Covariates and Distal Outcomes: 1-Step and 3-Step Approaches

Jeroen K. Vermunt
Department of Methodology and Statistics, Tilburg University
www.jeroenvermunt.nl

Introduction

- In the introductory video, I presented the three steps/goals of a LC analysis
  - Building a clustering model
  - Classifying individuals
  - Investigating the relationship with external/other variables
- This video deals with the last step/goal
  - Exploring the association between class membership and other variables
  - Predicting class membership using covariates
  - Using class membership as predictor of a (distal) outcome

Four possible approaches

1. <u>1-step approach</u>: covariates (and/or a distal outcome) are included in the estimated LC model

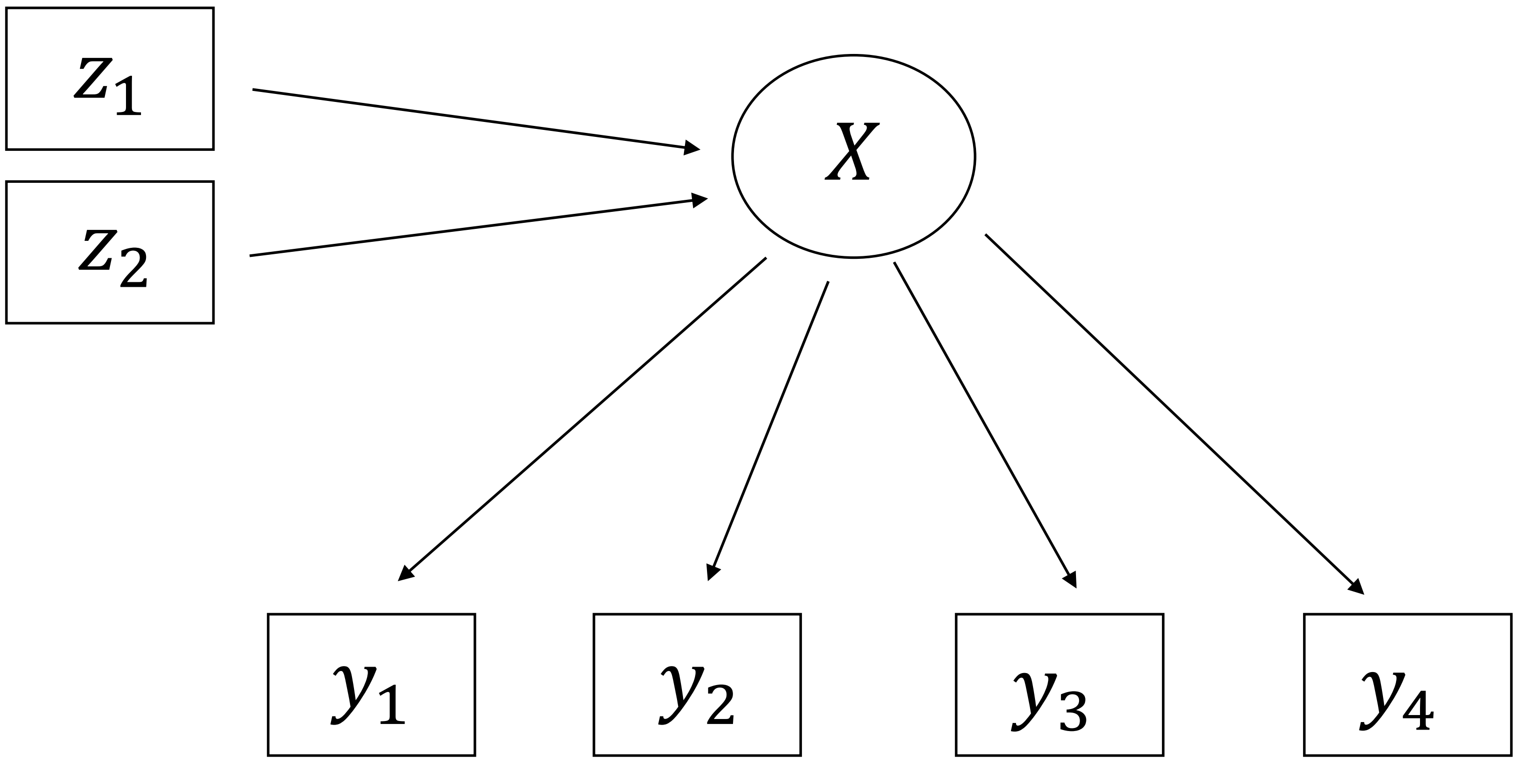
2a. <u>Inactive covariates</u> option in LatentGOLD: it computes Profile and ProbMeans output using posterior class membership probabilities

2b. <u>Standard 3-step</u> approach: after modal classification, analysis with other programs

2c. <u>Bias-adjusted 3-step</u> approach: classification to a file and subsequent analyses with the LatentGOLD Step3 module

1-step approach (covariates)

- LC model with two covariates z_1 and z_2



1-step approach (covariates)

- Formula for a LC model with two covariates z_1 and z_2:

$$P(y_1,\ldots,y_J \mid z_1,z_2) = \sum_{c=1}^{C} P(X=c \mid z_1,z_2)\prod_{j=1}^{J} P(y_j \mid X=c)$$

with a multinomial logit model for the latent variable:

