2017 Brazilian Stata Users Group meeting
São Paulo | 8 December
University of São Paulo

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

São Lucas
CLÍNICA & HOSPITAL

PPCR Principles and Practice of Clinical Research

Disclosures

- 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.

Introduction – 1

- 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".

Introduction – 2

- LCP models use the EM (expectation-maximization algorith.
- 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:
* Stata Finite Mixture Models Reference Manual.
** Wikipedia, at https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm)

Introduction – 3

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

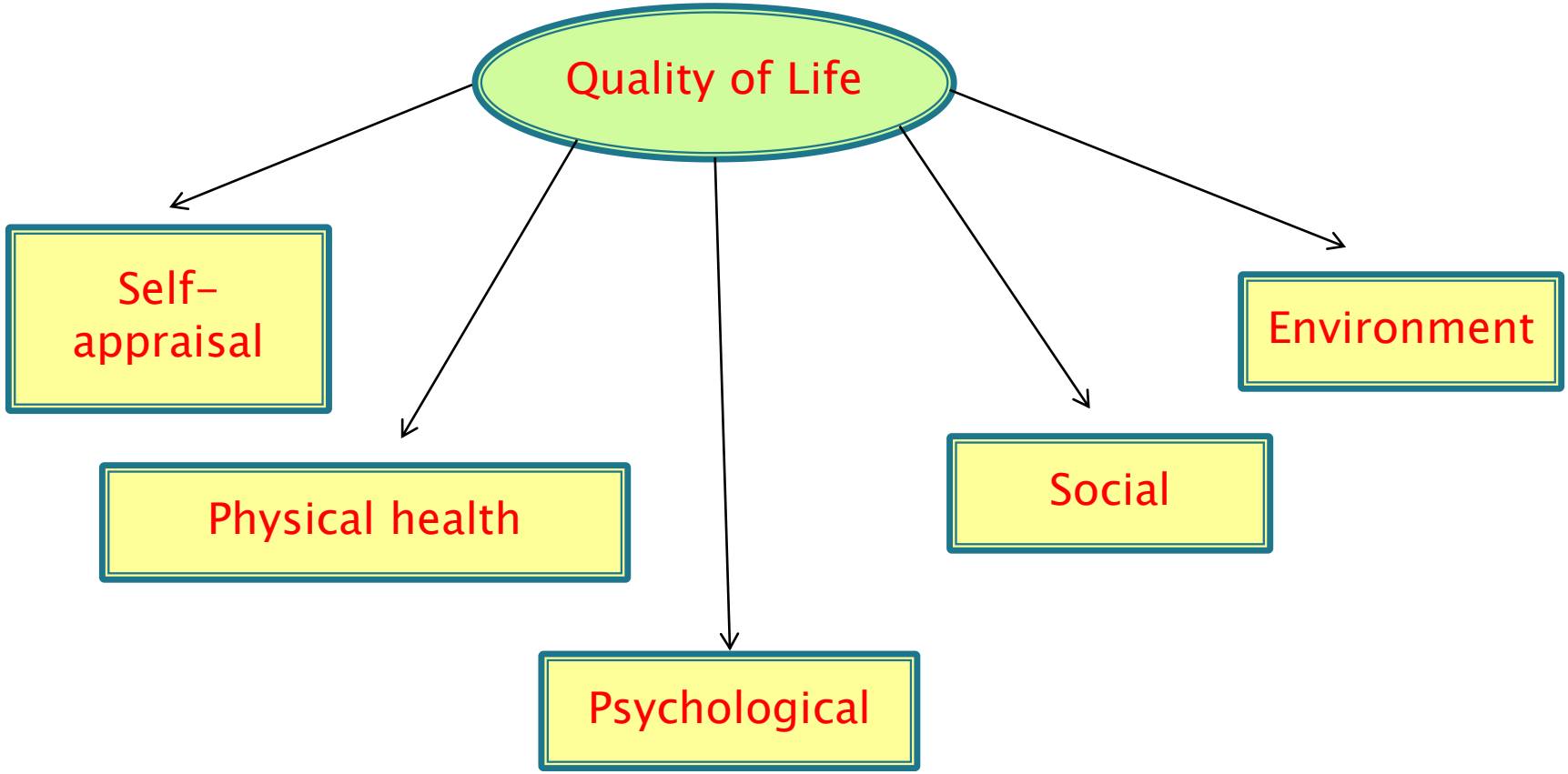
- In this case study enrolling 600 individuals, four domains of the questionnaire WHOQOL-BREF are the observed variables, whose scores we converted in a 0-100 scale.

