Latent class models applied to QOL scores: a case–study using GSEM to optimize a latent profile analysis

Marcos Almeida MD, MSc, PhD

Tenured Professor of the Faculty of Medicine and the Postgraduate Course (Masters and Doctorate) in Health and Environment at Tiradentes University (UNIT) – Brazil

General physician and cardiologist at Clínica & Hospital São Lucas – Aracaju (SE)

Senior Teaching Assistant in PPCR Course at Harvard T.H.Chan School of Public Health – USA
Marcos Almeida has no relevant conflict of interest related to the content of this presentation;

The views expressed in this presentation do not necessarily reflect the views of the institutions.
LCA (latent class analysis) is one of the highlights available in Stata 15.

This new feature allows identification of “unknown groups” (or classes) within a given population.

When dealing with continuous observed variables, a latent class model is named “latent profile analysis” (LPA) or “latent cluster analysis” or “Gaussian finite mixture models”.

Marcos Almeida – Latent class models applied to QOL scores
2017 Brazilian Stata Users Group Meeting
LCP models use the EM (expectation–maximization algorithm.

It is “an iterative procedure for refining starting values before maximizing the likelihood. The EM algorithm uses the complete-data likelihood as if we have observed values for the latent class indicator variable” (*).

“The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step” (**).

Source:
To check how it works, we used quality-of-life (QOL) scores in a LPA to fit a GSEM (generalized structural equation modeling).

In this case study enrolling 600 individuals, four domains of the questionnaire WHOQOL-BREF are the observed variables, whose scores we converted in a 0–100 scale.
Case-study “situation”

- Questionnaire WHOQOL-BREF:
- Quality of life – Developed by the WHO (1996);
- Number of questions: 26;
- Likert scale: scores from 1 to 5: (1 = not at all; 2 = not much; 3 = moderately; 4 = a great deal; 5 = completely).
- Negatively phrased items (3): Q3, Q4 and Q26;
- Four Domains + Self-appraisal:
  - Physical = mean (Q3r, Q4r, Q10, Q15, Q16, Q17, Q18);
  - Psychological = mean (Q5, Q6, Q7, Q11, Q19, Q26r);
  - Social relationships = mean (Q20, Q21, Q22);
  - Environment = mean (Q8, Q9, Q12, Q13, Q14, Q23, Q24, Q25);
  - Self-appraisal = mean (Q1, Q2).
- Scores lately *4 (range: 4–20) or a scale 0–100.
WHOOQOL-BREF: dimensions

Quality of Life

- Self-appraisal
- Physical health
- Psychological
- Social
- Environment
The goal of the modeling strategy was identifying the “most appropriate” number of classes.

To achieve this task, we specified different number of classes in a sequence of models.

After that, we estimated the marginal predicted means (with 95% confidence intervals) of each domain within each latent class.
We also estimated the *posterior probability* of individuals being in a given class.

The Akaike information criterion (AIC) as well as the *Bayesian* information criterion (BIC) were used as a measure to assess the relative quality of the model.

Plots of the parameters of the “best fit” model and interpretation for the results concerning the identification of (so far) “unknown” groups are presented.
### Latent class: 1

| Variable   | Coef.  | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|------------|--------|-----------|------|-----|---------------------|
| autoav100  | 75.7062| 1.036794  | 73.02| 0.000| 73.67754 – 77.7417  |
| phys100    | 73.8637| 0.896026  | 82.43| 0.000| 72.10755 – 75.61991 |
| psych100   | 73.5623| 1.005847  | 73.13| 0.000| 71.59087 – 75.53372 |
| social100  | 78.1319| 1.184029  | 65.99| 0.000| 75.81074 – 80.45204 |
| envir100   | 58.1636| 0.901369  | 64.53| 0.000| 56.39721 – 59.93052 |