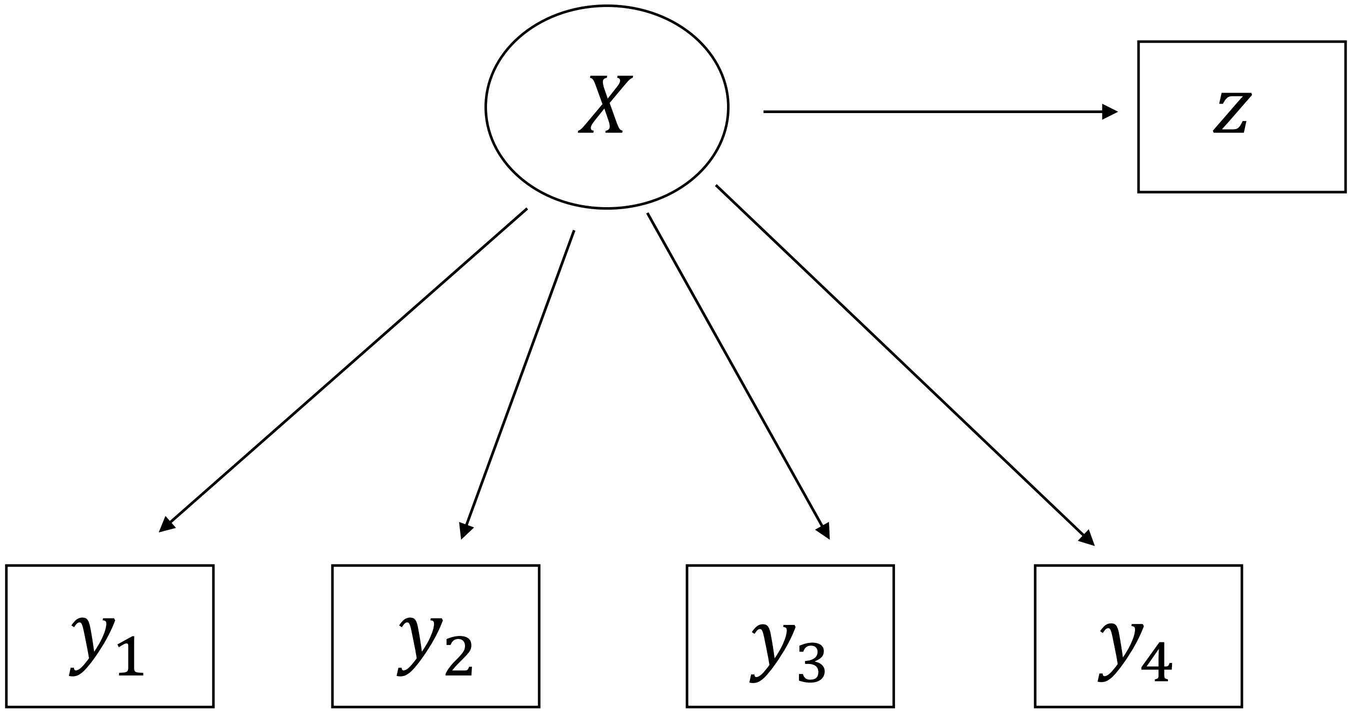
$$P(X=c \mid z_1,z_2) = \frac{\exp(\gamma_{0c}+\gamma_{1c}z_1+\gamma_{2c}z_2)}{\sum_{c'=1}^{C}\exp(\gamma_{0c'}+\gamma_{1c'}z_1+\gamma_{2c'}z_2)}$$

1-step approach (covariates)

- Seems to be straightforward, but this is clearly not the case
- Important to note the additional local independence assumption: covariates affect indicators only indirectly via the latent classes
- What if this does not hold in selected LC model:
  - Solution with original number of classes changes a lot
  - Number of classes needs to be increased
  - Direct effects need to be included for some covariate-indicator pairs
- But also if this holds, class definition will change somewhat, which in turn depends on the set of covariates included in the model

1-step approach (distal outcome)

- LC model with distal outcome *z*



1-step approach (distal outcome)

- Note from the graph that a distal outcome is in fact an additional indicator
- As in the covariate case, additional local independence assumptions need to made
- Moreover, when z is a continuous variable, distributional assumptions need to be made for $f(z|X=c)$, say normality, which may "distort" the classes if incorrect
- As in the covariates case, class definitions will always change somewhat when a distal outcome is included in the LC model

Other approaches

- Because of the practical issues associated with the 1-step approach, applied researchers prefer using approaches in which one does not include the covariates or distal outcome in the LC model itself
- Moreover, it looks a bit like cheating to allow the definition of the classes to depend on covariates and/or a distal outcome
- Other approaches:
  - Inactive covariates option in Latent GOLD
  - Standard 3-step analysis
  - Bias-adjusted 3-step analysis

Inactive covariates

- This is a feature specific for Latent GOLD, and yields a quick way to see how classes are related to other variables, including with plots.
- The LC model is estimated *without* covariates.
- Two-way covariate-class tables are created using the posterior class membership probabilities from this model.
- Profile and ProbMeans report the covariate distribution/mean given class and the class distribution given covariate value, respectively.
- Limitations: 1) associations are underestimated; 2) no statistical tests; 3) only bivariate relationships

