Latent Class Modelling

What is Latent Class Modelling?

Latent Class Modelling (LCM) comprises a set of techniques aimed to model situations where different subgroups (or, more generally, classes) of entities (for example organizations or individuals) are present in data and group membership is not directly observable, but has an impact on phenomena of interest. Depending on whether the observable variables are categorical or continuous, the models are labeled as Latent Class Analysis (LCA) or the Latent Profile Analysis (LPA); in both cases, differences between classes are bases on differences in means. Conversely, in Finite Mixture Models (FMM) classes affect the relationships between independent and dependent variables in a regression model. LCA is very close to factor analysis, but the underlying unobserved variables are not continuous but categorical.

Table 1. Name of different kinds of latent variable models (Oberski 2016)

	Models for distributions of observed variables		**Regression models (independent > dependent)**	
	Latent		*Latent*	
	Continuous	Discrete	Continuous	Discrete
Observed				
Continuous	Factor Analysis	Latent Profile Analysis	Random effects (RE)	Regression mixture
Discrete	Item response theory	Latent Class Analysis	Logistic RE	Logistic regression mixture

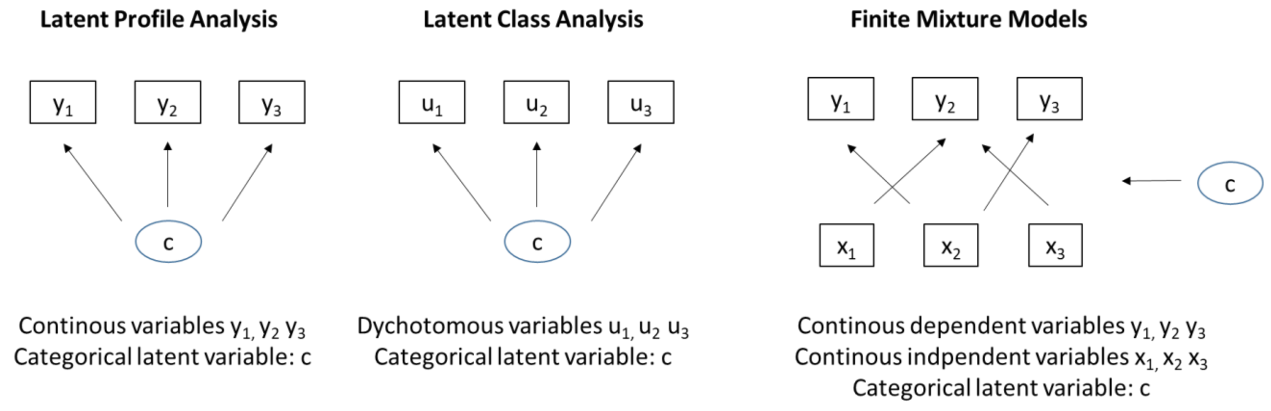
Heterogeneity

Latent Class Modelling permits to handle with unobservable heterogeneity. Heterogeneity is a crucial problem because it strongly affects the quality of any estimation, which could be imprecise or biased.

Heterogeneity mainly relies on differences between observation units (HEIs, people, organization) in terms of individual characteristics, organization, country factors or time. At the same time, heterogeneity is also an interesting source of information about the population investigated, and it is thus of a certain interest to model it.

Most social research studies, whether quantitative or qualitative, usually deal explicitly (or at least implicitly) with causal relationships between variables; in other words, how one or more circumstances or factors ('explanatory variables') cause one or more outcomes ('dependent variables'). In could happen that, in some cases, the causal relationship between two variables is due to a third, not directly observable, one (latent variable). In other words, observed correlations between variables may be due to each observed measure sharing an unobserved component (c). The value of the latent variable can then be inferred from measurable variables.

Figure 1. Relations between variables



Latent Profile Analysis

Latent Class Analysis

Finite Mixture Models

Continous variables $y_{1,}$ y_2 y_3
Categorical latent variable: c

Dychotomous variables $u_{1,}$ u_2 u_3
Categorical latent variable: c

Continous dependent variables $y_{1,}$ y_2 y_3
Continous indpendent variables $x_{1,}$ x_2 x_3
Categorical latent variable: c

Examples

As an example, you could be interested on investigate how scientists' innovativeness (not directly observable), combined with continuous or categorical exogenous covariates (ex: age, sex, citizenship, Ph.d. affiliation, etc..) affects outcomes, such as career progress, research network, citation score index, etc.. You might for example expect that age has a lower impact on citations when innovativeness is low than when it is high.

A Finite Mixture Model would classify individuals in groups by (latent) level of innovativeness and yield coefficients of covariates (for example age) by group. The model would therefore identify groups in the population that behave differently.

How it works?

Basically, latent class analysis:

- Fits the probabilities of which observations belongs to which class (probability class membership); this relationships can be also modeled based on some exogenous variables.
- Describes the relationship between the classes and the observed variables.

