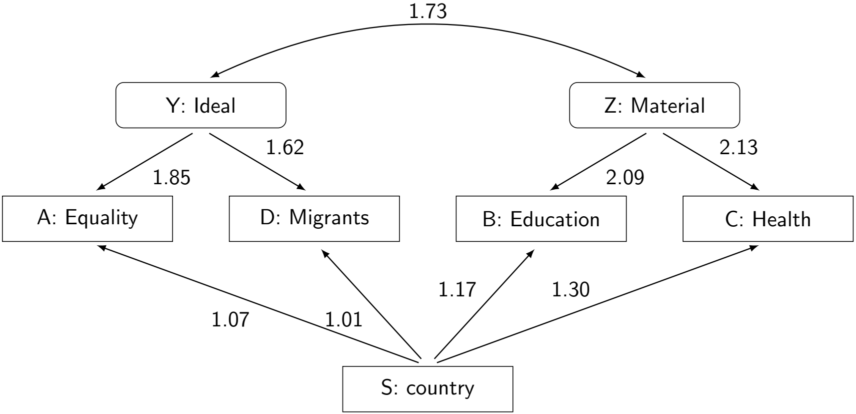
$L^2_{2a,1a} = 18.56 - 13.67 = 4.89,\ df = 12 - 8 = 4,\ p = 0.30$

$L^2_{2b,1a} = 22.40 - 13.67 = 8.73,\ df = 15 - 8 = 7),\ p = 0.27$

Best fitting model for Example 3

partially homogenous measurement model with heterogenous intercepts
identical structural model

Odds ratios, centered effects



Note: $L^2 = 22.40$, $df = 15$, $p = 0.10$, $X^2 = 22.22$.

catsem command for Example 3

- ```
  catsem ///
  (i.equality i.migrants <- i.ideell i.country) ///
  (i.education i.health <- i.materiell i.country) ///
  , lemdir("C:\lemwin") ///
  latent(ideell (2) materiell (2)) ///
  covstructure(i.ideell##i.materiell i.country)
  ```
- Stata: do, output
  ```

SEM with latent variables

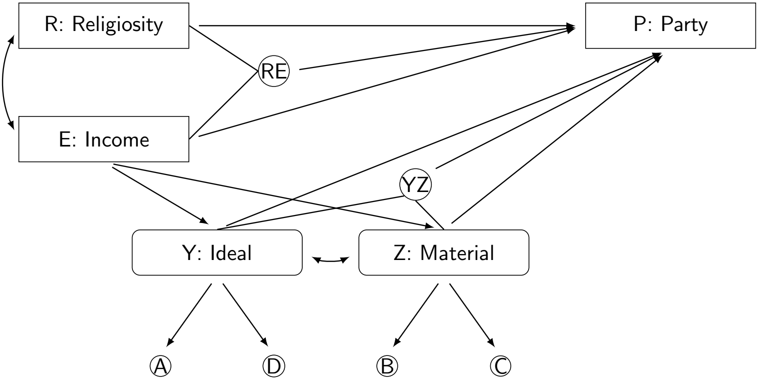
Example 4: Data on party preferences and welfare state attitudes in Germany

- P: Party preference
  1. left (SPD, DKP)
  2. center & right (CDU/CSU, FDP)
- R: Religiosity
  1. religious
  2. not religious
- E: Income
  1. less than 1,500 DM
  2. more than 1,500 DM
- Welfare state attitudes: gender equality (A), education (B), health (C), equal rights for migrants (D)
  1. yes
  2. no

Source: Political Action Study (1973-76), see Andreß et al. (1997, Tabelle 4.5).

Best fitting model for Example 4

measurement model could be restricted to 3 classes and Guttman structure



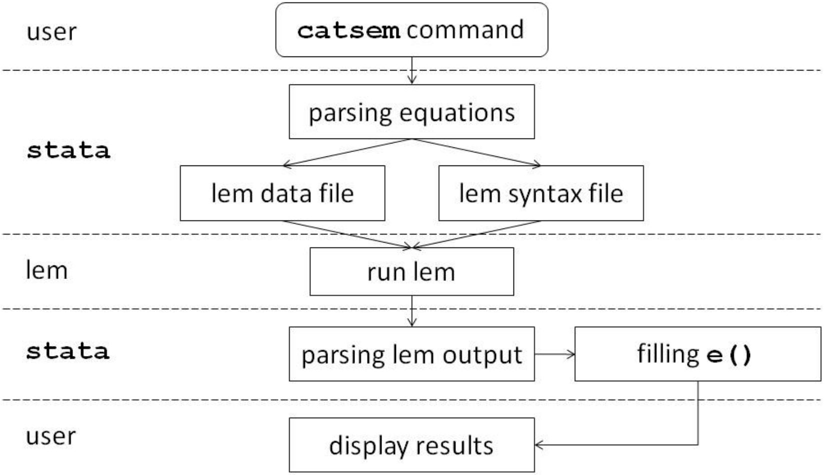
Note: $L^2 = 101.42$, $df = 105$, $p = 0.58$, $X^2 = 94.16$.

catsem command for Example 4