3-step LC analysis

Step 1: Building a clustering model

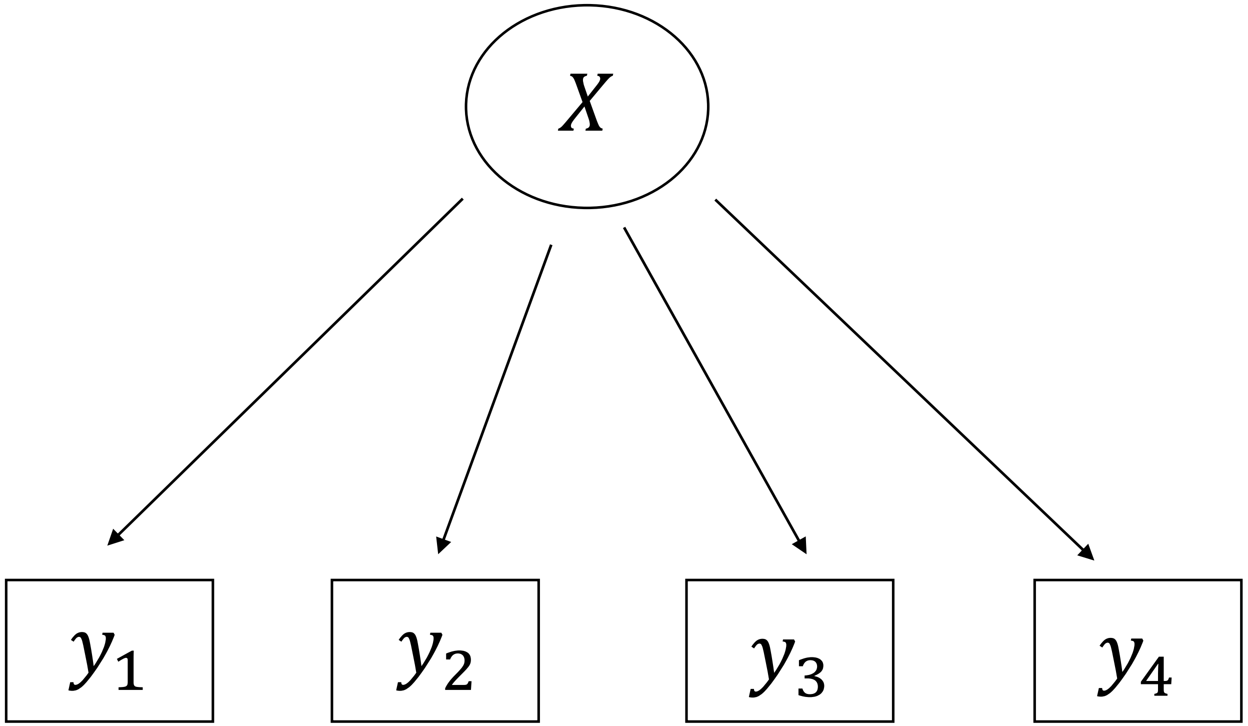
Step 2: Classifying individuals

Step 3: Investigating the relationship between the classifications and external/other variables

