Latent class models: basic formulation and extension for covariates

Silvia Bacci

silvia.bacci@unipg.it
Dipartimento di Economia - Università degli Studi di Perugia (IT)

Outline

Background

- Many real problems in the data analysis may be treated through the latent variable models
- A latent variable is a variable which is not directly observable and is assumed to affect the response variables (i.e., manifest variables)
- Examples: customer satisfaction, quality of life, mathematics ability, ...
- Example 1: the habit to the use of marijuana affects the probability of observing a given value of consumption along the time
- Example 2: the satisfaction for the health services affects the probability of answering in a certain way to a satisfaction questionnaire
- Example 3: beyond the effect of observed covariates, there remains a part of unexplained heterogeneity of the probability of a preterm delivery, which is due to unobservable woman/pregnancy characteristics

Use of latent variables

- Latent variables are typically included in a statistical model with different aims:
  - accounting for measurement errors: the latent variables represent the "true" outcomes and the manifest variables represent their "disturbed" versions (Example 1, Example 2)
  - representing the effect of unobservable covariates/factors and then accounting for the unobserved heterogeneity between subjects: latent variables are used to represent the effect of these unobservable factors (Example 3)

Classification of latent variable models

- The latent variable models are typically classified according to
  - nature of the response variables: discrete or continuous
  - nature of the latent variables: discrete or continuous
  - inclusion or not of individual covariates
- We focus on
  - Basic Latent Class (LC) models (Example 1, Example 2):

    models for categorical response variables based on a discrete latent variable, the levels of which correspond to latent classes in the population
  - LC models with covariates (Example 3):

    extension of basic LC models with observable covariates affecting the probability to belong to the latent classes
  - Generalized Linear Mixed Models (GLMMs) with discrete random-effects (Example 3):

    extension of the class of Generalized linear models (GLM) for (continuous or) categorical responses which account for unobserved heterogeneity, beyond the effect of observable covariates

The standard latent class model

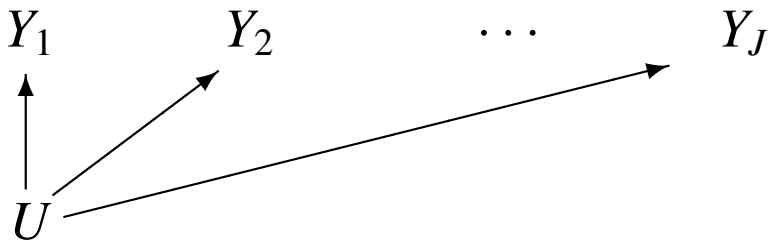
- LC models are based on the assumption that the population is composed by unobservable subgroups (or latent classes) of individuals, sharing common characteristics related to a latent variable of interest (e.g., the satisfaction for health services, the tendency to have a preterm delivery)
- Aim of LC models: clustering individuals in homogenous latent classes on the basis of observed responses to categorical variables (or items)
- In their standard version, LC models rule out covariates

Basic notation

- Y_{ij}: categorical response variable for subject i to item j, with $i = 1, \ldots, n$ and $j = 1, \ldots, J_i$
- y: value observed for Y_{ij}, with $y = 0, 1, \ldots, r_j - 1$
- $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$: vector of items for subject i
- U_i: discrete latent variable for subject i
- ξ_u: value assumed by U_i (support point), with $u = 1, \ldots, k$

Main assumptions

Basic LC:



- Local independence assumption: given latent class $U_i = u$, probability of answering Y_{ij} is independent of probability of answering Y_{il}, for $j \neq l;\ j, l = 1, \ldots, J$

Model formulation

- Manifest distribution of response vector $\boldsymbol{Y}_i$

$$p(\boldsymbol{y}) = p(\boldsymbol{Y}_i = \boldsymbol{y}) = \sum_{u=1}^{k} \pi_u \, p_u(\boldsymbol{y})$$

  - mass probability (or weight) that subject i belongs to class u ($u = 1, \ldots, k$):

$$\pi_u = p(U_i = \xi_u) = \frac{\exp(\psi_{0u})}{1 + \exp(\psi_{0u})} \quad \text{u.c.} \sum_u \pi_u = 1; \; \pi_u > 0$$

