Imputation of Latent Classes after Latent Class Analysis

Hacking Stata MI toolset

Stata Conference -- August 2, 2024

Stas Kolenikov, NORC

NORC at the University of Chicago

Latent Class Analysis

NORC at the University of Chicago

Latent Class Analysis: Discrete Random Variable(s)

LCA

- Discrete latent variable(s)
  - mixture models (`fmm`) are close relatives appropriate for single outcome
- Discrete outcomes
- "Classic" quantitative social sciences: sophisticated log-linear modeling of the full contingency table
- Stata implementation: variation of `gsem`

Latent Class Analysis: Discrete Random Variable(s)

Survey of medical residents

- Outcomes: program outcomes and satisfaction
- Two classes: happy vs. unhappy
- Uhm... maybe three classes, + happy with staff but not the facility?
- Uhm... maybe four classes, + happy with technical outcomes but feel isolated?
- Downstream analyses:
    - descriptive analysis of facility variables
    - classes as predictors in regression models

Latent Class Analysis: Example

Three binary variables, $2^3 = 8$ distinct outcomes, some (secret so far) model-based probabilities in the full 3-way table:

y1	y2	y3	Prob
0	0	0	0.096
0	0	1	0.084
0	1	0	0.104
0	1	1	0.116
1	0	0	0.224
1	0	1	0.096
1	1	0	0.176
1	1	1	0.104

Latent Class Analysis: Single class solution

One-class solution / marginal probabilities:

$$\mathbb{P}[y_1 = 1] = 0.6, \mathbb{P}[y_2 = 1] = 0.5, \mathbb{P}[y_3 = 1] = 0.4$$

Three-way probabilities:

y1	y2	y3	Prob	Prob(LCA 1)
0	0	0	0.096	0.12
0	0	1	0.084	0.08
0	1	0	0.104	0.12
0	1	1	0.116	0.08
1	0	0	0.224	0.18
1	0	1	0.096	0.12
1	1	0	0.176	0.18
1	1	1	0.104	0.12

Non-centrality: 0.03702 per observation; Pearson $\chi^2(4)$ will reject accordingly.

Latent Class Analysis: Single class solution

```
. gsem (y1 y2 y3 <-) [fw=Prob*1000], lclass(C 1) logit nodvheader nolog

Generalized structural equation model                 Number of obs = 1,000
Log likelihood = -2039.1705

------------------------------------------------------------------------------
             |  Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
1.C          |   (base outcome)
------------------------------------------------------------------------------

Class: 1

------------------------------------------------------------------------------
             |  Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
y1           |
       _cons |    .4054651   .0645497     6.28   0.000     .2789499    .5319802
-------------+----------------------------------------------------------------
y2           |
       _cons |    5.68e-17   .0632456     0.00   1.000     -.123959     .123959
-------------+----------------------------------------------------------------
y3           |
       _cons |   -.4054651   .0645497    -6.28   0.000    -.5319802   -.2789499
------------------------------------------------------------------------------
```

Latent Class Analysis: Single class solution

```
. estat lcmean

Latent class marginal means                           Number of obs = 1,000

-------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.     [95% conf. interval]
-------------+-----------------------------------------------
1            |
          y1 |         .6   .0154919      .5692888    .6299448
          y2 |         .5   .0158114      .4690499    .5309501
          y3 |         .4   .0154919      .3700552    .4307112
-------------------------------------------------------------

. estat lcgof

----------------------------------------------------------------------------
Fit statistic        |      Value   Description
---------------------+------------------------------------------------------
Likelihood ratio     |
          chi2_ms(4) |     39.245   model vs. saturated
            p > chi2 |      0.000
---------------------+------------------------------------------------------
Information criteria |
                 AIC |   4084.341   Akaike's information criterion
                 BIC |   4099.064   Bayesian information criterion
----------------------------------------------------------------------------
```

Latent Class Analysis: Two class solution

Two-class solution:

$$\mathbb{P}[y_1 = 1|C = 1] = 0.4, \mathbb{P}[y_2 = 1|C = 1] = 0.6, \mathbb{P}[y_3 = 1|C = 1] = 0.6$$

$$\mathbb{P}[y_1 = 1|C = 2] = 0.8, \mathbb{P}[y_2 = 1|C = 2] = 0.4, \mathbb{P}[y_3 = 1|C = 2] = 0.2$$