Fits naturally with the three goals/steps of LC modeling

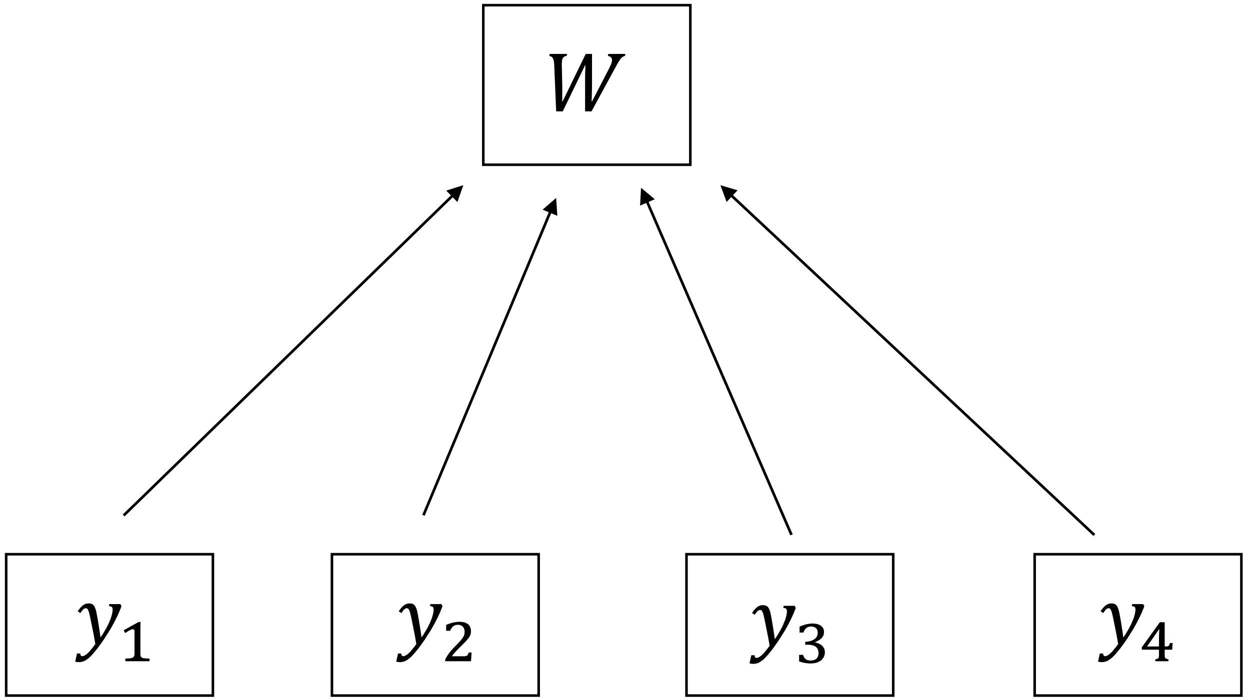
3-step LC analysis

- Step 1: Building a clustering model



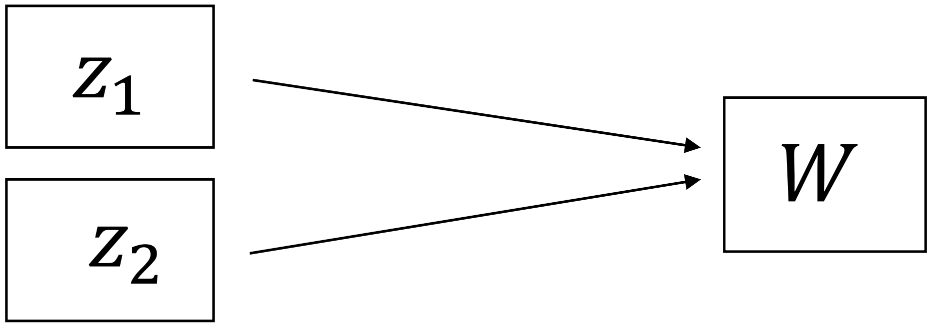
3-step LC analysis

- Step 2: obtaining classifications W (and adding these to the data file)

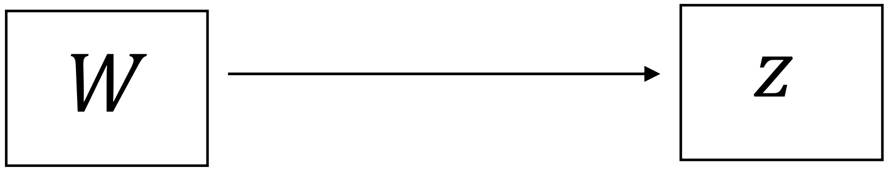


3-step LC analysis

- Step-3 model with covariates (say using SPSS logistic regression):



- Step-3 model with a distal outcome (say using SPSS Anova or Crosstabs):



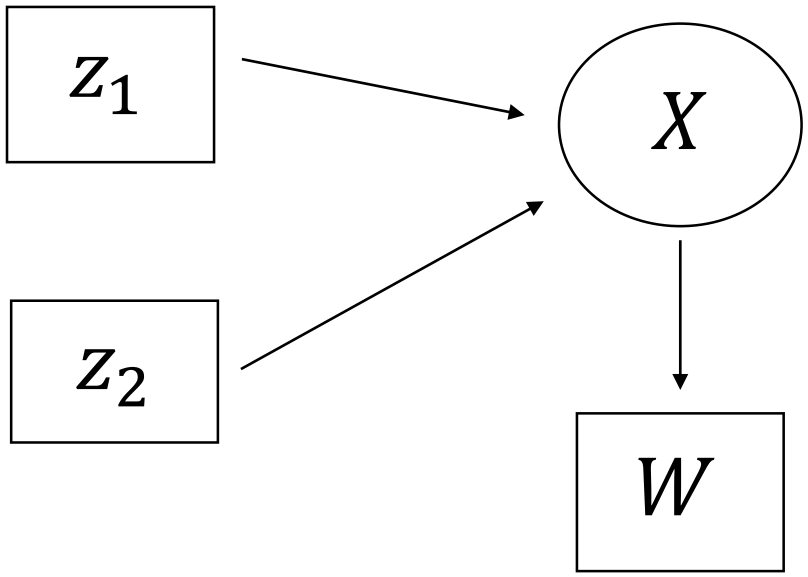
Problem with standard step-3 analysis

- You are investigating the relationship of *z*'s with *W* and not with *X.*
- *W* is an imperfect version of *X*, which contains classification errors.
- Result: associations/effects are underestimated (biased downwards).

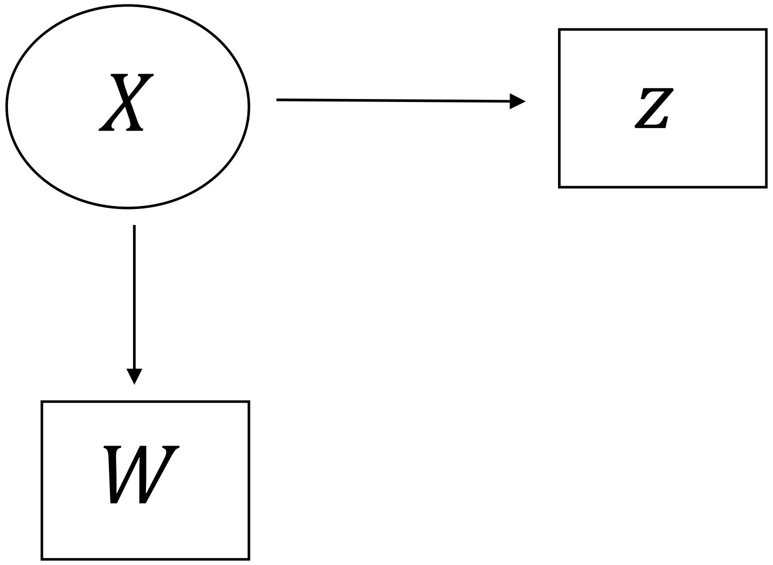
- Solution: correct for classification errors.
- Bolck, Croon and Hagenaars (BCH, 2004) proposed a bias-adjusted step-3 approach which was expanded and made practically applicable by Vermunt (2010).
- This has become the state-of-art approach.

