Latent Class Analysis

January, 2020

Boriana Pratt
Office of Population Research (OPR)

Main idea

Latent Class Analysis (LCA) is a statistical model in which individuals can be classified into mutually exclusive and exhaustive types, or latent classes, based on their pattern of answers on a set of (categorical) measured variables.

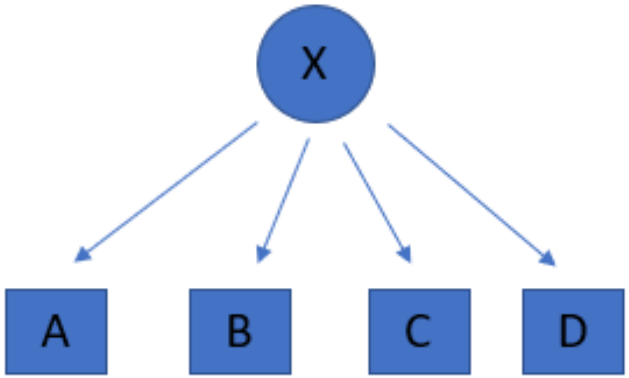
A measured variable (MV) is a variables that is directly measured whereas a latent variable (LV) is a construct that is not directly measured.

Roadmap

- Statistical Model
- Example Data
- Example model and results
- Assessing model fit
- Extension of the example
- Conclusion (and resources)

Latent Class Analysis model

Latent Class Analysis (LCA) is a way to uncover hidden groupings in data.

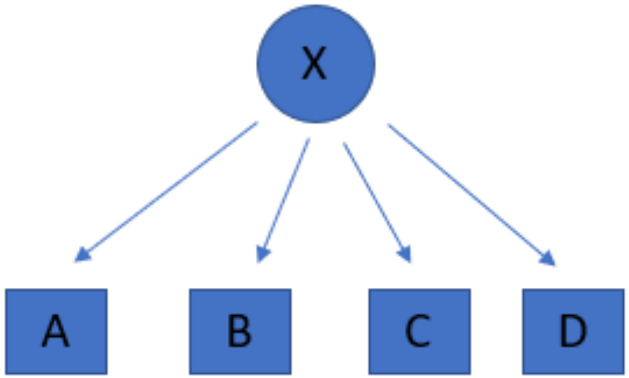


X – categorical latent variable

A, B, C, D – observed (categorical) variables

Latent Class Analysis model

Latent Class Analysis (LCA) is a way to uncover hidden groupings in data.



X – categorical latent variable

A, B, C, D – observed (categorical) variables

It is closely related to (a particular kind of) cluster analysis: used to discover groups of cases based on observed data, and, possibly, to also assign cases to groups.

Latent Class Analysis model

Main Model Assumptions: conditional independence

For two independent categorical variables - A (with J categories) and B (with K categories), the joint probability of being in category j and category k is:

$$P_{jk} = P_j^A P_k^B$$

Latent Class Analysis model

Main Model Assumptions: conditional independence

For two independent categorical variables - A (with J categories) and B (with K categories), the joint probability of being in category j and category k is:

$$P_{jk} = P_j^A P_k^B$$

If X is a latent (unobserved) variable with T classes, then (under conditional independence assumption):

$$\pi_{jkt} = \pi_t^X \pi_{jt}^{AX} \pi_{kt}^{BX}$$

where :

π_{jkt} – joint probability of being in category j, category k and class t

π_t^X – probability of being in class t

π_{jt}^{AX} – probability of being in category j (of A) conditional on being in class t (of X)

π_{kt}^{BX} - probability of being in category k (of B) conditional on being in class t (of X)

Latent Class Analysis – model estimation

Estimation is by Maximum Likelihood (ML) using the EM algorithm:

- Start with random split of people into classes.
- Reclassify based on a improvement criterion
- Reclassify until the best classification of people is found.

Latent Class Analysis – model estimation

Estimation is by Maximum Likelihood (ML) using the EM algorithm:

- Start with random split of people into classes.
- Reclassify based on a improvement criterion
- Reclassify until the best classification of people is found.