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.

WHOQOL-BREF: dimensions



Model design and steps of the analysis -1

- 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.

Model design and steps of the analysis – 2

- 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.

```
.gsem (autoav100 phys100 psych100 social100 envir100 <- _cons), family(gaussian)
link(identity) lclass(C 2) nolog vsquish
```

.

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
autoav100						
_cons	75.70962	1.036794	73.02	0.000	73.67754	77.7417
phys100						
_cons	73.86373	.896026	82.43	0.000	72.10755	75.61991
psych100						
_cons	73.5623	1.005847	73.13	0.000	71.59087	75.53372
social100						
_cons	78.13139	1.184029	65.99	0.000	75.81074	80.45204
envir100						
_cons	58.16386	.9013696	64.53	0.000	56.39721	59.93052
var(e.autoav100)	198.7334	12.50471			175.6757	224.8176
var(e.phys100)	148.0535	9.585563			130.4093	168.0849
var(e.psych100)	169.5343	11.50187			148.4256	193.6451
var(e.social100)	288.3311	18.16503			254.8386	326.2253
var(e.envir100)	132.9451	8.714268			116.917	151.1704

.

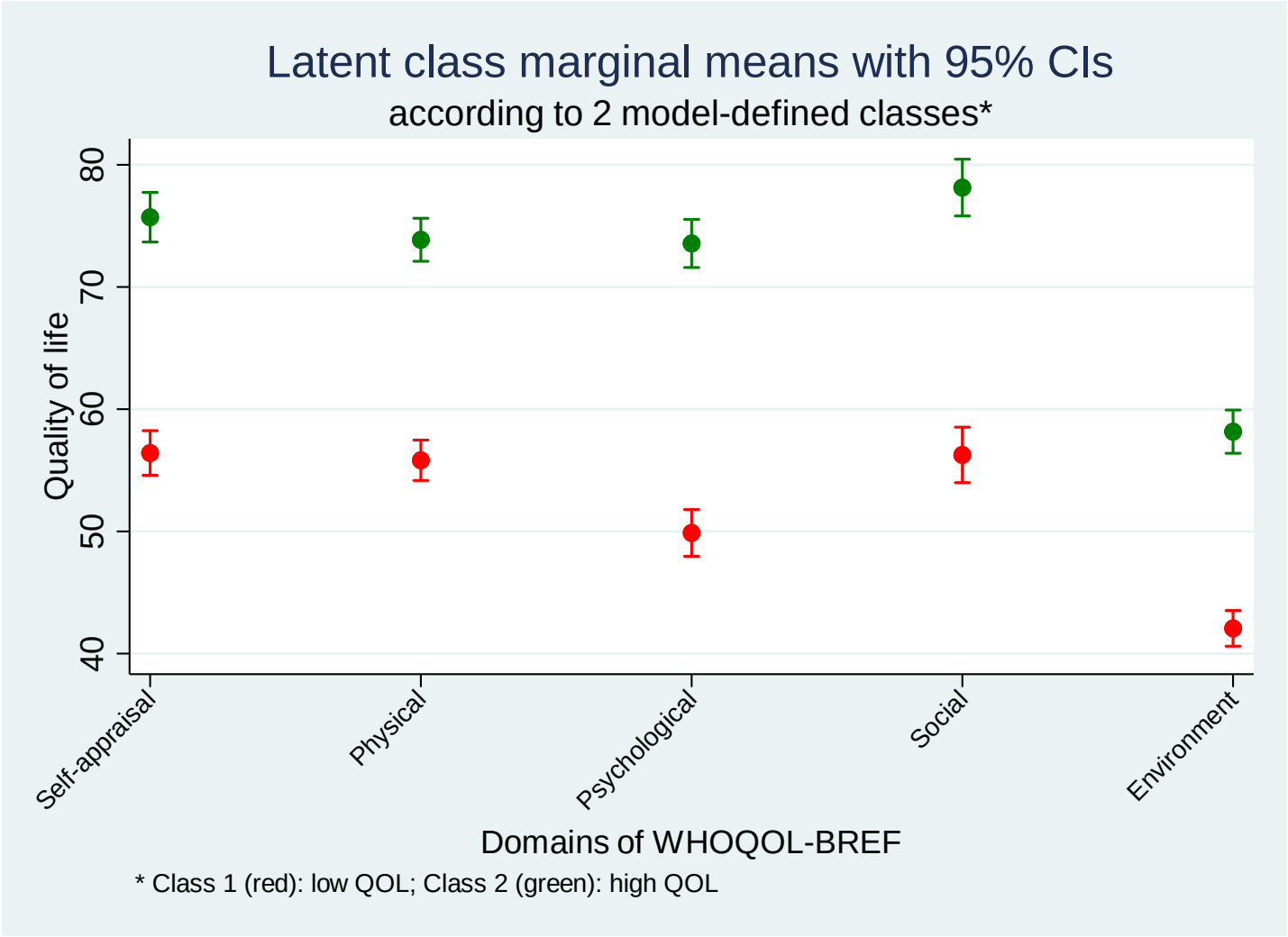
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
autoav100						
_cons	56.41726	.9315624	60.56	0.000	54.59143	58.24309
phys100						
_cons	55.82174	.8449315	66.07	0.000	54.16571	57.47778
psych100						
_cons	49.87889	.9763122	51.09	0.000	47.96536	51.79243
social100						
_cons	56.25947	1.157637	48.60	0.000	53.99054	58.5284
envir100						
_cons	42.06693	.745565	56.42	0.000	40.60565	43.52821
var(e.autoav100)	198.7334	12.50471			175.6757	224.8176
var(e.phys100)	148.0535	9.585563			130.4093	168.0849
var(e.psych100)	169.5343	11.50187			148.4256	193.6451
var(e.social100)	288.3311	18.16503			254.8386	326.2253
var(e.envir100)	132.9451	8.714268			116.917	151.1704