Bias-adjusted step-3 analysis

Covariates:



Distal outcome (dependent):



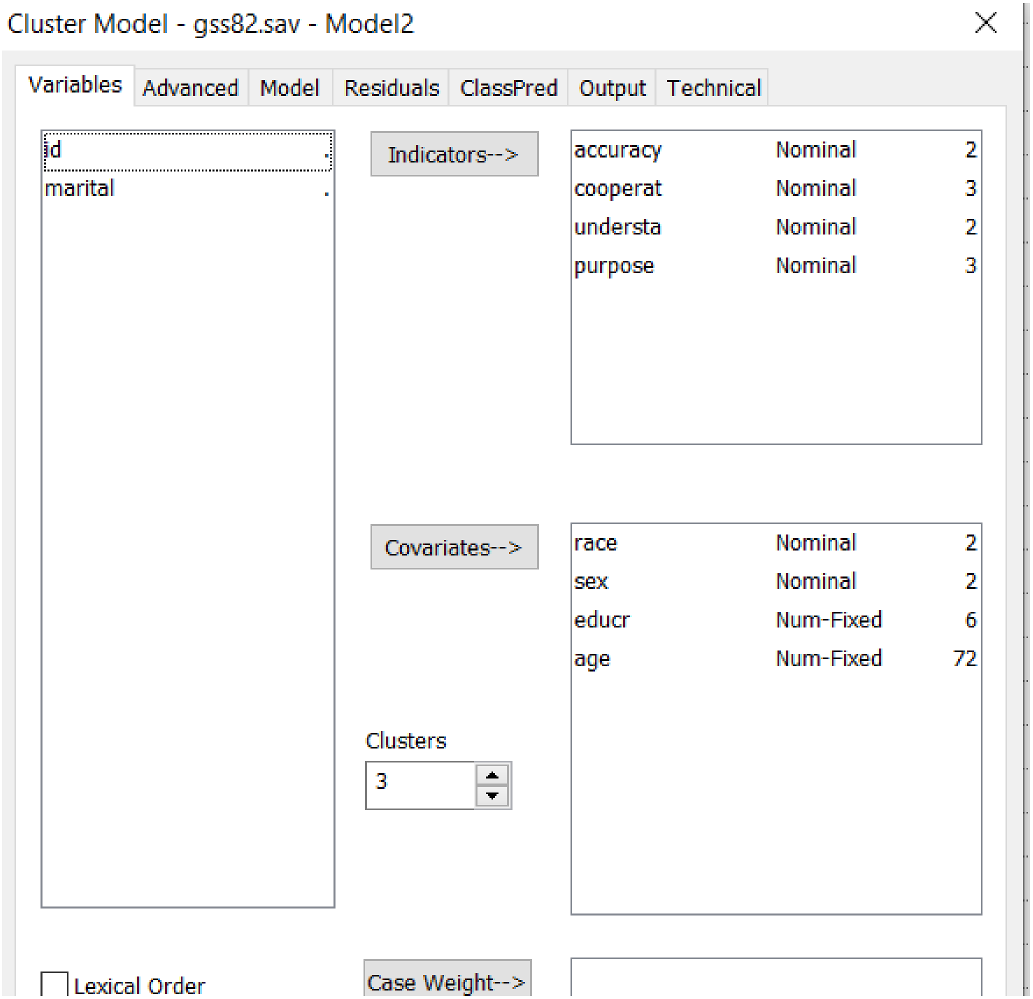
Bias-adjusted step-3 analysis

- As can be seen, we now model the relationship between z's and X, which is exactly what we want.
- We define a LC model in which W is use as a single indicator of X.
- Important: $P(W=t|X=c)$ is computed in step 2; it is the classification table rescaled to sum to 1 in the rows.
- One can use either modal or proportional classification.
- Two different estimation methods are ML and BCH, where BCH is mainly needed for continuous dependent variables.

Bias-adjusted 3-step analysis

- Also this approach has its limitations.
- It works less well when class separation is very bad; when the entropy R-squared is (much) smaller than .5.
- It is based on the same (additional) local independence assumptions as the 1-step approach (z's and y's are independent given X).
- Vermunt and Magidson (online first, Structural Equation Modeling) extended the 3-step approach to allow for direct effects between z's and y's (for differential item functioning).
- Many other extensions have been proposed to 3-step LC analysis.

GSS82.sav data: 1-step approach



GSS82.sav data: 1-step approach

- I included the predictors of class membership in the 3-class model (which was best according to BIC)
- On the positive side:
  - Covariate-indicator BVRs are all quite small
  - A 4-cluster models does not perform better according to BIC
- On the somewhat negative side:
  - Sample size slightly smaller because of missing values on age
  - Class definitions change somewhat and class one becomes quite a bit larger

GSS82.sav data: 1-step approach

- For interpretation, several output sections are of interest
- Parameters: Model for Clusters
  - Effect coded or dummy coded logit coefficients
  - Wald tests
  - See also Paired Comparisons (nested in Parameters)
- Profile: distribution/mean of covariates given class
- ProbMeans: class distribution given covariate values
- EstimatedValues: full model probabilities

GSS82.sav data: 1-step approach

- Parameters output (effect and dummy-first coding)