  - conditional probability of answering $\boldsymbol{y}$, given the latent class u (local independence assumption), $y = 0, \ldots, r_j - 1; \; u = 1, \ldots, k$:

$$p_u(\boldsymbol{y}) = p(\boldsymbol{Y}_i = \boldsymbol{y} | U_i = \xi_u) = \prod_{j=1}^{J_i} p(Y_{ij} = y | U_i = \xi_u)$$

Parameter estimation

- The number of free model parameters is equal to

$$\#\text{par} = \underbrace{(k-1)}_{\pi_u} + \underbrace{kJ(r_j-1)}_{p_u(\boldsymbol{y})}$$

- The LC model is estimated by the maximization of the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{Y}_i = \boldsymbol{y}) = \sum_{i=1}^{n} \log \sum_{u=1}^{k} \pi_u p_u(\boldsymbol{y}_i)$$

  where $\boldsymbol{\theta}$ is the vector of free model parameters

- The log-likelihood $\ell(\boldsymbol{\theta})$ may be efficiently maximized through an Expectation-Maximization (EM) algorithm

EM algorithm

- The EM algorithm treats the estimation of LC model parameters as an estimation problem in presence of missing data, being the belonging of individuals to the corresponding latent class a missing information
- The EM algorithm is based on the complete log-likelihood

$$\ell^*(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{u=1}^{k} \lambda_{ui} \left[\log \pi_u + \log p_u(\boldsymbol{y}_i) \right],$$

where λ_{ui} is an indicator equal to 1 if subject i belongs to latent class u and to 0 otherwise

EM algorithm

- The EM algorithm alternates two steps until convergence in $\ell(\boldsymbol{\theta})$

  E-step It consists of computing the expected value of $\ell^*(\boldsymbol{\theta})$, which is equivalent to computing the posterior probabilities λ_{ui}, under current estimates of model parameters:

$$\hat{\lambda}_{ui} = p(U_i = \xi_u | \boldsymbol{Y}_i = \boldsymbol{y}) = \frac{\pi_u \, p_u(\boldsymbol{y}_i)}{\sum_{c=1}^{k} \pi_c \, p_c(\boldsymbol{y}_i)}$$

  M-step It consists in updating the model parameters by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$, obtained by substituting the values of $\hat{\lambda}_{ui}$

Warnings . . .

- The maximization process requires a set of starting values to initialize the EM algorithm
- The log-likelihood of a LC model is usually characterized by local maximum points: we suggest to try several randomly chosen (e.g., from an $U(0,\ 1)$) initializations of the EM algorithm to detect the global maximum solution
- The number k of latent classes is not a model parameter, but it has to be a priori fixed: we suggest to select it on the basis of some information criteria, such as BIC index

$$BIC = -2\hat{\ell} + \#\text{par}\log(n)$$

In practice, we fit the model for increasing values of BIC until does not start to increase and then we take the previous value of k as optimal one

Use of standard LC model

- After the parameter estimation, each individual i may be allocated to one of the k latent classes on the basis of the highest estimated posterior probability

$$\hat{\lambda}_{ui} = \frac{\hat{\pi}_u \, \hat{p}_u(\boldsymbol{y}_i)}{\sum_{c=1}^{k} \hat{\pi}_c \, \hat{p}_c(\boldsymbol{y}_i)}$$

- Note that the allocation to latent classes depends only on the observed configuration of item responses