$$\mathbb{P}[C = 1] = 0.5, \mathbb{P}[C = 2] = 0.5$$

Latent Class Analysis: Two class solution

```
. qui gsem (y1 y2 y3 <-) [fw=Prob*1000], lclass(C 2) logit nodvheader nolog

. estat lcmean

Latent class marginal means                         Number of obs = 1,000

--------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.     [95% conf. interval]
-------------+------------------------------------------------
1            |
          y1 |   .8000078   .1080591      .5156459     .937619
          y2 |   .3999953   .0563854      .2960959    .5137439
          y3 |   .1999922   .1080591       .062381    .4843541
-------------+------------------------------------------------
2            |
          y1 |   .4000128   .1106099      .2127046      .62196
          y2 |   .5999942   .0596549       .479579    .7094289
          y3 |   .5999872   .1106099        .37804    .7872954
--------------------------------------------------------------

. estat lcgof

----------------------------------------------------------------------------
Fit statistic        |      Value   Description
---------------------+------------------------------------------------------
Likelihood ratio     |
         chi2_ms(0)  |      0.000   model vs. saturated
           p > chi2  |          .
---------------------+------------------------------------------------------
Information criteria |
                AIC  |   4053.096   Akaike's information criterion
                BIC  |   4087.451   Bayesian information criterion
----------------------------------------------------------------------------
```

Class Predictions

What if you want to use classes in subsequent analyses?

- Summarize variables not in the model by class
- Use classes as predictors in downstream models

You... don't get them

- Classes are latent variables: you can never be sure about class membership
- Any prediction of the class labels is subject to a (prediction) error
- Subsequent use of single predictions would lead to measurement error biases

Posterior probablity predictions

You can get $\hat{p}[C|\text{pattern of } y] = \frac{\hat{p}[y|\mathrm{C}]\times\hat{p}[C]}{\sum_c \hat{p}[y|c]\times\hat{p}[c]}$:

```
. predict post_1, classposterior class(1)

. predict post_2, classposterior class(2)

. list, sep(0)
```

	y1	y2	y3	Prob	post_1	post_2
1.	0	0	0	.096	.49996262	.50003738
2.	0	0	1	.084	.14283939	.85716061
3.	0	1	0	.104	.30766144	.69233856
4.	0	1	1	.116	.0689565	.9310435
5.	1	0	0	.224	.85712399	.14287601
6.	1	0	1	.096	.49996262	.50003738
7.	1	1	0	.176	.72724308	.27275692
8.	1	1	1	.104	.30766144	.69233856

What do we do???

Is there a practical solution to the problem of class prediction after LCA?

Multiple imputation

Multiple imputation is the worst missing data method except all others that have been tried

(Winston Churchill The Statistician)

MI algorithm

1. Formulate a multivariate predictive model of the world (including outcomes)
2. For $m = 1, \ldots, M$:
    1. Obtain estimates $\hat{\beta}$ and standard errors $s(\hat{\beta})$
    2. Predict from "model + parameter uncertainty" $\hat{\beta} + z \times s(\hat{\beta})$
    3. Add noise from $y \sim f(y|\hat{\beta} + z \times s(\hat{\beta}))$
    4. Refit the model until some sort of distribution convergence
    5. Retain the last set of imputations $Y^{(m)}$
3. Estimate the model of substantive interest $\theta^{(m)} = g(Y^{(m)})$ for each m.
4. Overall estimate: $\theta_{\mathrm{MI}}^{(M)} = \frac{1}{M} \sum_{m=1}^{M} \theta^{(m)}$
5. Overall variance (Rubin's formula):

$$T = \bar{U} + (1 + 1/M)B, \; \bar{U} = \frac{1}{M} \sum_{m=1}^{M} v^{(m)} \left[\theta^{(m)}\right]$$

$$B = \frac{1}{M-1} \sum_{m=1}^{M} \left(\theta^{(m)} - \bar{\theta}\right) \left(\theta^{(m)} - \bar{\theta}\right)'$$

Worthwhile references

- Original: Rubin (1977)
- Review: after 18+ years Rubin (1996)
- Most practical: van Buuren FIMD 2nd edn (2018)
- Stata resources:
  - MI manual
  - SJ MI diagnostics: Eddings and Marchenko (2012)

Hacking Stata MI engine

NORC at the University of Chicago

MI for the people

1. Study MI manual.
2. Study `help mi_technical`.
3. Write your custom imputation code (Stas likes `mi set wide`).
4. Make sure it satisfies `mi` internal standards and expectations: `mi update`.
5. Cross fingers and run `mi estimate: whatever`.

Turns out there is more: Stata freaks out about omitted entries in `e(b)`, zero variances, and other oddities.

postlca_class_predpute

```
. mi describe

Style: wide
       last mi update 01aug2024 06:54:07, 0 seconds ago

Observations:
   Complete              0
   Incomplete        1,000  (M = 50 imputations)
   ---------------------
   Total             1,000

Variables:
   Imputed: 1; lclass(1000)

   Passive: 0

   Regular: 0

   System:  1; _mi_miss

   (there are 6 unregistered variables)
```

mi estimate

```
. mi estimate: mean y* , over(lclass)

Multiple-imputation estimates     Imputations     =         50
Mean estimation                   Number of obs   =      1,000
                                  Average RVI     =     0.5253
                                  Largest FMI     =     0.4290
                                  Complete DF     =        999
DF adjustment:   Small sample     DF:     min     =     184.84
                                          avg     =     267.13
Within VCE type:      Analytic            max     =     406.93

-------------------------------------------------------------
             |       Mean   Std. err.     [95% conf. interval]
-------------+-----------------------------------------------
c.y1@lclass  |
           1 |    .800361   .0236115      .7537783    .8469437
           2 |    .401218   .0271327      .3477782    .4546578
             |
c.y2@lclass  |
           1 |   .4014008   .0272393      .3477516      .45505
           2 |   .5977264   .0268503      .5448579    .6505949
             |
c.y3@lclass  |
           1 |   .1977063   .0220926      .1541956    .2412171
           2 |   .6006077   .0250412      .5513814     .649834
-------------------------------------------------------------
Note: Numbers of observations in e(_N) vary among imputations.
```