Model for Clusters					
Intercept	Cluster1	Cluster2	Cluster3	Wald	p-value
	0.4273	0.6938	-1.1211	16.3582	0.00028
Covariates	Cluster1	Cluster2	Cluster3	Wald	p-value
race					
WHITE	0.1347	-0.0953	-0.0394	8.7468	0.013
BLACK	-0.1347	0.0953	0.0394		
sex					
MALE	0.0217	0.0304	-0.0521	0.7271	0.70
FEMALE	-0.0217	-0.0304	0.0521		
educr					
	0.2831	-0.4965	0.2134	128.8204	1.1e-28
age					
	-0.0058	0.0032	0.0026	5.6375	0.060

Model for Clusters					
Intercept	Cluster1	Cluster2	Cluster3	Wald	p-value
	-0.0000	0.0453	-1.7963	18.1943	0.00011
Covariates	Cluster1	Cluster2	Cluster3	Wald	p-value
race					
WHITE	0.0000	-0.0000	-0.0000	8.7468	0.013
BLACK	-0.0000	0.4599	0.3481		
sex					
MALE	0.0000	-0.0000	-0.0000	0.7271	0.70
FEMALE	-0.0000	-0.0174	0.1477		
educr					
	-0.0000	-0.7797	-0.0698	128.8204	1.1e-28
age					
	-0.0000	0.0089	0.0084	5.6375	0.060

GSS82.sav data: 1-step approach

- ProbMeans and Profile

Covariates			
race			
WHITE	0.6943	0.1597	0.1460
BLACK	0.5774	0.2529	0.1698
sex			
MALE	0.6773	0.1823	0.1404
FEMALE	0.6517	0.1869	0.1614
educr			
0 - 1	0.4077	0.4603	0.1320
2 - 2	0.6391	0.2247	0.1362
3 - 3	0.7093	0.1255	0.1652
4 - 4	0.7930	0.0615	0.1455
5 - 5	0.8063	0.0197	0.1739
age			
18 - 26	0.7118	0.1508	0.1373
27 - 34	0.7519	0.0995	0.1485
35 - 47	0.6959	0.1462	0.1579
48 - 61	0.6035	0.2144	0.1821
62 - 89	0.5515	0.3134	0.1351

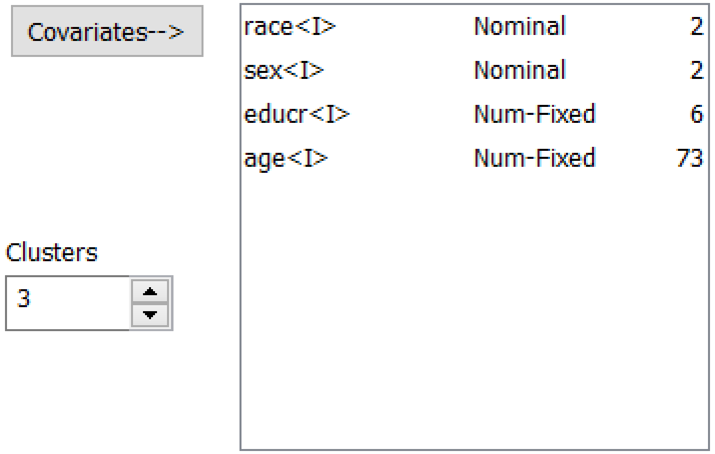
Covariates			
race			
WHITE	0.7640	0.6297	0.6985
BLACK	0.2360	0.3703	0.3015
sex			
MALE	0.4355	0.4200	0.3923
FEMALE	0.5645	0.5800	0.6077
educr			
0 - 1	0.1478	0.5980	0.2081
2 - 2	0.0696	0.0876	0.0644
3 - 3	0.3664	0.2323	0.3710
4 - 4	0.2399	0.0666	0.1913
5 - 5	0.1763	0.0155	0.1652
Mean	3.1975	1.5463	3.0063
age			
18 - 26	0.1970	0.1496	0.1651
27 - 34	0.2379	0.1128	0.2043
35 - 47	0.2118	0.1594	0.2089
48 - 61	0.1848	0.2353	0.2424
62 - 89	0.1684	0.3429	0.1793
Mean	42.0402	51.4287	44.7274

GSS82.sav data: 1-step approach

- EstimatedValues (or Classification-Model)

				Cluster		
race	sex	educr	age	1	2	3
WHITE	MALE	<8	28	0.3917	0.5261	0.0822
WHITE	MALE	<8	37	0.3729	0.5428	0.0844
WHITE	MALE	<8	40	0.3667	0.5482	0.0851
WHITE	MALE	<8	44	0.3585	0.5555	0.0860
WHITE	MALE	<8	48	0.3504	0.5626	0.0870
WHITE	MALE	<8	49	0.3484	0.5644	0.0872
WHITE	MALE	<8	50	0.3464	0.5662	0.0874
WHITE	MALE	<8	51	0.3444	0.5680	0.0876
WHITE	MALE	<8	56	0.3345	0.5768	0.0888
WHITE	MALE	<8	57	0.3325	0.5785	0.0890
WHITE	MALE	<8	60	0.3266	0.5837	0.0896
WHITE	MALE	<8	66	0.3150	0.5940	0.0909
WHITE	MALE	<8	67	0.3131	0.5957	0.0911
WHITE	MALE	<8	68	0.3112	0.5974	0.0914
WHITE	MALE	<8	69	0.3093	0.5991	0.0916
WHITE	MALE	<8	70	0.3075	0.6008	0.0918
WHITE	MALE	<8	74	0.3000	0.6074	0.0926
WHITE	MALE	<8	75	0.2981	0.6091	0.0928
WHITE	MALE	<8	79	0.2908	0.6157	0.0936
WHITE	MALE	<8	81	0.2871	0.6189	0.0940
WHITE	MALE	<8	82	0.2853	0.6205	0.0942
WHITE	MALE	<8	83	0.2835	0.6221	0.0944