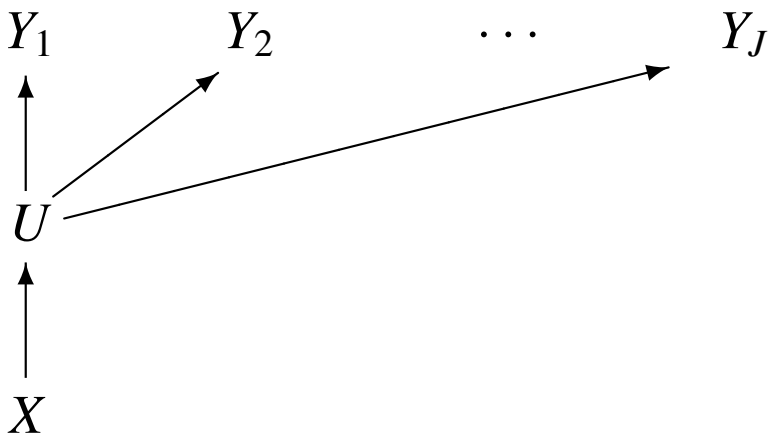
LC models with covariates

- Covariates may be included in an LC model in two different ways

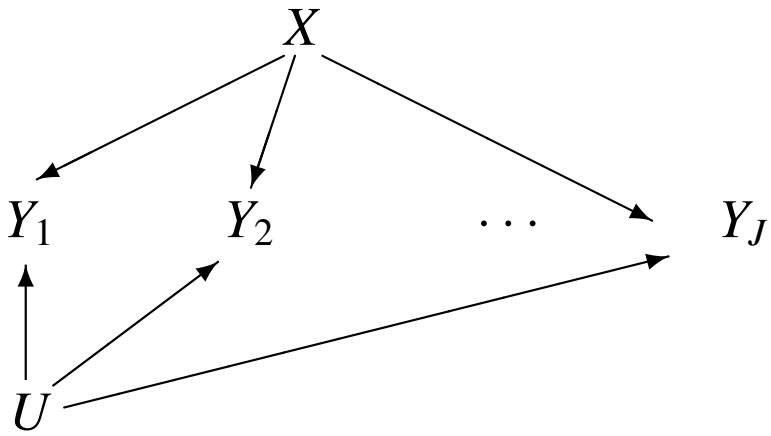
  Case 1 on the mass probabilities, that is, on the model for the distribution of the latent variable U_i, via a multinomial logit model (or global logit in case of ordered classes)

  Case 2 on the observed item responses, via a logit type parameterization

Case 1



Case 2



- LC models based on the two extensions have a different interpretation

  Case 1 the main interest is on the discrete latent variable which is measured through the observable response variables and on how this latent variable depends on the covariates; covariates directly affect the probability of belonging to a given class

  Case 2 the discrete latent variable is used to account for the unobserved heterogeneity and then the model may be seen as a discrete version of a logit type model with random intercept; covariates directly affect the probability of observing a given pattern of item responses

LC models with covariates - Case 1

- Mass probabilities π_u now are subject-specific and depend on the vector of M-covariates $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{Mi})$, through a multinomial logit model

$$\pi_{ui}(\boldsymbol{x}) = p(U_i = \xi_u | \boldsymbol{X}_i = \boldsymbol{x}) = \frac{\exp(\psi_{0u} + \boldsymbol{x}_i' \boldsymbol{\psi}_u)}{1 + \exp(\psi_{0u} + \boldsymbol{x}_i' \boldsymbol{\psi}_u)}$$

$$\frac{\exp(\psi_{0u} + \sum_{m=1}^{M} \psi_{mu} x_{mi})}{1 + \exp(\psi_{0u} + \sum_{m=1}^{M} \psi_{mu} x_{mi})} \quad u = 2, \ldots, k$$

or, equivalently,

$$\log \frac{\pi_{ui}(\boldsymbol{x})}{\pi_{1i}(\boldsymbol{x})} = \psi_{0u} + \boldsymbol{x}_i' \boldsymbol{\psi}_u = \psi_{0u} + \sum_{m=1}^{M} \psi_{mu} x_{mi} \quad u = 2, \ldots, k$$

- Regression coefficient ψ_{mu} means the effect of an increase of an unit of m-th covariate on the logit of belonging to class u with respect to the reference class (e.g., class 1)

Model formulation and estimation - Case 1

- Manifest distribution of response vector $\boldsymbol{Y}_i$

$$p(\boldsymbol{Y}_i = \boldsymbol{y} | \boldsymbol{X}_i = \boldsymbol{x}) = \sum_{u=1}^{k} \pi_{ui}(\boldsymbol{x}) \; p_u(\boldsymbol{y})$$

- The number of free model parameters is equal to

$$\#\text{par} = (k-1)(M+1) + kJ(r_j - 1)$$

- The model is estimated by the maximization of the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{u=1}^{k} \pi_{ui}(\boldsymbol{x}) p_u(\boldsymbol{y}_i)$$

- The log-likelihood $\ell(\boldsymbol{\theta})$ may be efficiently maximized through an EM algorithm, where the M-step is modified relying on standard algorithms for the maximization of the likelihood of a multinomial logit model