Estimation is by Maximum Likelihood (ML) using the EM algorithm:

(1) Start with (random) initial probabilities.

(2) Maximize the log-likelihood (LL) function.

(3) Update the probabilities (based on the posterior distribution).

(4) Repeat (2) and (3) until can not improve any more (LL is at max value).

Problems you might run into: finding local maximum (instead of global).

Example Data

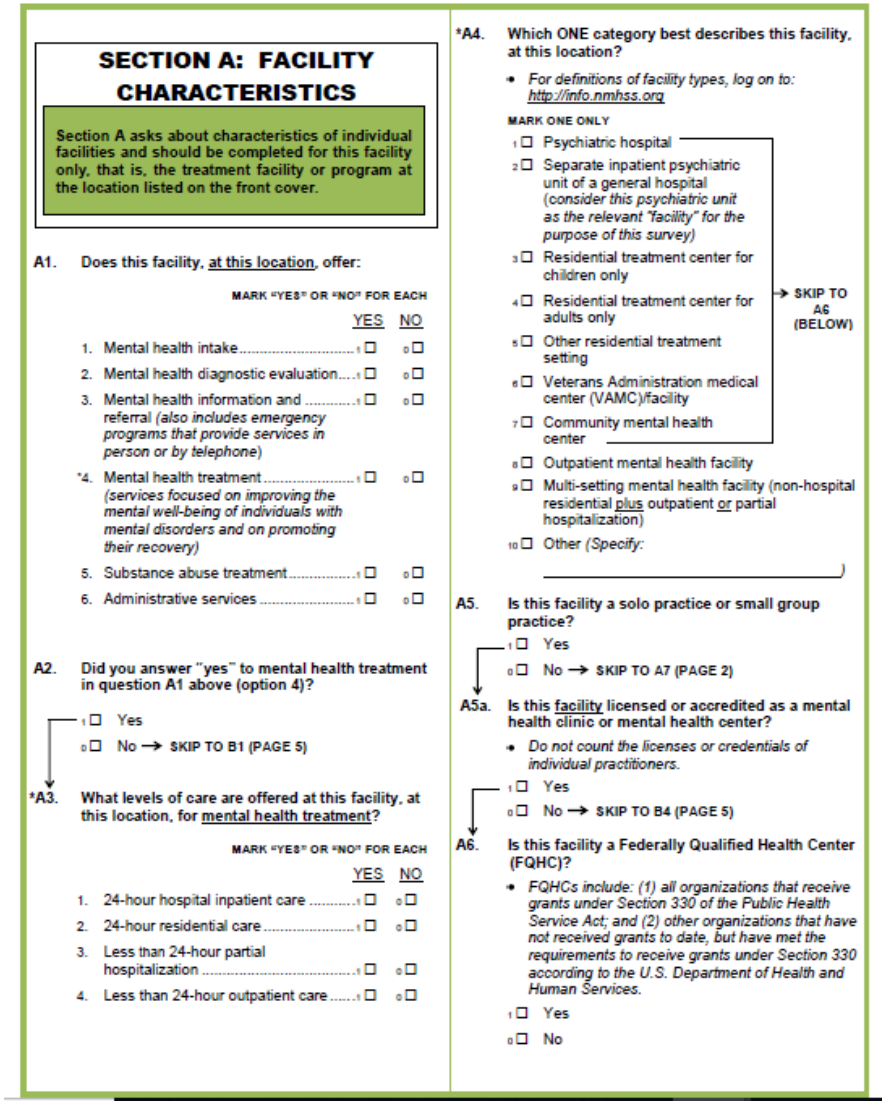
Data on Mental health services facilities

- 12,671 facilities responded to the survey in 2015
- In 50 states (the 5 territories are excluded)
- 137 variables (characteristics)

The *National Mental Health Services Survey (N-MHSS)* is an annual survey designed to collect statistical information on the numbers and characteristics of all known mental health treatment facilities within the 50 States, the District of Columbia, and the U.S. territories. In every other year, beginning in 2015, the survey also collects statistical information on the numbers and demographic characteristics of persons served in these treatment facilities as of a specified survey reference date.