Summary of the missing data impact

```
. mi estimate, dftable

Multiple-imputation estimates     Imputations     =         50
Mean estimation                   Number of obs   =      1,000
                                  Average RVI     =     0.5253
                                  Largest FMI     =     0.4290
                                  Complete DF     =        999
DF adjustment:   Small sample     DF:     min     =     184.84
                                          avg     =     267.13
Within VCE type:     Analytic             max     =     406.93

-------------------------------------------------------------
             |                                      % increase
             |      Mean   Std. err.          df    std. err.
-------------+-----------------------------------------------
 c.y1@lclass |
          1  |   .800361   .0236115        184.8        31.76
          2  |   .401218   .0271327        248.2        23.96
             |
 c.y2@lclass |
          1  |  .4014008   .0272393        248.5        23.92
          2  |  .5977264   .0268503        263.6        22.60
             |
 c.y3@lclass |
          1  |  .1977063   .0220926        250.7        23.72
          2  |  .6006077   .0250412        406.9        14.46
-------------------------------------------------------------
Note: Numbers of observations in e(_N) vary among imputations.
```

`mi estimate` failures

```
. cap noi mi estimate: mean y* [fw=Prob*1000], over(lclass)
mi estimate: no observations in some imputations
    This is not allowed.  To identify offending imputations, you can use mi xeq to run the command
    on individual imputations or you can reissue the command with mi estimate, noisily

. cap noi mi estimate: reg y1 i.lclass
mi estimate: omitted terms vary
    The set of omitted variables or categories is not consistent between m=1 and m=11; this is not
    allowed.  To identify varying sets, you can use mi xeq to run the command on individual
    imputations or you can reissue the command with mi estimate, noisily

.
```

Stas' intuition:

- more of a problem when you have small multi-way cells
- less of a problem with continuous variables

More and better work

NORC at the University of Chicago

More comprehensive coverage

Stata Journal (formatted) paper

- More rigorous methodology overview
- Full documentation of the new command, its options and its use
- Simulations

https://github.com/skolenik/Stata.post.LCA.class.predimpute

Quasi-real example

```
. webuse nhanes2.dta, clear

. qui svy , subpop(if hlthstat<8) : gsem (heartatk diabetes highbp <-, logit) ///
>          (hlthstat <-, ologit) , lclass(C 2) nolog  startvalues(randomid, draws(5) seed(101))

. est tab . , keep(highbp:1.C highbp:2.C heartatk:1.C heartatk:2.C)

-----------------------------
    Variable |    Active
-------------+---------------
highbp       |
           C |
          1  |  .42449212
          2  | -.81661048
-------------+---------------
heartatk     |
           C |
          1  | -1.8749666
          2  | -6.0813072
-----------------------------
```

Quasi-real example

```
. postlca_class_predpute, lcimpute(lclass) addm(62) seed(9752)
(10,351 missing values generated)
(62 imputations added; M = 62)

Sampling weights: finalwgt
             VCE: linearized
     Single unit: missing
        Strata 1: strata
 Sampling unit 1: psu
           FPC 1: <zero>
```

Quasi-real example

```
. mi estimate , dftable : prop lclass, over(race)

Multiple-imputation estimates     Imputations     =         62
Proportion estimation             Number of obs   =     10,351
                                  Average RVI     =     0.4413
                                  Largest FMI     =     0.3509
                                  Complete DF     =      10350
DF adjustment:   Small sample     DF:      min    =     468.08
                                           avg    =     655.93
Within VCE type:     Analytic              max    =     967.06

--------------------------------------------------------------
             |                                          Normal
             |  Proportion   Std. err.           df   std. err.
-------------+------------------------------------------------
 lclass@race |
     1 White |    .2563973   .0052444         967.1       14.36
     1 Black |    .3732549   .0178635         532.7       21.74
     1 Other |    .2393548   .0373245         468.1       23.87
     2 White |    .7436027   .0052444         967.1       14.36
     2 Black |    .6267451   .0178635         532.7       21.74
     2 Other |    .7606452   .0373245         468.1       23.87
--------------------------------------------------------------
```

Questions slide

Thank you.

Stas Kolenikov
Principal Statistician

kolenikov-stas@norc.org

Research You Can Trust™