LC models with covariates - Case 2

- Conditional probability $p_u(\boldsymbol{y})$ depends also on the vector of T-covariates $\boldsymbol{W}_{ij} = (W_{1ij}, \ldots, W_{Tij})$, through a logit or probit type parameterization

$$p_{uw}(\boldsymbol{y}) = p(\boldsymbol{Y}_i = \boldsymbol{y} | U_i = \xi_u, \boldsymbol{W}_{ij} = \boldsymbol{w}) = \prod_{j=1}^{J_i} p(Y_{ij} = y | U_i = \xi_u, \boldsymbol{W}_{ij} = \boldsymbol{w})$$

- A logit type model with random intercept is usually adopted for $p(Y_{ij} = y | U_i = \xi_u, \boldsymbol{W}_{ij} = \boldsymbol{w})$. For instance, in case of binary responses we have

$$\log \frac{p(Y_{ij} = 1 | U_i = \xi_u, \boldsymbol{W}_{ij} = \boldsymbol{w})}{p(Y_{ij} = 0 | U_i = \xi_u, \boldsymbol{W}_{ij} = \boldsymbol{w})} = \xi_u + \boldsymbol{w}'_{ij}\boldsymbol{\beta}$$

- Regression coefficients in $\boldsymbol{\beta}$ mean the effect of an increase of an unit of the t-th covariate ($t = 1, \ldots, T$) on the logit of answering $y = 1$ rather than $y = 0$ to item j

Model formulation and estimation - Case 2

- Manifest distribution of response vector $\boldsymbol{Y}_i$

$$p(\boldsymbol{Y}_i = \boldsymbol{y} | \boldsymbol{W}_{ij} = \boldsymbol{w}) = \sum_{u=1}^{k} \pi_u \, p_{uw}(\boldsymbol{y})$$

- The number of free model parameters is equal to

$$\#\text{par} = (k-1) + kJ(r_j - 1) + TJ$$

- The model is estimated by the maximization of the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{u=1}^{k} \pi_u \, p_{uw}(\boldsymbol{y}_i)$$

- The log-likelihood $\ell(\boldsymbol{\theta})$ may be efficiently maximized through an EM algorithm, where the M-step is modified relying on standard algorithms for the maximization of a weighted likelihood of a logit type model

Applications

- Example 1: analysis of marijuana consumption
  - We adopt a basic LC model without covariates
  - We have five response variables ($J = 5$) with 3 ordered levels ($y = 0, 1, 2$ for all j)
- Example 2: analysis of the satisfaction for the health services
  - We adopt an LC model with covariates on the mass probabilities (Case 1)
  - We have one response variable ($J = 1$) with 3 ordered levels ($y = 0, 1, 2$)
- Example 3: analysis of determinants of preterm deliveries
  - We adopt an LC model with covariates both on the mass probabilities and on the response variable (Case 1 + Case 2)
  - We account for the multilevel structure of data, consisting in babies/pregnancies nested in women
  - We have one response variable with 2 levels ($y = 0, 1$)

Data source and study population

- Data is based on 5 annual waves of the National Youth Survey concerning the consumption of marijuana among individuals who were aged 13 years in 1976
- Sample size: 237 individuals followed up for 5 years
- Ordinal response variable for each wave measuring the marijuana consumption: 0 = never in the past year, 1 = no more once in a month in the past year, 2 = more than once a month in the past year
- Aim of the study: detecting homogenous classes of individuals with respect to the level of marijuana consumption

Model selection and latent class weights

Table : Selection of the number of latent classes: maximum log-likelihood value, number of parameters, and BIC index

k	$\hat{\ell}$	# par	BIC
2	-695.2882	21	1505.406
3	-658.2381	32	1491.454
4	-652.8618	43	1540.850

On the basis of BIC index we select $k = 3$ latent classes

Table : Estimates of latent class weights $\pi_u = p(U_i = \xi_u)$ for model with $k = 3$

u	$\hat{\pi}_u$
1	0.618
2	0.215
3	0.167

Interpretation of latent classes

Table : *Estimates of conditional response probabilities* $p(Y_{ij} = y|U_i = \xi_u)$