Data downloaded from here: https://www.datafiles.samhsa.gov/study-series/national-mental-health-services-survey-n-mhss-nid13521

Example Data - continued



SECTION A: FACILITY CHARACTERISTICS

Section A asks about characteristics of individual facilities and should be completed for this facility only, that is, the treatment facility or program at the location listed on the front cover.

A1. Does this facility, at this location, offer:

MARK "YES" OR "NO" FOR EACH

	YES	NO
1. Mental health intake	1 ☐	0 ☐
2. Mental health diagnostic evaluation	1 ☐	0 ☐
3. Mental health information and referral *(also includes emergency programs that provide services in person or by telephone)*	1 ☐	0 ☐
*4. Mental health treatment *(services focused on improving the mental well-being of individuals with mental disorders and on promoting their recovery)*	1 ☐	0 ☐
5. Substance abuse treatment	1 ☐	0 ☐
6. Administrative services	1 ☐	0 ☐

A2. Did you answer "yes" to mental health treatment in question A1 above (option 4)?

1 ☐ Yes
0 ☐ No → SKIP TO B1 (PAGE 5)

***A3. What levels of care are offered at this facility, at this location, for mental health treatment?**

MARK "YES" OR "NO" FOR EACH

	YES	NO
1. 24-hour hospital inpatient care	1 ☐	0 ☐
2. 24-hour residential care	1 ☐	0 ☐
3. Less than 24-hour partial hospitalization	1 ☐	0 ☐
4. Less than 24-hour outpatient care	1 ☐	0 ☐

***A4. Which ONE category best describes this facility, at this location?**

- *For definitions of facility types, log on to: http://info.nmhss.org*

MARK ONE ONLY

1 ☐ Psychiatric hospital
2 ☐ Separate inpatient psychiatric unit of a general hospital *(consider this psychiatric unit as the relevant "facility" for the purpose of this survey)*
3 ☐ Residential treatment center for children only
4 ☐ Residential treatment center for adults only
5 ☐ Other residential treatment setting
6 ☐ Veterans Administration medical center (VAMC)/facility
7 ☐ Community mental health center

→ SKIP TO A6 (BELOW)

8 ☐ Outpatient mental health facility
9 ☐ Multi-setting mental health facility (non-hospital residential plus outpatient or partial hospitalization)
10 ☐ Other *(Specify:* _____ *)*

A5. Is this facility a solo practice or small group practice?

1 ☐ Yes
0 ☐ No → SKIP TO A7 (PAGE 2)

A5a. Is this facility licensed or accredited as a mental health clinic or mental health center?