Latent class: 1

Latent class: 2

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*
Quality of life
40
50
60
70
80
Self-appraisal
Physical
Psychological
Social
Environment
Domains of WHOQOL-BREF
* Class 1 (red): low QOL; Class 2 (green): high QOL

Modeling (code):

```
.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
```

```
. estimates stats twoclasses threeclasses
fourclasses fiveclasses
```

Akaike's information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
twoclasses	599	.	-12335.39	16	24702.77	24773.1
threeclasses	599	.	-12210.88	22	24465.77	24562.46
fourclasses	599	.	-12191.79	28	24439.57	24562.64
fiveclasses	599	.	-12177.98	34	24423.97	24573.41

Checking it all (with 3 classes)

```
. estat lcmean

Latent class marginal means                     Number of obs     =        599
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
1						
autoav100	49.69824	1.425156	34.87	0.000	46.90499	52.4915
phys100	47.64035	1.221265	39.01	0.000	45.24672	50.03399
psych100	39.49719	1.307835	30.20	0.000	36.93388	42.0605
social100	45.56859	1.763395	25.84	0.000	42.1124	49.02478
envir100	37.37635	1.14273	32.71	0.000	35.13664	39.61606
2						
autoav100	64.36672	.9258747	69.52	0.000	62.55204	66.18141
phys100	63.81917	.7574375	84.26	0.000	62.33462	65.30372
psych100	60.74826	.8972173	67.71	0.000	58.98975	62.50678
social100	67.13671	1.107349	60.63	0.000	64.96635	69.30708
envir100	48.23475	.7378856	65.37	0.000	46.78852	49.68098
3						
autoav100	81.56046	1.318448	61.86	0.000	78.97635	84.14457
phys100	79.67181	1.157432	68.84	0.000	77.40329	81.94034
psych100	80.0374	1.180748	67.79	0.000	77.72318	82.35163
social100	83.14302	1.489364	55.82	0.000	80.22392	86.06212
envir100	63.32276	1.126052	56.23	0.000	61.11574	65.52979

Getting the predicted values