		Observed response		
j	u	$y = 0$	$y = 1$	$y = 2$
1	1	0.9732	0.0199	0.0068
1	2	0.9401	0.0599	0.0000
1	3	0.6959	0.2030	0.1011
2	1	0.9913	0.0087	0.0000
2	2	0.6761	0.2523	0.0716
2	3	0.3873	0.3256	0.2870
3	1	1.0000	0.0000	0.0000
3	2	0.2838	0.6025	0.1137
3	3	0.1522	0.2610	0.5868
4	1	0.9414	0.0375	0.0212
4	2	0.3547	0.6453	0.0000
4	3	0.0000	0.0672	0.9328
5	1	0.8244	0.1251	0.0504
5	2	0.3171	0.5867	0.0962
5	3	0.0265	0.0959	0.8775

- The three classes correspond to an increasing tendency to marijuana consumption
- The biggest class (class 1) collects individuals with the smallest tendency to marijuana consumption
- The smallest class (class 3) collects individuals with the highest tendency to marijuana consumption
- We observe a general tendency to increase the marijuana consumption along the time, for all the latent classes
- Note that individuals belong to the same latent class during all the time period, but the conditional distribution of responses is allowed to change

Data source and study population

- Data comes from the survey on Italian families about Health conditions and use of health services - 2013 administered by the Italian National Institute of Statistics (ISTAT) in the period 2012-2013
- Main topics of the survey: health conditions (self-evaluated health status, presence of chronic diseases, ...), presence of disabilities, life styles (smoking habits, physical activity, ...), prevention, use of health services, use of unusual drugs or therapies
- We accounted for individuals who received specialized medical examinations; individuals older than at least 18 years
- We adopt $k = 3$ latent classes
- Sample size: 50,871 individuals
- Aim of the study: detecting homogenous classes of individuals with respect to the level of satisfaction for the received service and determinants of belonging to the latent classes

Variables of interest

Table : *Response variable*

Satisfaction level	insufficient	0.056
	sufficient/good	0.522
	excellent	0.422

Table : *Covariates*

Variable description	Category	Proportion/Average
Gender	male	0.41
	female	0.59
Age	average years	55.36
Education level	Compulsory diploma	0.53
	High school diploma	0.35
	Degree or above	0.12
Geographic area	South	0.35
	Centre	0.19
	North	0.46
Physical state index	discomfortable condition (≤ 41)	0.27
	comfortable condition > 41	0.73
Psychological state index	discomfortable condition (≤ 41)	0.25
	comfortable condition > 41	0.75
Economic status	discomfortable	0.37
	comfortable	0.63
Smoking habit	No	0.82
	Yes	0.18
BMI	average value	25.27

Estimates of conditional probabilities and average weights

Table : *Estimates of* $p_u(\boldsymbol{y}) = p(\boldsymbol{Y}_i = \boldsymbol{y} | U_i = \xi_u)$ *and* $\bar{\pi}_{ui}(\boldsymbol{x}) = p(U_i = \xi_u | \boldsymbol{X}_i = \boldsymbol{x})$

	Latent class		
y	$u = 1$	$u = 2$	$u = 3$
0	0.231	0.001	0.000
1	0.709	0.730	0.105
2	0.060	0.270	0.895
$\hat{\bar{\pi}}_{ui}(\boldsymbol{x})$	0.239	0.437	0.324

Estimates of regression coefficients ψ

	$\hat{\psi}_{m2}$	$\hat{\psi}_{m3}$	$\hat{\text{se}}_{m2}$	$\hat{\text{se}}_{m3}$
intercept ψ_{0u}	0.64	-0.52	0.34	0.29
female $(m = 1)$	-0.11	0.29*	0.06	0.07
age $(m = 2)$	0.01	0.00	0.00	0.00
ed_high_school $(m = 3)$	0.19*	0.18*	0.08	0.05
ed_degree $(m = 4)$	0.33*	0.39*	0.11	0.08
area_north $(m = 5)$	0.05*	0.64*	0.02	0.10
area_centre $(m = 6)$	-0.04	0.30*	0.08	0.08
physical_disc $(m = 7)$	-0.16*	-0.33*	0.07	0.06
psychol_disc $(m = 8)$	-0.48*	-0.45*	0.10	0.05
economic_good $(m = 9)$	0.44*	0.51*	0.09	0.05
foreign $(m = 10)$	-0.43*	-0.39*	0.19	0.09
smoke $(m = 11)$	-0.30*	-0.00	0.08	0.07
bmi $(m = 12)$	-0.02*	-0.00	0.01	0.01