The estimation of Latent Class parameters used the maximum likelihood approach. Since latent class membership is not observable, the likelihood function is complex; then an expectation-maximization (EM) algorithm will be run. From a practical point of view, estimation with EM algorithm starts with a random split of the individuals considered into a defined number of classes then, based on an improvement criterion, reclassified until the best classification is found. It can be repeated using different starting values.

The model has to be run with a fixed number of classes, but fit statistics such as Maximum Likelihood, AIC and BIC can be used to select the optimal number of classes.

The model convergence and robustness request samples with high number of observations and high degree of freedom, as the number of parameters grows rapidly with the number of classes. Hence, a parsimonious specification of the model is important. Another important point is the identification of highly "discriminative" variables at the base of the creation of classes and subgroups. The latter is also relevant for classes' interpretability.

References

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg,& M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (Ch. 6; pp. 311-359). New York: Plenum.

Hagenaars J, McCutcheon A (Eds) (due Feb. 2001). *Applied Latent Class Analysis*. Cambridge University Press.

Lazarsfeld, P. F., and Henry, N. W. (1968), *Latent Structure Analysis*, Boston: Houghton Mifflin.

McCutcheon, A. C. (1987). *Latent Class Analysis*. Beverly Hills: Sage Publications.

Annex. Stata scripts

A useful software package that can be used for the estimation is GSEM (Generalized Structural Equation Model Estimation) in STATA. Scripts that can be used are structured as follows:

Latent Profile Analysis

The basic Stata command syntax for this type of model is:

gsem(y1 y2 y3 y4 <cons) (C<z1, z2, …), regress lclass(C 2)

This fits a latent class model with one categorical latent variable, C, that has two classes. Both the name of the latent variable and the number of classes is specified in the lclass() option.

This basic latent class model can be extended in many ways.

- We can specify that C has different numbers of latent classes.
- We are not limited to having observed variables that are categorical. When variables are continuous we have a latent profile analysis (LPA).
- We can include predictors of C—predictors of the probabilities of being in the different classes.
- We can include more than one categorical latent variable.

An example of LPA (Lepori et al. 2018 "The heterogeneity of European Higher Education Institutions. A typological approach"):

```
gsem (lnstaff research_intensity education_intensity citationsstaff masterorientation
herfindahlindexstudentsisced57 ssh patentintensity <- _cons) (C <- i.phdawarding i.legalstatus), regress
lclass(C 8) lcinvariant(none) nonrtolerance startvalues(randomid, draws(15) seed(5))
```

with:

covariates: lnstaff research_intensity education_intensity citationsstaff masterorientation herfindahlindexstudentsisced57 ssh patentintensity

C <- i.phdawarding i.legalstatus: classes are defined with reference to the presence or not of a research mandate (phdawarding) and to the public/private legal status (legalstatus)

lclass(C 8): this option specifies that the name of our categorical latent variable is C and that it has eight latent classes

lcinvariant(none): it allows all parameters to vary across classes

nonrtolerance: is a maximize option

startvalues(randomid, draws(15) seed(5): this option requests that starting values be computed using random class assignments. In this option, draws(5) specifies that five random draws be taken and that the one with the best log likelihood after the EM iterations be selected.

Finite Mixture Model

Here we focus on finite mixture regression models in which you can fit any regression model allowed by gsem and estimate the parameters of that model separately for each latent class. For a linear regression of y on x1 and x2, the command syntax for a two-class model is:

(y <- x1 x2), lclass(C 2)

The intercept and the coefficients on x1 and x2 will be estimated separately for the two classes.
In addition, we estimate the probability of being in each class. If we have a variable z that predicts class membership, the command syntax becomes:

(y <- x1 x2) (C <- z), lclass(C 2)

An example of FMM (Lepori Antonioli 2018, “Funding Policies in Higher Education, their Interaction with Organizational Heterogeneity and the Implications for Public Management”)

```
gsem (ln_budget <- norm_gdp i.universityhospital lnstudent research_volume_1
herfindahlindexstudentsisced57, regress) ( ln_third_party ln_student_fees <- norm_gdp i.universityhospital
lnstudent research_volume_1 herfindahlindexstudentsisced57, regress) (C <- i.dummyphd i.dummylegst),
regress lclass(C 6) lcinvariant(coef) nonrtolerance
```

with:

dependent variable: ln_budget, ln_third_party ln_student_fees

covariates: norm_gdp i.universityhospital lnstudent research_volume_1 herfindahlindexstudentsisced57

C <- i.dummyphd i.dummylegst: classes are defined with reference to the presence or not of a research mandate (dummyphd) and to the public/private legal status (dummylegst)

lcinvariant(coef): it states all coefficients constrained to be equal across classes.

lclass(C 6): this option specifies that the name of our categorical latent variable is C and that it has six latent classes

nonrtolerance: is a maximize option.