```
. estat lcprob

Latent class marginal probabilities             Number of obs
```

	Margin	Delta-method Std. Err.	[95% Conf.	Interval]
C				
1	.2065115	.0227046	.1655291	.2545452
2	.5511042	.0271832	.4974427	.6036017
3	.2423844	.0254649	.1960066	.2956998

Expected “posterior” classification – for each individual

Working with predictions

.

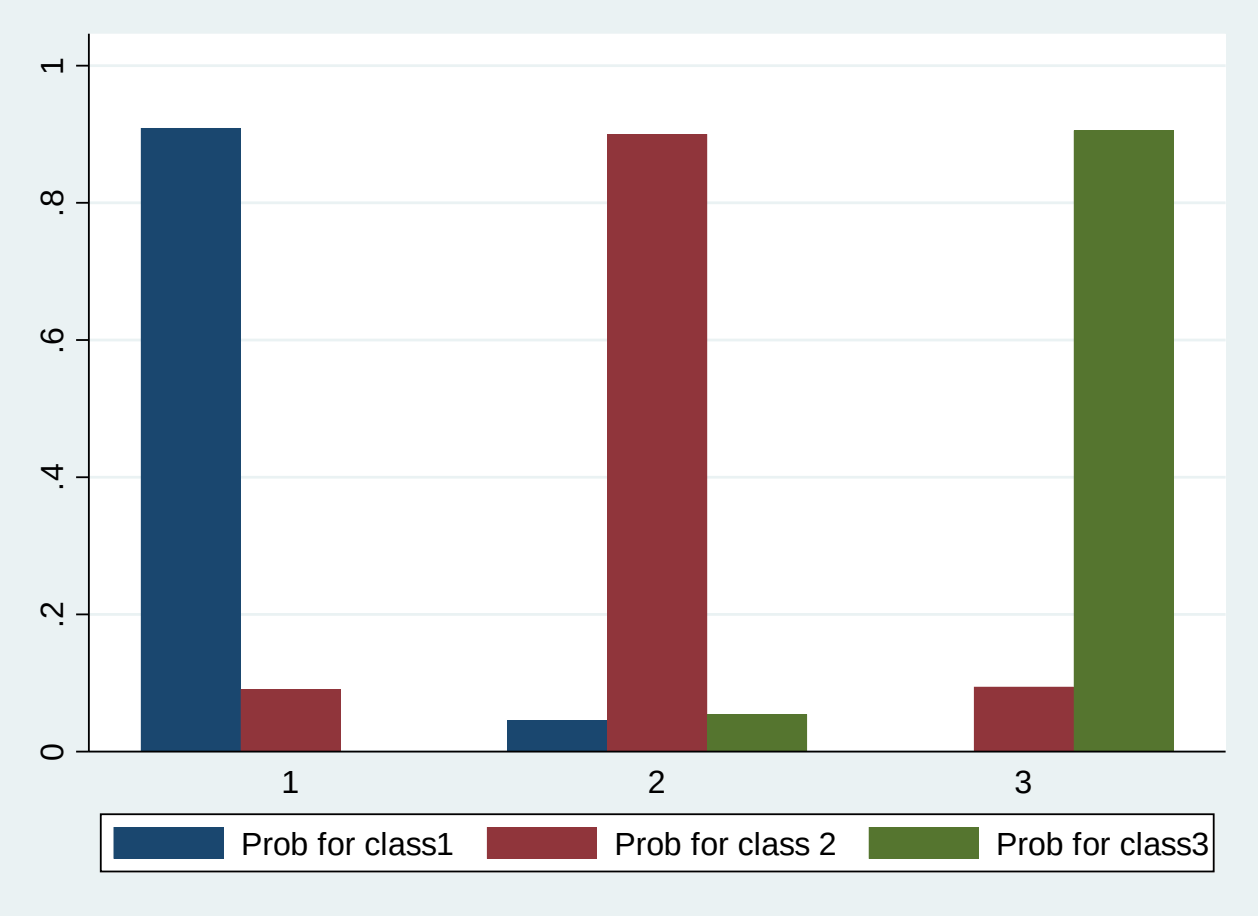
	cpost3~1	cpost3~2	cpost3~3	max3cl~s	p~3~s
1.	3.34e-06	.5406071	.4593896	.5406071	2
2.	3.53e-09	.0100017	.9899983	.9899983	3
3.	.0329318	.9665484	.0005198	.9665484	2
4.	.0000105	.8185978	.1813917	.8185978	2
5.	.03552	.9642501	.0002299	.9642501	2
6.	6.50e-09	.0371493	.9628507	.9628507	3
7.	.0185071	.9805173	.0009756	.9805173	2
8.	.9947197	.0052803	3.10e-10	.9947197	1
9.	.99993	.00007	9.53e-14	.99993	1
10.	7.10e-11	.002671	.9973291	.9973291	3