*: statistically significant at 5% level

Estimates of average latent class weights

Table : *Average estimates of* $p(U_i = \xi_u | \boldsymbol{X}_i = \boldsymbol{x})$

Pattern	Latent class $u=1$	$u=2$	$u=3$
Female, Italian, Degree	0.17	0.42	0.41
Female, Foreign, Degree	0.26	0.35	0.39
Female, Italian, Compulsory educ.	0.26	0.42	0.32
Male, Italian, Compulsory educ.	0.26	0.47	0.26
Male, Italian, North	0.20	0.45	0.35
Male, Italian, South	0.28	0.51	0.21
Female, Discomf. physical status	0.24	0.45	0.31
Female, Discomf. psychological status	0.29	0.36	0.34
Female, Discomf. psycho-physical status	0.35	0.39	0.25
Female, Comfortable psycho-physical status	0.20	0.42	0.39
Male, high BMI	0.27	0.44	0.29
$\bar{\hat{\pi}}_{ui}(\boldsymbol{x})$	0.24	0.44	0.32

Conclusions

- The proposed LC model with covariates allows us to classify an individual in one of the three latent classes, on the basis of some individual characteristics
- Each latent class is homogenous with respect to the satisfaction level for the specific type of received health service
- Individuals in the same latent class detect a set of shared needs
- Future developments: we intend to formulate an appropriate LC model that accounts for
  - all the types of health services (day hospital, surgery, diagnostics checks)
  - the multilevel structure of data: individuals within families
  - the presence of informative missing data

Data source and study population

- Data comes from the Standard Certificate of Live Birth (SCLB) administered in the Umbria Region since 2005 until 2013
- SCLB is a questionnaire compulsorily filled in all Italian birth centres within ten days after the delivery by one of the attendants the birth (e.g., doctor, midwife)
- SCLB collects information both on infants and their parents (mainly mothers), other than on the course of pregnancy
- We accounted for: women that delivered for the first time during years 2005-2013 and that delivered at least twice in this time interval; only singleton births; infants with a birthweight of at least 500 grams and a gestational age between 24 and 42 (included) weeks
- Sample size: 12,157 babies/deliveries within 5,865 women
- Aim of the study: detecting determinants of preterm births and clustering pregnant women on the basis of the probability of delivering preterm

Variables of interest

Table : *Response variable*

Variable description	Category	Proportion
Preterm birth (y_{ij})	$y = 0$: normoterm birth	0.957
	$y = 1$: preterm birth ($<$ 37 weeks)	0.043

Table : *Characteristics developed during the pregnancy*

Variable description	Category	Proportion
First medical check	within the first pregnancy quarter	0.981
	after the first pregnancy quarter	0.019
Course of pregnancy	physiologic	0.955
	pathologic	0.045
Baby's gender	female	0.479
	male	0.521
Birthweight	low birthweight ($<$ 2500 grams)	0.037
	normal weight	0.963

Table : *Characteristics known at the beginning of the pregnancy*

Variable description	Category	Proportion
Woman's age	$<$ 20 years	0.013
	between 20 and 35 years	0.750
	$>$ 35 years	0.237
Woman's citizenship	Italian or other western citizenship	0.814
	foreign citizenship	0.186
Woman's job condition	working woman	0.672
	not working woman	0.328
Partner's job condition	working partner	0.970
	not working partner	0.030
Woman's educational level	middle school or less	0.183
	high school	0.495
	degree or above	0.322
Previous miscarriages	none	0.835
	1 miscarriage	0.133
	$\geq$ 2 miscarriages	0.031
Voluntary interruptions	none	0.962
	$\geq$ 1	0.038
Type of conception	natural conception	0.991
	assisted fertilisation	0.009
Siblings	first born	0.518
	at least second born	0.482

Model specification

- Data has a hierarchical structure with babies/pregnancies nested within women
- We denote by j a generic baby/pregnancy (instead of the item) and by i a generic woman
- We assume that the probability of a preterm delivery is
  - directly affected by some observable baby/pregnancy characteristics ($\boldsymbol{W}_{ij}$), which develop during the pregnancy, and by some unobservable (latent) characteristics of the woman
  - indirectly affected by some characteristics of the woman and her childbearing history ($\boldsymbol{X}_i$), which are known at the beginning of the pregnancy and directly affect the probability of belonging to a given latent class, which is homogeneous in terms of risk of preterm delivery