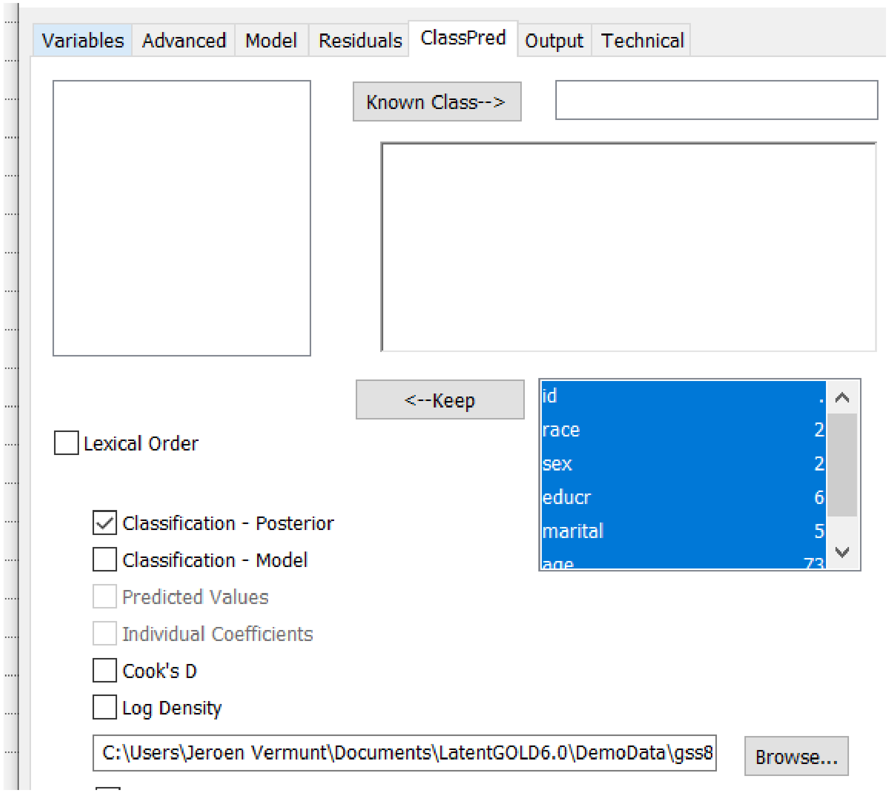
GSS82.sav data: inactive covariates



- Pattern of associations is the same as with 1-step approach, but associations are somewhat weaker
- See, for example, mean age of the classes (Profile) or the class probabilities for educational categories (ProbMeans)

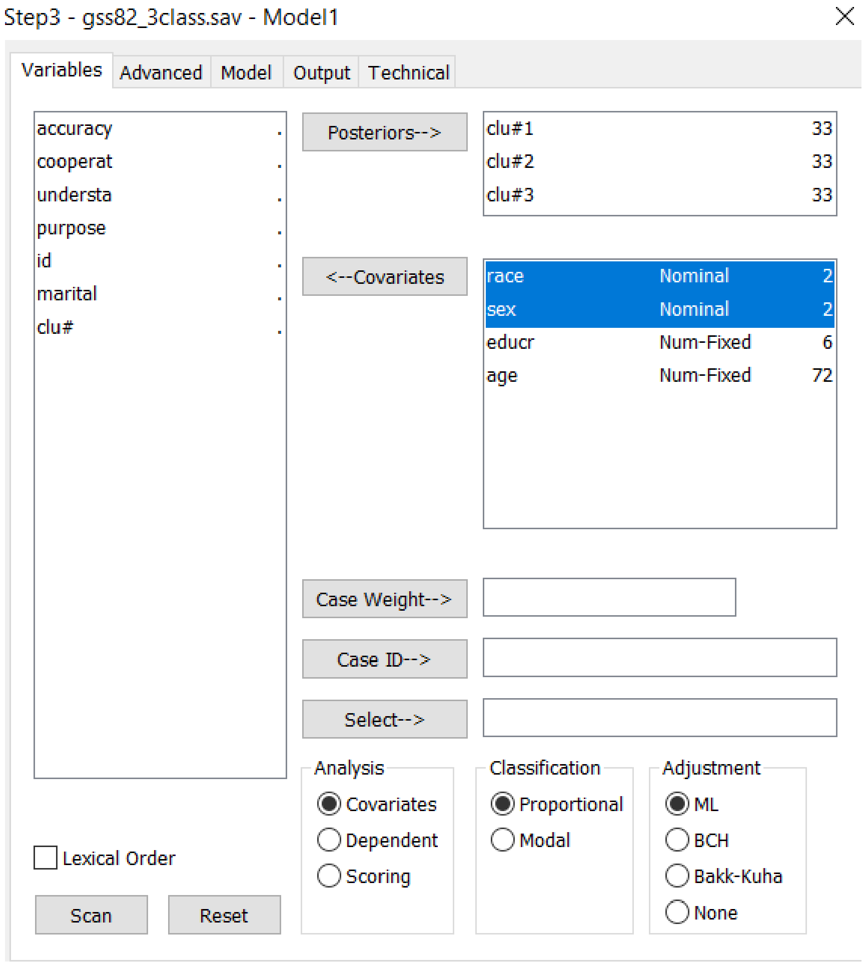
GSS82.sav data: bias-adjusted 3-step method