```
. tab pred3class
```

pred3class	Freq.	Percent	Cum.
1	119	19.87	19.87
2	340	56.76	76.63
3	140	23.37	100.00
Total	599	100.00	

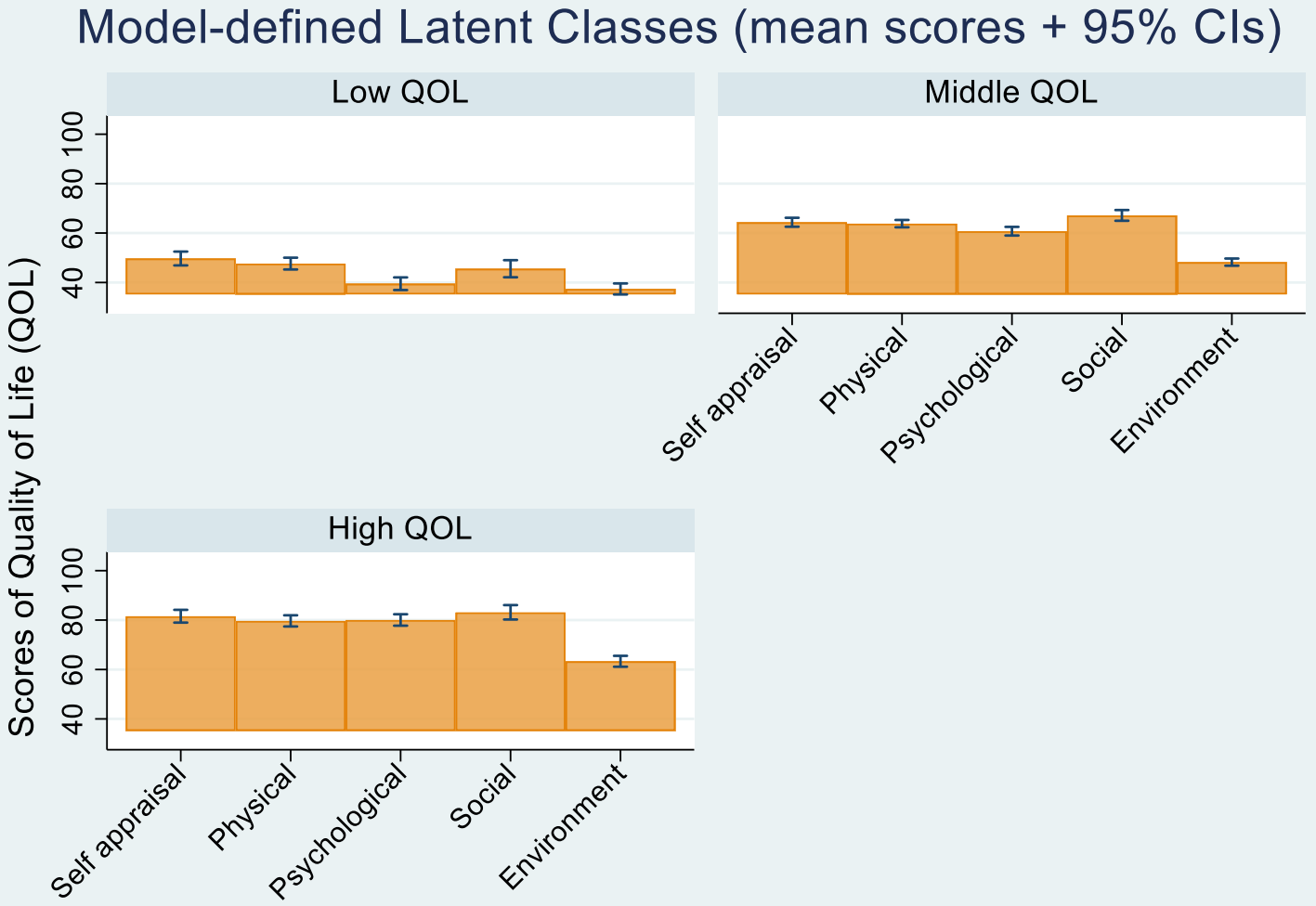
```
.predict cpost3class*, classposteriorpr
.egen max3class = rowmax(cpost3class*)
.generate pred3class = 1 if cpost3class1==max3class
.replace pred3class = 2 if cpost3class2==max3class
.replace pred3class = 3 if cpost3class3==max3class
.list cpost3class1-pred3class in 1/10, compress
.tab pred3class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