### Latent class: 2

| Variable   | Coef.  | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|------------|--------|-----------|------|-----|---------------------|
| autoav100  | 56.4172| 0.931562  | 60.56| 0.000| 54.59143 – 58.24309 |
| phys100    | 55.8217| 0.844931  | 66.07| 0.000| 54.16571 – 57.47778 |
| psych100   | 49.8789| 0.976312  | 51.09| 0.000| 47.96536 – 51.79243 |
| social100  | 56.2594| 1.157637  | 48.60| 0.000| 53.99054 – 58.5284  |
| envir100   | 42.0669| 0.745565  | 56.42| 0.000| 40.60565 – 43.52821 |

Marcos Almeida – Latent class models applied to QOL scores
2017 Brazilian Stata Users Group Meeting
The EM algorithm

Fitting class model:

Iteration 0:  (class) log likelihood = -658.06709
Iteration 1:  (class) log likelihood = -658.06709

Fitting outcome model:

Iteration 0:  (outcome) log likelihood = -11722.393
Iteration 1:  (outcome) log likelihood = -11722.393

Refining starting values:

Iteration 0:  (EM) log likelihood = -12439.784
Iteration 1:  (EM) log likelihood = -12419.491
Iteration 2:  (EM) log likelihood = -12396.125
Iteration 3:  (EM) log likelihood = -12379.434
Iteration 4:  (EM) log likelihood = -12368.783
Iteration 5:  (EM) log likelihood = -12362.24
Iteration 6:  (EM) log likelihood = -12358.264
Iteration 7:  (EM) log likelihood = -12355.849
Iteration 8:  (EM) log likelihood = -12354.377
Iteration 9:  (EM) log likelihood = -12353.474
Iteration 10:  (EM) log likelihood = -12352.919
Iteration 11:  (EM) log likelihood = -12352.577
Iteration 12:  (EM) log likelihood = -12352.367
Iteration 13:  (EM) log likelihood = -12352.238
Iteration 14:  (EM) log likelihood = -12352.159
Iteration 15:  (EM) log likelihood = -12352.112
Iteration 16:  (EM) log likelihood = -12352.083
Iteration 17:  (EM) log likelihood = -12352.067
Iteration 18:  (EM) log likelihood = -12352.058

Fitting full model:

Iteration 0:  log likelihood = -12210.885
Iteration 1:  log likelihood = -12210.884
Iteration 2:  log likelihood = -12210.884
Latent class marginal means with 95% CIs
according to 2 model-defined classes*

* Class 1 (red): low QOL; Class 2 (green): high QOL

Domains of WHOQOL-BREF
Modeling (code):

```stata
.gsem (autoav100 phys100 psych100 social100 envir100 <- _cons), family(gaussian) link(identity) lclass(C 2) estimates store twoclasses
.gsem (autoav100 phys100 psych100 social100 envir100 <- _cons), family(gaussian) link(identity) lclass(C 3) estimates store threeclasses
.gsem (autoav100 phys100 psych100 social100 envir100 <- _cons), family(gaussian) link(identity) lclass(C 4) estimates store fourclasses
.gsem (autoav100 phys100 psych100 social100 envir100 <- _cons), family(gaussian) link(identity) lclass(C 5)
/* due to slow convergence with further classes, we may add:
.gsem (autoav100 phys100 psych100 social100 envir100 <- _cons), family(gaussian) link(identity) lclass(C 5) startvalues(randomid, draws(5) seed(12345)) emopts(iter(20)) estimates store fiveclasses
.estimates stats twoclasses threeclasses fourclasses fiveclasses
```
Akaike's information criterion and Bayesian information criterion