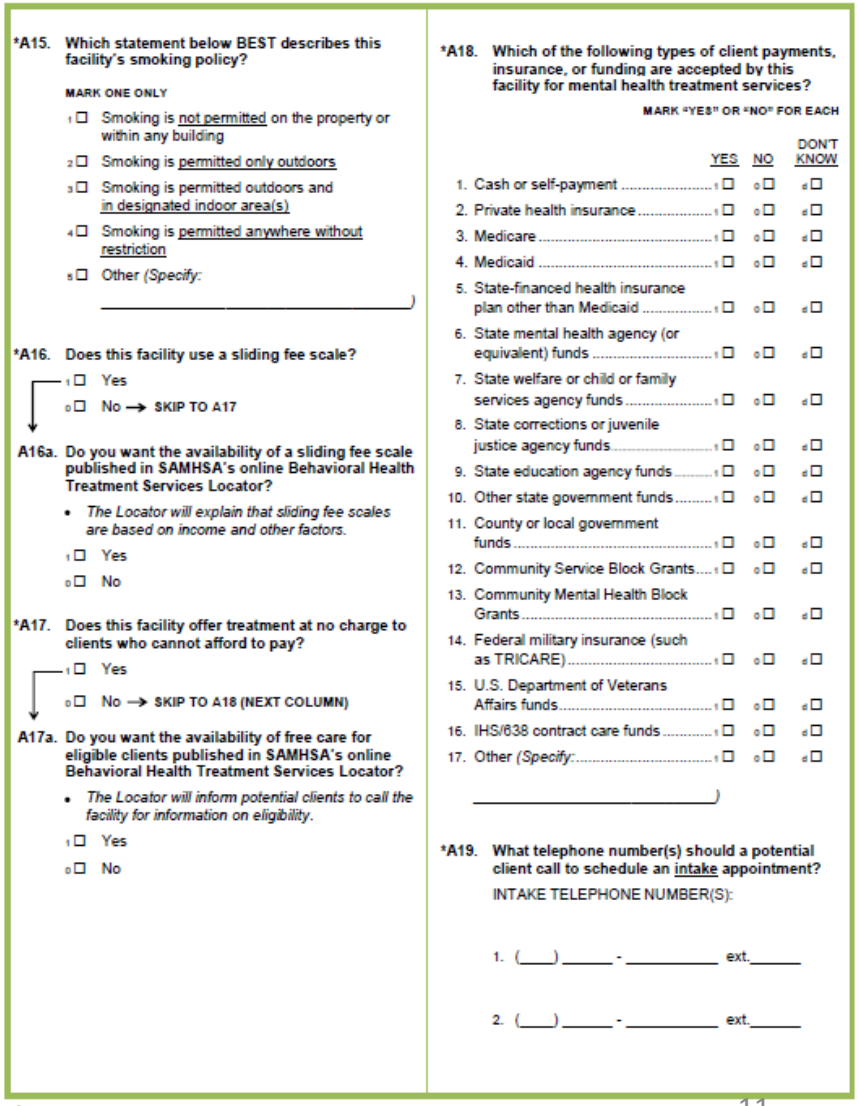
- *Do not count the licenses or credentials of individual practitioners.*

1 ☐ Yes
0 ☐ No → SKIP TO B4 (PAGE 5)

A6. Is this facility a Federally Qualified Health Center (FQHC)?

- *FQHCs include: (1) all organizations that receive grants under Section 330 of the Public Health Service Act; and (2) other organizations that have not received grants to date, but have met the requirements to receive grants under Section 330 according to the U.S. Department of Health and Human Services.*

1 ☐ Yes
0 ☐ No



***A15. Which statement below BEST describes this facility's smoking policy?**

MARK ONE ONLY

1 ☐ Smoking is not permitted on the property or within any building
2 ☐ Smoking is permitted only outdoors
3 ☐ Smoking is permitted outdoors and in designated indoor area(s)
4 ☐ Smoking is permitted anywhere without restriction
5 ☐ Other *(Specify:* _____ *)*

***A16. Does this facility use a sliding fee scale?**

1 ☐ Yes
0 ☐ No → SKIP TO A17

A16a. Do you want the availability of a sliding fee scale published in SAMHSA's online Behavioral Health Treatment Services Locator?

- *The Locator will explain that sliding fee scales are based on income and other factors.*

1 ☐ Yes
0 ☐ No

***A17. Does this facility offer treatment at no charge to clients who cannot afford to pay?**

1 ☐ Yes
0 ☐ No → SKIP TO A18 (NEXT COLUMN)

A17a. Do you want the availability of free care for eligible clients published in SAMHSA's online Behavioral Health Treatment Services Locator?

- *The Locator will inform potential clients to call the facility for information on eligibility.*

1 ☐ Yes
0 ☐ No

***A18. Which of the following types of client payments, insurance, or funding are accepted by this facility for mental health treatment services?**