- First we should save the classifications to an output data file:



GSS82.sav data: bias-adjusted 3-step method

- With the file containing the classifications and the Step3 module …

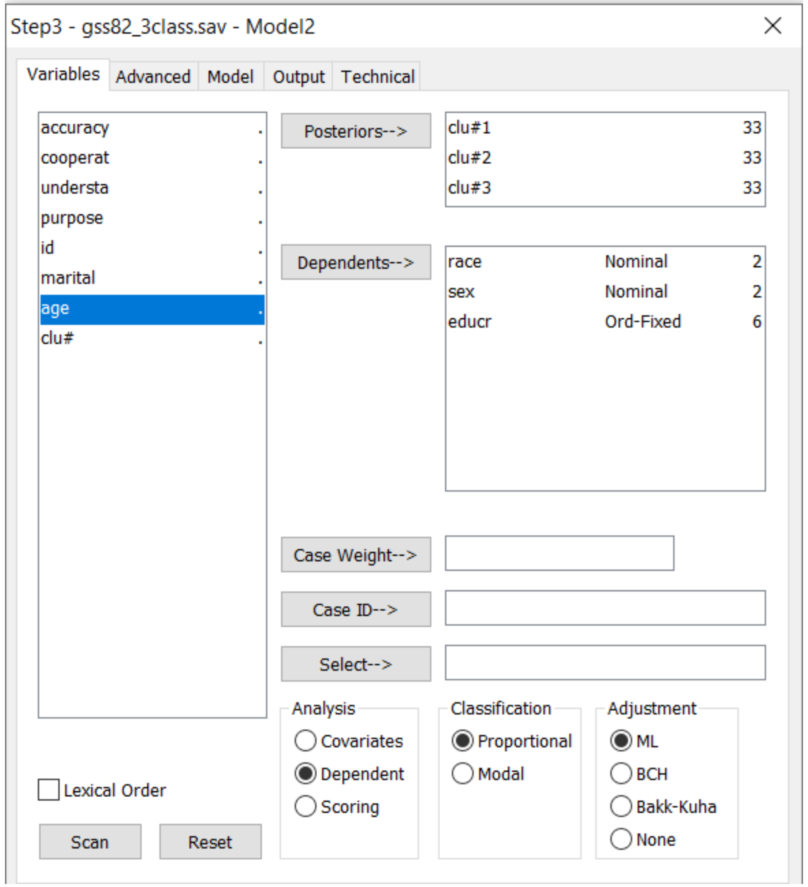


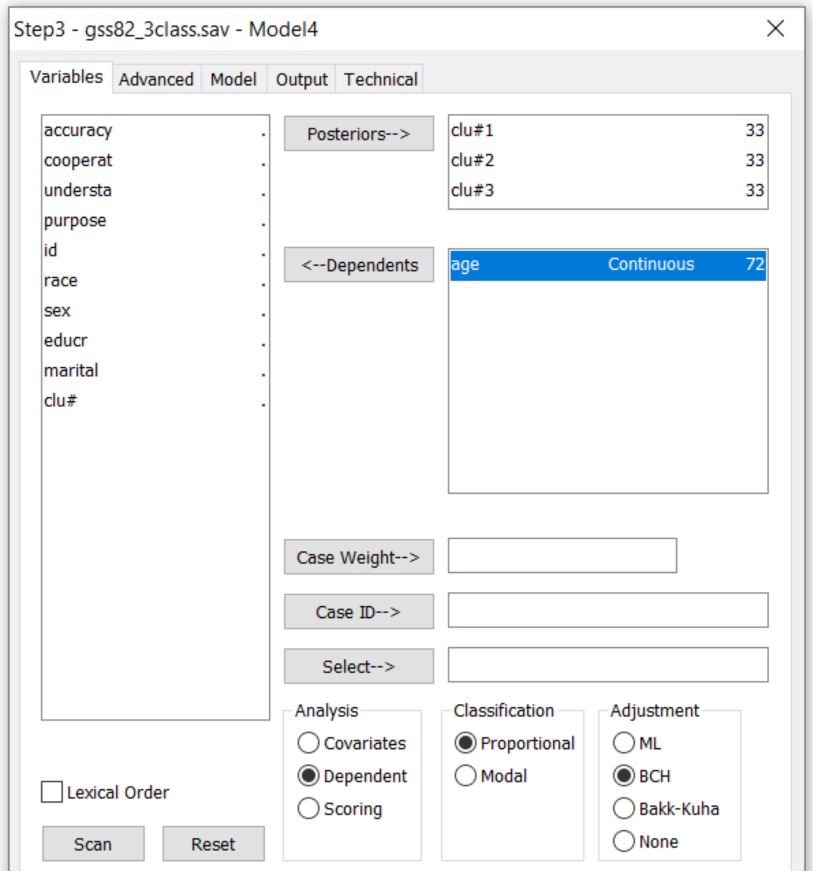
Note that I am using the options:

- Covariates
- Proportional
- ML

GSS82.sav data: bias-adjusted 3-step method

- Instead you may use the dependent (=bivariate analysis) option, where for age using BCH would be preferred.





GSS82.sav data: bias-adjusted 3-step method

- Relevant output for Step3-Covariates: same as for 1-step approach
  - Parameters, Wald tests, and Paired comparisons
  - Profile
  - ProbMeans
  - EstimatedValues
- Most relevant output for Step3-Dependent
  - Wald tests and Paired comparisons
  - Profile