LC model with covariates (Case 1 + Case 2)

- The resulting model is a logit model with a discrete random intercept:

$$\log \frac{p(Y_{ij}=1|U_i=\xi_u, \boldsymbol{W}_{ij}=\boldsymbol{w})}{p(Y_{ij}=0|U_i=\xi_u, \boldsymbol{W}_{ij}=\boldsymbol{w})} = u_i + \boldsymbol{w}'_{ij}\boldsymbol{\beta}$$

- where value assumed by random intercept u_i is obtained by:

$$\log \frac{\pi_{ui}(\boldsymbol{x})}{\pi_{1i}(\boldsymbol{x})} = \log \frac{p(U_i=\xi_u|\boldsymbol{X}_i=\boldsymbol{x})}{p(U_i=\xi_1|\boldsymbol{X}_i=\boldsymbol{x})} = \psi_{0u} + \boldsymbol{x}'_i\boldsymbol{\psi}_u \quad u=2,\ldots,k$$

Selection of the number k of latent classes

- Estimated log-likelihood ($\hat{\ell}$), number of parameters (# par), and BIC values, for $k = 1, 2, 3, 4$

	$k = 1$	$k = 2$	$k = 3$	$k = 4$
$\hat{\ell}$	-1,532.362	-1,511.625	-1,508.457	-1,508.344
# par	18	20	22	24
BIC	3,222.644	3,198.718	3,209.929	3,227.248

- After having chosen the value of k, we select the set of statistically significant covariates by using a backward selection strategy

Estimates of support points and average weights

Table : *Estimates of* ξ_u *and* $\bar{\pi}_{ui}(\boldsymbol{x})$

	$u=1$	$u=2$
$\hat{\xi}_u$	-5.537	-2.395
$\hat{\bar{\pi}}_{ui}(\boldsymbol{x})$	0.796	0.204

- Latent class 1 detects women with a smaller risk of preterm delivery with respect to women belonging to latent class 2

Estimates of regression coefficients β

	$\hat{\beta}_t$	$\hat{\text{se}}_t$	t-value	p-value	OR
`course` $(t = 1)$	2.053	0.233	8.803	0.000	7.790
`lbw` $(t = 2)$	4.581	0.493	9.287	0.000	97.570

- `course`: pathological course of pregnancy
- `lbw`: low birthweight

Estimates of regression coefficients ψ

	$\hat{\psi}_{m2}$	$\hat{\text{se}}_{m2}$	t-value	p-value	OR
intercept ψ_{02}	-1.619	–	–	–	0.198
`miscar2` $(m = 1)$	0.723	0.177	4.092	0.000	2.060
`p_nonjob` $(m = 2)$	1.392	0.481	2.893	0.004	4.023
`w_compeduc` $(m = 3)$	0.855	0.292	2.927	0.003	2.352

- `miscar2`: at least 2 miscarriages
- `p_nonjob`: non working partner
- `w_compeduc`: woman with a compulsory education level at most

Conclusions

- The proposed LC model with covariates allows us
  - to classify a woman in one of the two latent classes, on the basis of some characteristics known at the beginning of the pregnancy
  - to monitor the possibility that the risk of a preterm delivery modifies on the basis of some characteristics, which develop during the pregnancy
- Odds of belonging to class 2 with respect to class 1 (odds_{h2}) and odds of preterm delivery for class 1 ($\text{odds}_{y1,h1}$) and for class 2 ($\text{odds}_{y1,h2}$), for some specific covariate patterns

course	lbw	miscar2	p_nonjob	w_compeduc	odds_{h2}	$\text{odds}_{y1,h1}$	$\text{odds}_{y1.h2}$
0	0	0	0	0	0.198	0.004	0.091
1	1	1	1	1	3.860	2.997	69.310
1	0	0	1	1	1.873	0.031	0.710

R package MultiLCIRT

- Example 1 and Example 2 are performed through R package MultiLCIRT
- Package MultiLCIRT provides a flexible framework for the LC and Item Response Theory (IRT) analysis of dichotomous and ordinal polytomous outcomes under the assumption of multidimensionality and discreteness of latent traits. Every level of the abilities identify a latent class of subjects. The fitting algorithms are based on the EM paradigm and allow for missing responses and for different item parametrizations. The package also allows for the inclusion of individual covariates affecting the class weights
- The main function for the model estimation is est_multi_poly