MARK "YES" OR "NO" FOR EACH

	YES	NO	DON'T KNOW
1. Cash or self-payment	1 ☐	0 ☐	d ☐
2. Private health insurance	1 ☐	0 ☐	d ☐
3. Medicare	1 ☐	0 ☐	d ☐
4. Medicaid	1 ☐	0 ☐	d ☐
5. State-financed health insurance plan other than Medicaid	1 ☐	0 ☐	d ☐
6. State mental health agency (or equivalent) funds	1 ☐	0 ☐	d ☐
7. State welfare or child or family services agency funds	1 ☐	0 ☐	d ☐
8. State corrections or juvenile justice agency funds	1 ☐	0 ☐	d ☐
9. State education agency funds	1 ☐	0 ☐	d ☐
10. Other state government funds	1 ☐	0 ☐	d ☐
11. County or local government funds	1 ☐	0 ☐	d ☐
12. Community Service Block Grants	1 ☐	0 ☐	d ☐
13. Community Mental Health Block Grants	1 ☐	0 ☐	d ☐
14. Federal military insurance (such as TRICARE)	1 ☐	0 ☐	d ☐
15. U.S. Department of Veterans Affairs funds	1 ☐	0 ☐	d ☐
16. IHS/638 contract care funds	1 ☐	0 ☐	d ☐
17. Other *(Specify:* _____ *)*	1 ☐	0 ☐	d ☐

***A19. What telephone number(s) should a potential client call to schedule an intake appointment?**

INTAKE TELEPHONE NUMBER(S):

1. (____) _____ - _____ ext._____

2. (____) _____ - _____ ext._____

4

Example Data - continued

Characteristics collected:

- Types of services offered
- Ownership
- Type of setting
- Focus
- Treatment options available (types of therapy used)
- …

The survey includes both publicly and privately-operated mental health treatment facilities; includes for-profit and non-for-profit facilities in three types of settings: outpatient, inpatient and/or residential.

Facility locator: https://findtreatment.samhsa.gov/

Latent Class Analysis – example in R

Using poLCA package in R:

```
. install.packages("poLCA")
. library(poLCA)
```

Read the data in:

```
. samhsa2015 <- read.table(file="samhsa_2015F.csv", header=T, as.is=T, sep=",")
```

Using the first five characteristics collected (services offered), run a model with 2 latent classes:

```
. f1 <- as.formula(cbind(mhintake, mhdiageval, mhreferral, treatmt, adminserv)~1)

. LCA2 <- poLCA(f1, data=samhsa2015, nclass=2)
```

poLCA expects all variables to start at level 1 (dichotomous variables should be 1/2 , not 0/1!)

All five variables are dichotomous. So for latent variable with just one class there are 5 parameters to estimate, for a latent variable with two classes there will be 11 parameters to estimate, (three classes – 17 parameters to estimate) and so on.

Latent Class Analysis – example results

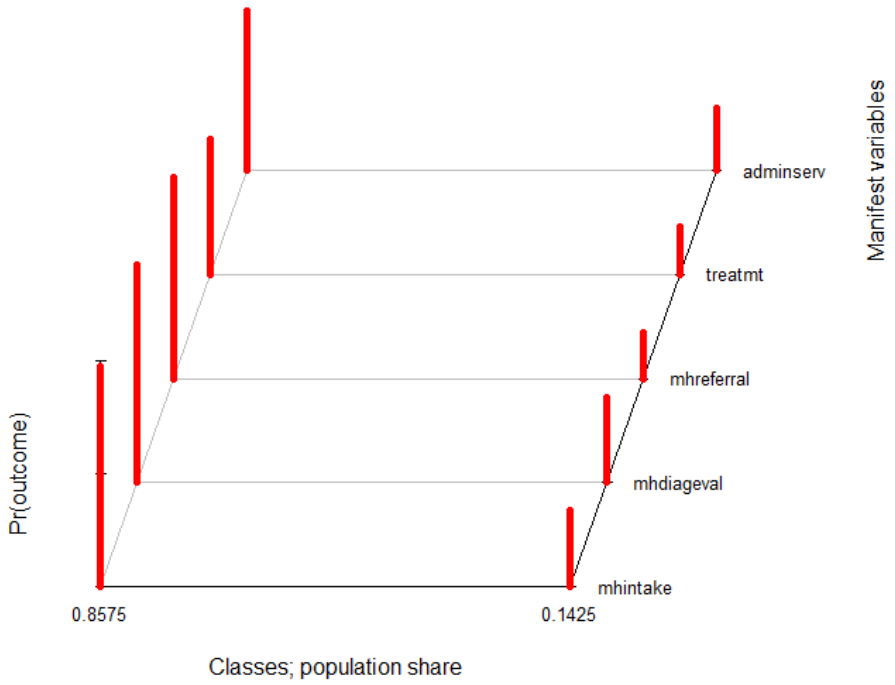
```
. LCA2 <- poLCA(f1, data=samhsa2015, nclass=2)
$`mhintake`
           Pr(1)  Pr(2)
class 1:  0.0261 0.9739
class 2:  0.6644 0.3356

$mhdiageval
           Pr(1)  Pr(2)
class 1:  0.0360 0.9640
class 2:  0.6249 0.3751

$mhreferral
           Pr(1)  Pr(2)
class 1:  0.1090 0.8910
class 2:  0.7973 0.2027

$treatmt
           Pr(1)  Pr(2)
class 1:  0.4031 0.5969
class 2:  0.7877 0.2123

$adminserv
           Pr(1)  Pr(2)
class 1:  0.2926 0.7074
class 2:  0.7240 0.2760
```