<table>
<thead>
<tr>
<th>Model</th>
<th>Obs</th>
<th>ll(null)</th>
<th>ll(model)</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>twoclasses</td>
<td>599</td>
<td>.</td>
<td>-12335.39</td>
<td>16</td>
<td>24702.77</td>
<td>24773.1</td>
</tr>
<tr>
<td>threeclasses</td>
<td>599</td>
<td>.</td>
<td>-12210.88</td>
<td>22</td>
<td>24465.77</td>
<td>24562.46</td>
</tr>
<tr>
<td>fourclasses</td>
<td>599</td>
<td>.</td>
<td>-12191.79</td>
<td>28</td>
<td>24439.57</td>
<td>24562.64</td>
</tr>
<tr>
<td>fiveclasses</td>
<td>599</td>
<td>.</td>
<td>-12177.98</td>
<td>34</td>
<td>24423.97</td>
<td>24573.41</td>
</tr>
</tbody>
</table>
Checking it all (with 3 classes)

```
. estat lcmean

Latent class marginal means          Number of obs  =        599

<table>
<thead>
<tr>
<th>Delta-method</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Margin</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>autoav100</td>
<td>49.69824</td>
<td>1.425156</td>
<td>34.87</td>
<td>0.000</td>
</tr>
<tr>
<td>phys100</td>
<td>47.64035</td>
<td>1.221265</td>
<td>39.01</td>
<td>0.000</td>
</tr>
<tr>
<td>psych100</td>
<td>39.49719</td>
<td>1.307835</td>
<td>30.20</td>
<td>0.000</td>
</tr>
<tr>
<td>social100</td>
<td>45.56859</td>
<td>1.763395</td>
<td>25.84</td>
<td>0.000</td>
</tr>
<tr>
<td>envir100</td>
<td>37.37635</td>
<td>1.14273</td>
<td>32.71</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>autoav100</td>
<td>64.36672</td>
<td>0.925847</td>
<td>69.52</td>
<td>0.000</td>
</tr>
<tr>
<td>phys100</td>
<td>63.81917</td>
<td>0.757437</td>
<td>84.26</td>
<td>0.000</td>
</tr>
<tr>
<td>psych100</td>
<td>60.74826</td>
<td>0.897217</td>
<td>67.71</td>
<td>0.000</td>
</tr>
<tr>
<td>social100</td>
<td>67.13671</td>
<td>1.107349</td>
<td>60.63</td>
<td>0.000</td>
</tr>
<tr>
<td>envir100</td>
<td>48.23475</td>
<td>0.737886</td>
<td>65.37</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>autoav100</td>
<td>81.56046</td>
<td>1.318448</td>
<td>61.86</td>
<td>0.000</td>
</tr>
<tr>
<td>phys100</td>
<td>79.67181</td>
<td>1.157432</td>
<td>68.84</td>
<td>0.000</td>
</tr>
<tr>
<td>psych100</td>
<td>80.0374</td>
<td>1.180748</td>
<td>67.79</td>
<td>0.000</td>
</tr>
<tr>
<td>social100</td>
<td>83.14302</td>
<td>1.489364</td>
<td>55.82</td>
<td>0.000</td>
</tr>
<tr>
<td>envir100</td>
<td>63.32276</td>
<td>1.126052</td>
<td>56.23</td>
<td>0.000</td>
</tr>
</tbody>
</table>
```
Getting the predicted values

. estat lcprob

Latent class marginal probabilities

<table>
<thead>
<tr>
<th>C</th>
<th>Delta-method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Margin</td>
</tr>
<tr>
<td>1</td>
<td>.2065115</td>
</tr>
<tr>
<td>2</td>
<td>.5511042</td>
</tr>
<tr>
<td>3</td>
<td>.2423844</td>
</tr>
</tbody>
</table>

Expected “posterior” classification – for each individual

Marcos Almeida – Latent class models applied to QOL scores
2017 Brazilian Stata Users Group Meeting
Working with predictions

```
cpost3~1  cpost3~2  cpost3~3  max3cl~s  p~3~s
1. 3.34e-06  .5406071  .4593896  .5406071  2
2. 3.53e-09  .0100017  .9899983  .9899983  3
3. 0.0329318  .9665484  .0005198  .9665484  2
4. 0.0000105  .8185978  .1813917  .8185978  2
5. 0.03552  .9642501  .0002299  .9642501  2
6. 6.50e-09  .0371493  .9628507  .9628507  3
7. 0.0185071  .9805173  .0009756  .9805173  2
8. 0.9947197  .0052803  3.10e-10  .9947197  1
9. 0.99993  .00007  9.53e-14  .99993  1
10. 7.10e-11  .002671  .9973291  .9973291  3
```

```
.predict cpost3class*, classposterior
.egen max3class = rowmax(cpost3class*)
generate pred3class = 1 if cpost3class1==max3class
.replace pred3class = 2 if cpost3class2==max3class
.replace pred3class = 3 if cpost3class3==max3class
.list cpost3class1-pred3class in 1/10, compress
tab pred3class
```

