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: 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
```

	Margin	Delta-method 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

NORC at the University of Chicago

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

```
. 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

NORC at the University of Chicago

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: $\theta_{\text{MI}}^{(M)} = \frac{1}{M} \sum_{m=1}^{M} \theta^{(m)}$
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

```
. postlca_class_predpute, lcimpute(lclass) addm(50) seed(20103)
(8 missing values generated)
(50 imputations added; M = 50)

. mi describe

Style: wide
       last mi update 31jul2024 12:46:07, 0 seconds ago

Observations:
   Complete            0
   Incomplete          8  (M = 50 imputations)
   ---------------------
   Total               8

Variables:
   Imputed: 1; lclass(8)

   Passive: 0

   Regular: 0

   System:  1; _mi_miss

   (there are 6 unregistered variables)
```

mi estimate

```
. mi estimate, imp(`safelist'): mean y* [fw=Prob*1000], over(lclass)

Multiple-imputation estimates     Imputations      =         24
Mean estimation                   Number of obs    =      1,000
                                  Average RVI      =    61.6889
                                  Largest FMI      =     0.9904
                                  Complete DF      =        999
DF adjustment:   Small sample     DF:     min      =       7.85
                                          avg      =      10.03
Within VCE type:     Analytic             max      =      13.72

------------------------------------------------------------------------------
             |      Mean   Std. err.      [95% conf. interval]
-------------+----------------------------------------------------------------
c.y1@lclass  |
           1 |   .7797935   .1393365       .4657843    1.093803
           2 |    .385061   .1496405       .0573946    .7127275
             |
c.y2@lclass  |
           1 |   .4135727   .1792912       .0032355     .82391
           2 |    .626088   .1958488       .1729077    1.079268
             |
c.y3@lclass  |
           1 |   .2517946   .1074943       .0208018    .4827873
           2 |   .6084094   .1716698       .2232136    .9936051
------------------------------------------------------------------------------
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 with smaller models
- less of a problem with continuous variables

More and better work

NORC at the University of Chicago

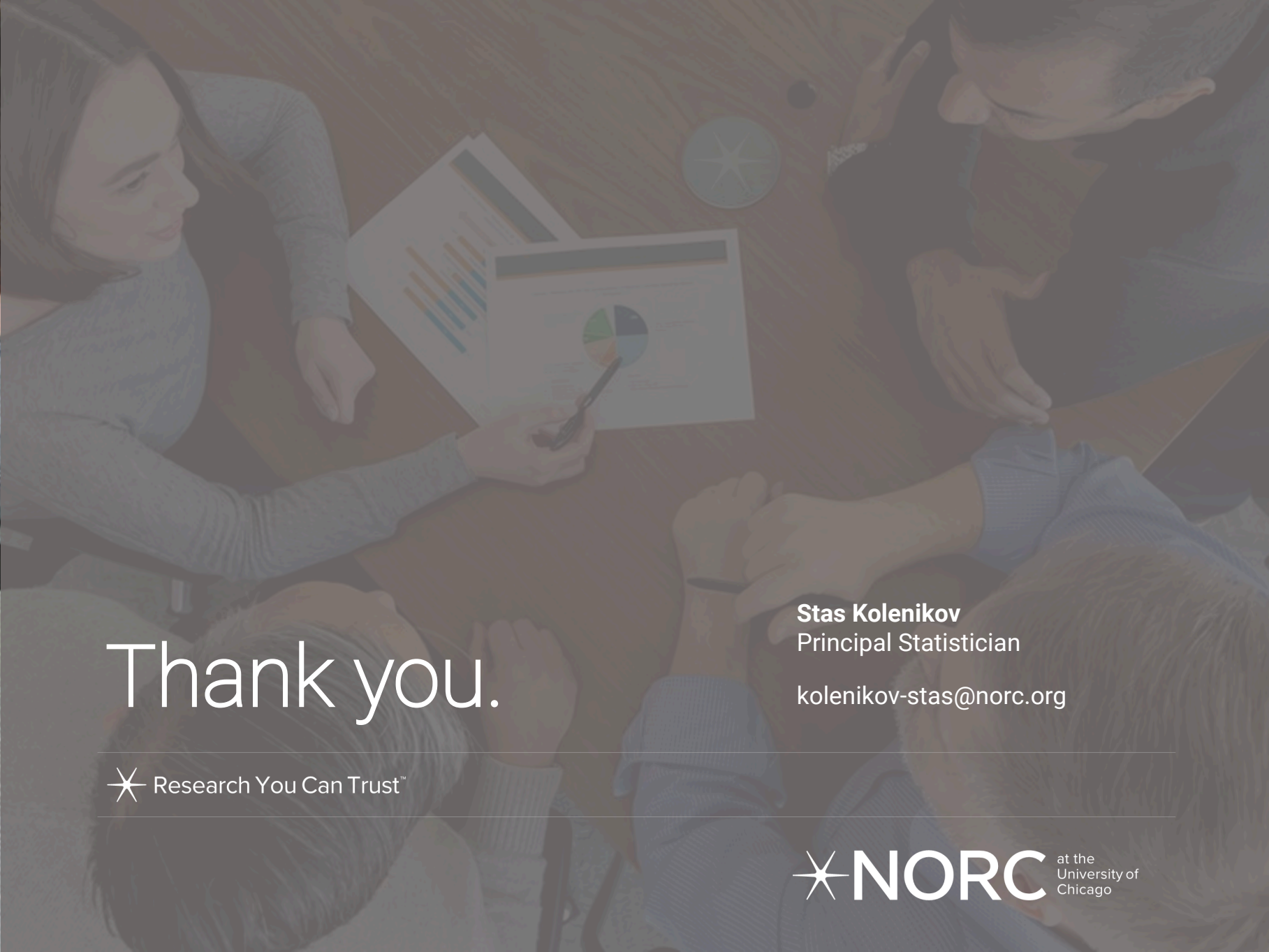
More comprehensive coverage

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

Questions slide



Thank you.

Stas Kolenikov
Principal Statistician

kolenikov-stas@norc.org

Research You Can Trust™

NORC at the University of Chicago