Latent Class Analysis – example results

```
Estimated class population shares
 0.8575 0.1425

Predicted class memberships (by modal posterior prob.)
 0.8751 0.1249

=========================================================
Fit for 2 latent classes:
=========================================================
number of observations: 12671
number of estimated parameters: 11
residual degrees of freedom: 20
maximum log-likelihood: -29740.22

AIC(2): 59502.44
BIC(2): 59584.36
G^2(2): 557.3724 (Likelihood ratio/deviance statistic)
X^2(2): 582.4467 (Chi-square goodness of fit)
```

Latent Class Analysis – example results

```
. plot(LCA2)
```



Latent Class Analysis – example

Let's also run models with 3 and 4 classes and look at the results, and compare:

```
. LCA3 <- poLCA(f1, data=samhsa2015, nclass=3)
```

…

ALERT: iterations finished, MAXIMUM LIKELIHOOD NOT FOUND

Latent Class Analysis – example

Let's also run models with 3 and 4 classes and look at the results, and compare:

```
. LCA3 <- poLCA(f1, data=samhsa2015, nclass=3)
```

…

ALERT: iterations finished, MAXIMUM LIKELIHOOD NOT FOUND

poLCA(formula, data, nclass = 2, maxiter = 1000, graphs = FALSE, tol = 1e-10, na.rm = TRUE, probs.start = NULL, nrep = 1, verbose = TRUE, calc.se = TRUE)

Optional parameters to use/tweak if you get the above alert:

.maxiter – The maximum number of iterations through which the estimation algorithm will cycle.

.nrep - Number of times to estimate the model, using different values of probs.start. (default is one)

Latent Class Analysis – example

Let's run models with 3 and 4 classes also and look at the results:

```
. LCA3 <- poLCA(f1, data=dt2015, nclass=3, maxiter=3000)
. LCA3 <- poLCA(f1, data=samhsa2015, nclass=3, nrep=5)
. LCA3 <- poLCA(f1, data=samhsa2015, nclass=3, maxiter=3000, nrep=5)
```

Latent Class Analysis – example

Let's run models with 3 and 4 classes also and look at the results:

```
. LCA3 <- poLCA(f1, data=dt2015, nclass=3, maxiter=3000)
. LCA3 <- poLCA(f1, data=samhsa2015, nclass=3, nrep=5)
. LCA3 <- poLCA(f1, data=samhsa2015, nclass=3, maxiter=3000, nrep=5)
```

```
. LCA4 <- poLCA(f1, data=samhsa2015, nclass=4 , maxiter=3000, nrep=5)
```

Try also:

```
. LCA5 <- poLCA(f1, data=samhsa2015, nclass=5 , maxiter=5000, nrep=10)
```

Latent Class Analysis – assessing model fit

	2 classes	3 classes	4 classes
AIC	59502.44	59119.43	58987.27
BIC	59584.36	59246.03	59158.55
G^2	557.3724	162.3576	18.19505
X^2	582.4475	180.5809	18.11012
Df	20	14	8

Let's look closer at the model with 3 classes, even though model with 4 classes does better based on AIC and BIC criteria (model with 5 classes has higher BIC than model with 4 classes, so it's definitely not an improvement).