Example 1

- Data structure

```
> library(LMest)
> data(data_drug)
> head(data_drug)

  V1 V2 V3 V4 V5  V6
1  1  1  1  1  1 111
2  1  1  1  1  2  18
3  1  1  1  1  3   7
4  1  1  1  2  1   6
5  1  1  1  2  2   6
6  1  1  1  2  3   1

> Y=data_drug[,1:5]-1 # matrix of item responses
> ww=data_drug[,6] # vector of weights
```

- Model estimation

```
out = est_multi_poly(S=Y,yv=ww,k=3,output=T,disp=T)
```

  - `S`: data matrix (one record for each response pattern)
  - `yv`: vector of weights of response patterns
  - `k`: number of latent classes
  - `output=T`: to return additional outputs (e.g., conditional response probabilities)
  - `disp=T`: to display the likelihood evolution step by step

- Output

```
> out$piv #  class weights
> out$Phi # conditional response probabilities
```

Example 2

- Data structure

```
> head(Y) # matrix of item responses
     [,1]
[1,]    1
[2,]    2
[3,]    2
[4,]    1
[5,]    2
[6,]    1

> head(XX)[1:8] # matrix of covariates
  SESSO ETA EDU2 EDU3 rip1 rip2 INDFIS1 INDMENT1
1     1  56    0    0    1    0       1        1
2     0  49    1    0    1    0       1        1
3     1  44    0    0    1    0       1        1
4     1  87    0    0    1    0       0        0
5     1  56    0    0    1    0       1        1
6     1  66    0    0    1    0       0        0
```

- Model estimation

```
> out = est_multi_poly(S=Y,k=3,X=XX,output=T,out_se=T,
disp=T,tol=10^-6)
```

  - `S`: data matrix (one record for each individual; if you have one record for each response pattern you must specify option `yv`)
  - `k`: number of latent classes
  - `X`: matrix of observed covariates that affects the weights
  - `output=T`: to return additional outputs (e.g., conditional response probabilities)
  - `out_se`: to return the standard errors
  - `disp=T`: to display the likelihood evolution step by step
  - `tol`: to set the tolerance level for the convergence of the algorithm measured by the relative difference between consecutive log-likelihoods

- Output

```
> out$piv # average class weights
> out$Phi # conditional response probabilities
> out$De # coefficients of covariates
> out$seDe # standard errors
# matrix of weights for every covariate pattern configur
> out$Piv
# matrix of weights for a given profile
> prof1 = (XX[,1]==1 & XX[,7]==1 & XX[,8]==1)
> Piv_prof1 = out$Piv[prof1]
> colMeans(Piv_prof1)
# classification of individuals
> class = apply(out$Piv,1,which.max)
# matrix of the posterior response probabilities for
each covariate configuration and latent class
> out$Pp
```

Example 3

- Example 3 is performed through an `R` function specific for the estimation of GLLMs with discrete random effects, named `est_lc_bin_ext.R`
- Model estimation

```
> out =  est_lc_bin_ext(y, X, Z, id, k, resp="bin")
```

  - `y`: binary or ordinal response variable
  - `X`: matrix of cluster-varying (or time-varying) covariates (e.g., characteristics of pregnancy that change from one pregnancy to another one)
  - `Z`: optional matrix of cluster-constant (or time-constant) covariates (e.g., citizenship of woman); they may also be obtained by averaging on variables in `X`
  - `id`: cluster-level units indicator
  - `k`: number of latent classes at cluster-level
  - `resp`: type of response (binary, ordinal)

Main references

- Bartolucci F., Bacci S., Gnaldi M. (2014). MultiLCIRT: An R package for multidimensional latent class item response models, Computational Statistics and Data Analysis, 71, 971?985
- Dempster, A. P., Laird, N. M., Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society, Series B, 39, 1-38
- Hagenaars , J. A., McCutcheon, A. L. (2002). Applied latent class analysis. Cambridge University Press
- Lazarsfeld, P. F., Henry, N. W. (1968). Latent Structure Analysis. Houghton Mifflin, Boston
- McCulloch, C. E., Searle, S. R., Neuhaus, J. M. (2008). Generalized, Linear, and Mixed Models. Wiley
- McLachlan, G. and Peel, D. (2000). Finite Mixture Models. Wiley