- ```
  catsem ///
  (i.income | i.religiosity -> i.materiell##i.ideell) ///
  (i.materiell -> i.education i.health) ///
  (i.ideell -> i.equality i.migrants) ///
  (i.materiell i.ideell i.religiosity##i.income ->
  i.party) ///
  , lemdir("C:\lemwin") ///
  latent(ideell (2) materiell (2))
  ```
- Stata: do, output
  ```

Flow chart of catsem ado

- uses external program ℓEM (Vermunt 1997) for estimation



catsem syntax

`catsem paths [if] [in] [, options]`

- `paths`
  - same syntax as Stata gsem command
  - possibility to specify "control" variables using '|'
  - possibility to specify "combined" endogenous variables using '##'
- `options`
  - `lemdir(path)`: directory of external program ℓEM; default: working directory of do file
  - latent(name(#) name(#) ...): specify latent variable(s) and their number of categories; default: no latent variables
  - covstructure(model): log-linear model for relationships among exogenous variables; default: saturated model
  - seed(#): specify a seed for random starting values; default: seed is derived from computer clock
  - iterations(#): specify max. number of iterations of EM algorithm; default: 5000
  - `lemout(fn)`, `leminp(fn)`, `lemcovar(fn)`, `lemlog(fn)`: specify a filename `fn` in the working directory for ℓEM input and output

What to do next

- Store latent class output in suitable Stata objects (similar to matrix of factor loadings in factor analysis)
- Enable predict command to show latent class probabilities (ℓEM: `wpo`)
- Flexible handling of base categories
- Restrictions on latent class probabilities and regression coefficients
- Ordinal dependent and continuous independent variables
- WLS estimation (Grizzle et al. 1969) for models including only categorical variables and no latent variables
- More options
  - ...
- Technicalities
  - improved reading of ℓEM's var-cov-matrix
  - error checking of user input
  - ...
- Implement EM algorithm within Stata

How to install ℓEM and catsem?

1. Download `lemwin.zip` from Jeroen Vermunt's website
   - http://members.home.nl/jeroenvermunt/lemfiles
2. Install LEM95.EXE on your computer
   - important: the path to the EXE must not include any blanks
   - specify the path in the `catsem` command with the option `lemdir(path)`, otherwise `catsem` will search for the EXE in your working directory
3. Install `catsem` in the directory for ado's
4. Check it out and report errors and problems to hja@wiso.uni-koeln.de

Thank you for your attention

Special thanks to
Jeroen K. Vermunt (Tilburg University)
who wrote this powerful program ℓEM
and answered all our stupid questions

Want to become our beta tester?
hja@wiso.uni-koeln.de

References

Abbott, A. (1988): Transcending general linear reality. *Sociological Theory*, *6*, 169–186.

Andreß, H.J./ Hagenaars, J.A./ Kühnel, S. (1997): *Analyse von Tabellen und kategorialen Daten: Log-lineare Modelle, latente Klassenanalyse, logistische Regression und GSK-Ansatz*. Springer-Lehrbuch, Berlin et al.: Springer.

Angrist, J./ Pischke, J. (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Grizzle, J./ Starmer, C./ Koch, G. (1969): Analysis of categorical data by linear models. *Biometrics*, *25*, 489–504.

Vermunt, J.K. (1997): *ℓEM: A general program for the analysis of categorical data*. Tilburg University.