Latent Class Analysis – interpreting results

	Class 1	Class 2	Class 3
mhintake	0.1907	0.9429	0.9916
mhdiageval	0.2447	0.9360	0.9787
mhreferral	0.1626	0.7660	1.0000
treatmt	0.2043	0.4503	0.7559
adminserv	0.2694	0.5328	0.8986

```
Estimated class population shares
 0.1046 0.5108 0.3846

Predicted class memberships (by modal posterior prob.)
 0.1048 0.5582 0.337
```

Latent Class Analysis – interpreting results

	Class 1	Class 2	Class 3
mhintake	0.1907	0.9429	0.9916
mhdiageval	0.2447	0.9360	0.9787
mhreferral	0.1626	0.7660	1.0000
treatmt	0.2043	0.4503	0.7559
adminserv	0.2694	0.5328	0.8986

```
Estimated class population shares
 0.1046 0.5108 0.3846

Predicted class memberships (by modal posterior prob.)
 0.1048 0.5582 0.337
```

The three classes roughly represent:
Class1 – facilities that don't offer any of the five services
Class2 – facilities that offer primarily mental health services
Class3 – facilities that offer all of the five services

Note: classes are unordered!

Latent Class Analysis – interpreting results

Class probabilities:

```
Predicted class memberships (by modal posterior prob.)
 0.1048 0.5582 0.337
```

Probabilities of membership in each class – these sum to one as the classes are assumed to be mutually exclusive.

Class probabilities are stored in 'predclass' element of the returned object – `LCA3$predclass`.

Latent Class Analysis – interpreting results

Class probabilities:

```
Predicted class memberships (by modal posterior prob.)
 0.1048 0.5582 0.337
```

Probabilities of membership in each class – these sum to one as the classes are assumed to be mutually exclusive.

Conditional probabilities:

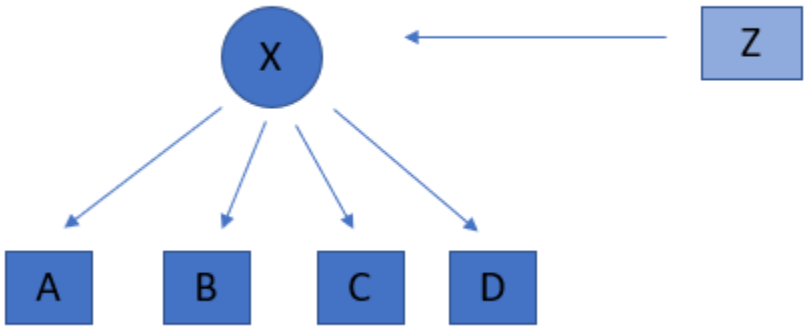
```
$mhreferral
           Pr(1)  Pr(2)
class 1:  0.8374 0.1626
class 2:  0.2340 0.7660
class 3:  0.0000 1.0000
```

Relationship between each item and each class – estimates of the probability for a particular response given membership in a certain class.

Because of conditional independence assumption within each class probabilities sum to 1.

Latent Class Analysis – extension

LCA model with covariates:



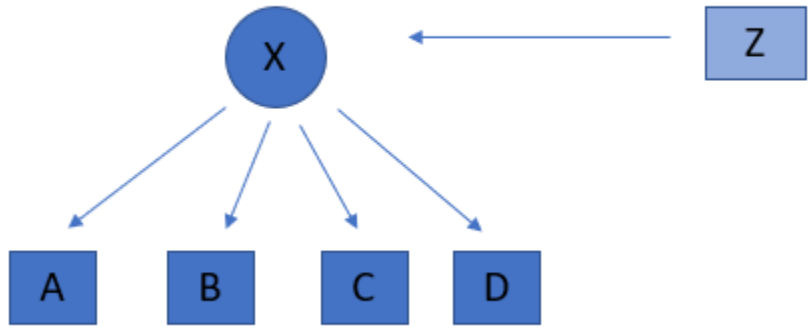
Z is observed covariate/s (categorical or continuous)