Marcos Almeida – Latent class models applied to QOL scores
2017 Brazilian Stata Users Group Meeting
Graph with the predictions

Marcos Almeida – Latent class models applied to QOL scores
2017 Brazilian Stata Users Group Meeting
A “view” of the unobserved classes

Model-defined Latent Classes (mean scores + 95% CIs)

Low QOL

Middle QOL

High QOL

Scores of Quality of Life (QOL)

Self appraisal  Physical  Psychological  Social  Environment

Marcos Almeida – Latent class models applied to QOL scores
2017 Brazilian Stata Users Group Meeting
Working with the matrix

.estat lcmean, post
.ereturn list
.matrix D = e(b)
.svmat D, names(var)
.matrix list D

D[1,15]
autoav100 phys100 psych100 social100 envir100 autoav100 phys100 psych100 social100 envir100
y1  49.698243  47.640352  39.497187  45.568586  37.37635  64.366725  63.819174  60.748262  67.136715  48.234751

3:     3:     3:     3:     3:
autoav100 phys100 psych100 social100 envir100
y1  81.56046  79.671812  80.037403  83.143017  63.322764

A couple of “rename”, “generate”, “reshape”, “label” and “replace” commands later....
Means of the unobserved classes

Mean scores of QOL parameters
according to 3 model-defined classes

Scores of quality of life
Low QOL  Middle QOL  High QOL

Mean scores of QOL parameters
self_appraisal  physical
psychological  social
environment

Marcos Almeida – Latent class models applied to QOL scores
2017 Brazilian Stata Users Group Meeting
Mean scores of QOL parameters according to 3 model-defined latent classes

Scores of Quality of Life (QOL)

Effects with Respect to

- Low QOL
- Middle QOL
- High QOL
Mean scores of QOL parameters according to the latent classes

Scores of Quality of Life (QOL)

Effects with Respect to

- Low QOL
- Middle QOL
- High QOL

Marcos Almeida – Latent class models applied to QOL scores
2017 Brazilian Stata Users Group Meeting
A few caveats

- Structural equation modeling (SEM as well as GSEM), dubbing an expression used in the Stata Manual, is a “way of thinking”.
- *Nota bene*: not the only one!
- To some extent, we can rely on LCP under GSEM strategy in order to embrace several issues.
- That being said, LCP shall not be taken as a “jack of all trades”, rather, it is a resource to approach a specific problem.
Latent class profile (LCP) analysis may be performed in Stata 15, under the GSEM “umbrella”.

A step-by-step approach in terms of command and modeling was hereby presented.

The number of unobserved classes can be defined after the empirical examination of the data set.

Also, LCP analysis gives information about the probability of an individual being classified within a given class.
AIC and BIC are helpful tools to select the most appropriate model.

LCP analysis displays point estimates as well as 95% confidence intervals for all calculations.

Convergence issues may be curbed by the appropriate selection of starting numbers and the limit of iteration for the EM (expectation–maximization) algorithm.
There is much to learn from LCP analysis.

Such a remarkable method can be further used to tackle complex models, for example, by integrating latent constructs with a panoply of regression analyses as well as a strategy to cope with unobserved heterogeneity.

This notwithstanding, neither a wrongly-defined study question nor a carelessly-measured questionnaire will suffice with the overarching family of Latent Class Models.
Extended regression analysis

Generalized structural equation models

Finite mixture models

Latent class models

Latent class profile


Thank you!

Contact:

- Marcos Almeida, MD PhD
- Email: virtual.596@gmail.com