X – categorical latent variable

A, B, C, D – observed (categorical) variables

Latent Class Analysis – extension

LCA model with covariates:



Z is observed covariate/s (categorical or continuous)

X – categorical latent variable

A, B, C, D – observed (categorical) variables

```
. f2 <- as.formula(cbind(mhintake, mhdiageval, mhreferral, treatmt,
adminserv)~payasst)

. LCA3c <- poLCA(f2, data=samhsa2015, nclass=3, maxiter=3000, nrep=5)
```

Note: assigning class membership based on the base model (without covariates) and then using those classes to model relationship with a covariate gives biased estimates; it is better to include the covariate/s directly in the LCA model.

Latent Class Analysis – extension

```
Predicted class memberships (by modal posterior prob.)
 0.5798 0.0889 0.3312
=========================================================
Fit for 3 latent classes:
=========================================================
2 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -1.28826     0.14405   -8.943         0
payasst        -0.69675     0.08190   -8.507         0
=========================================================
3 / 1
            Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)     0.11932     0.19466    0.613     0.551
payasst        -0.66159     0.07089   -9.332     0.000
=========================================================
number of observations: 12278
number of estimated parameters: 19
residual degrees of freedom: 12
maximum log-likelihood: -28726.57

AIC(3): 57491.14
BIC(3): 57632.03
X^2(3): 193.2728 (Chi-square goodness of fit)
```

In this case class1 are the facilities that offer all five services, class2 are those facilities that don't offer any of the five services and class3 are the facilities that offer primarily mental health services.

Latent Class Analysis – extension

```
Predicted class memberships (by modal posterior prob.)
 0.5798 0.0889 0.3312
=========================================================
Fit for 3 latent classes:
=========================================================
2 / 1
             Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)    -1.28826     0.14405   -8.943         0
payasst        -0.69675     0.08190   -8.507         0
=========================================================
3 / 1
             Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)     0.11932     0.19466    0.613     0.551
payasst        -0.66159     0.07089   -9.332     0.000
=========================================================
number of observations: 12278
number of estimated parameters: 19
residual degrees of freedom: 12
maximum log-likelihood: -28726.57

AIC(3): 57491.14
BIC(3): 57632.03
X^2(3): 193.2728 (Chi-square goodness of fit)
```

In this case class1 are the facilities that offer all five services, class2 are those facilities that don't offer any of the five services and class3 are the facilities that offer primarily mental health services.

The results show that facilities that don't offer any of these five services are much less likely to offer pay assistance, as also are facilities that offer mostly mental health services, compared to facilities that offer all five services.

Conclusion

LCA is a method of finding subtypes within a sample through use of multivariate categorical data.

Main differences with cluster analysis are:

- LCA is model based rather than distance based grouping of data.
- Class/group membership is assigned probabilistically (rather than deterministically).

LCA could be used for dimension reduction.

Conclusion

LCA is a method of finding subtypes within a sample through use of multivariate categorical data.

Main differences with cluster analysis are:

- LCA is model based rather than distance based grouping of data.
- Class/group membership is assigned probabilistically.

LCA could be used for dimension reduction.

Steps for LCA:

- Run models with different number of classes
- Compare models fit to choose “best” model
- Include covariate(s) directly in the model

Resources:

- " Applied Latent class analysis" Allan L. McCutcheon (2002).
- "Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences" Collins, L. & Lanza, S. (2010).
- "Mixture models: Latent profile and latent class analysis" D. Obersky (2016).
- https://www.jstatsoft.org/article/view/v042i10 (to download the poLCA article)
- http://www.john-uebersax.com/stat/faq.htm
- "Categorical Data Analysis" Allan Agresti
- "Latent class analysis: The empirical study of latent types, latent variables, and latent structures. In Applied Latent Class Analysis" Goodmand (2002).