Graph with the predictions




```
. graph bar (mean) cpost3class1 cpost3class2 cpost3class3 , over(pred3class)
legend(label(1 "Prob for class1") label(2 "Prob for class 2") label(3 "Prob
for class3") cols(3))
```

A “view” of the unobserved classes



Working with the matrix

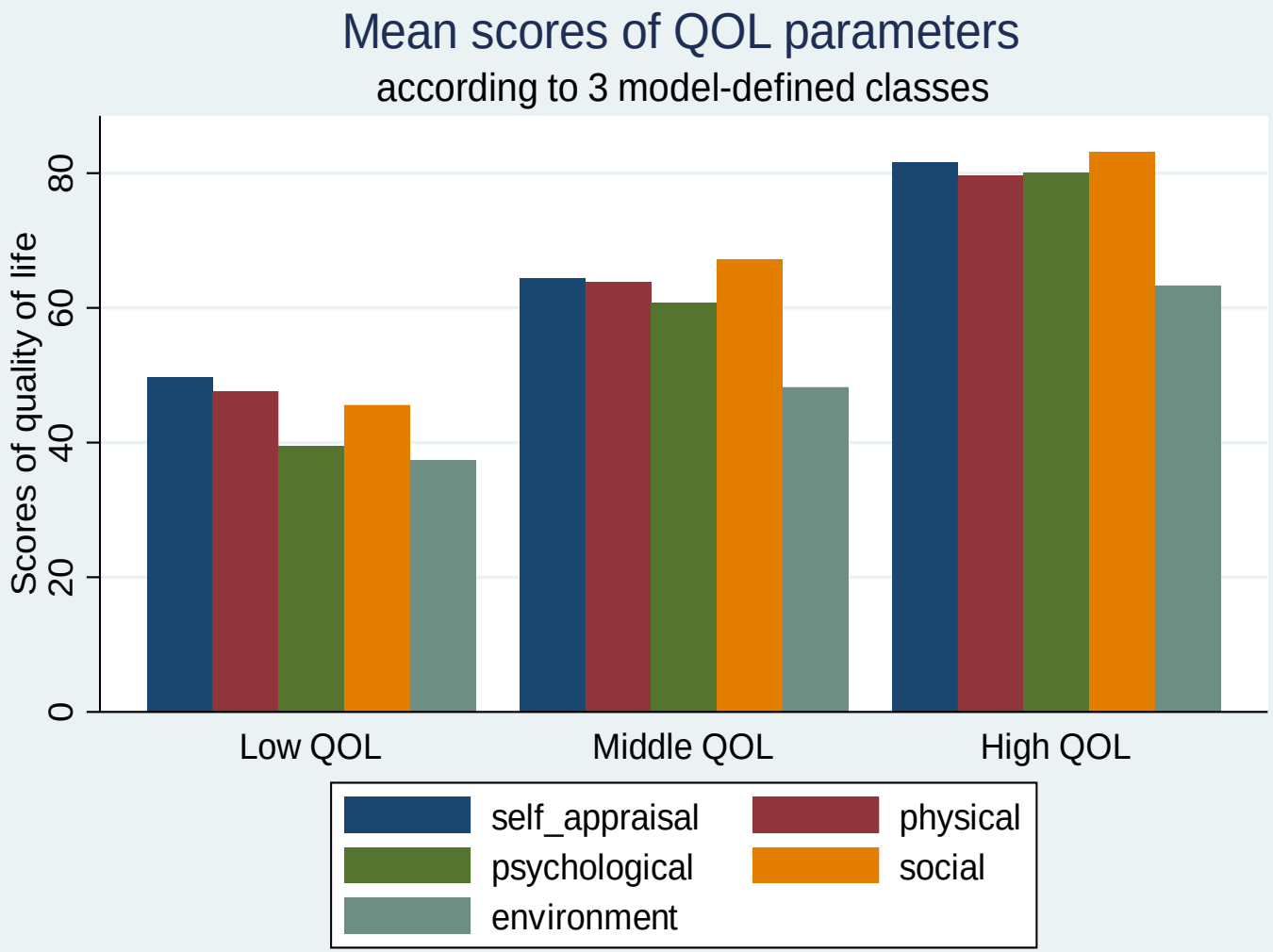
```
.estat lcmean, post
.ereturn list
.matrix D = e(b)
.svmat D, names(var)
.matrix list D
```

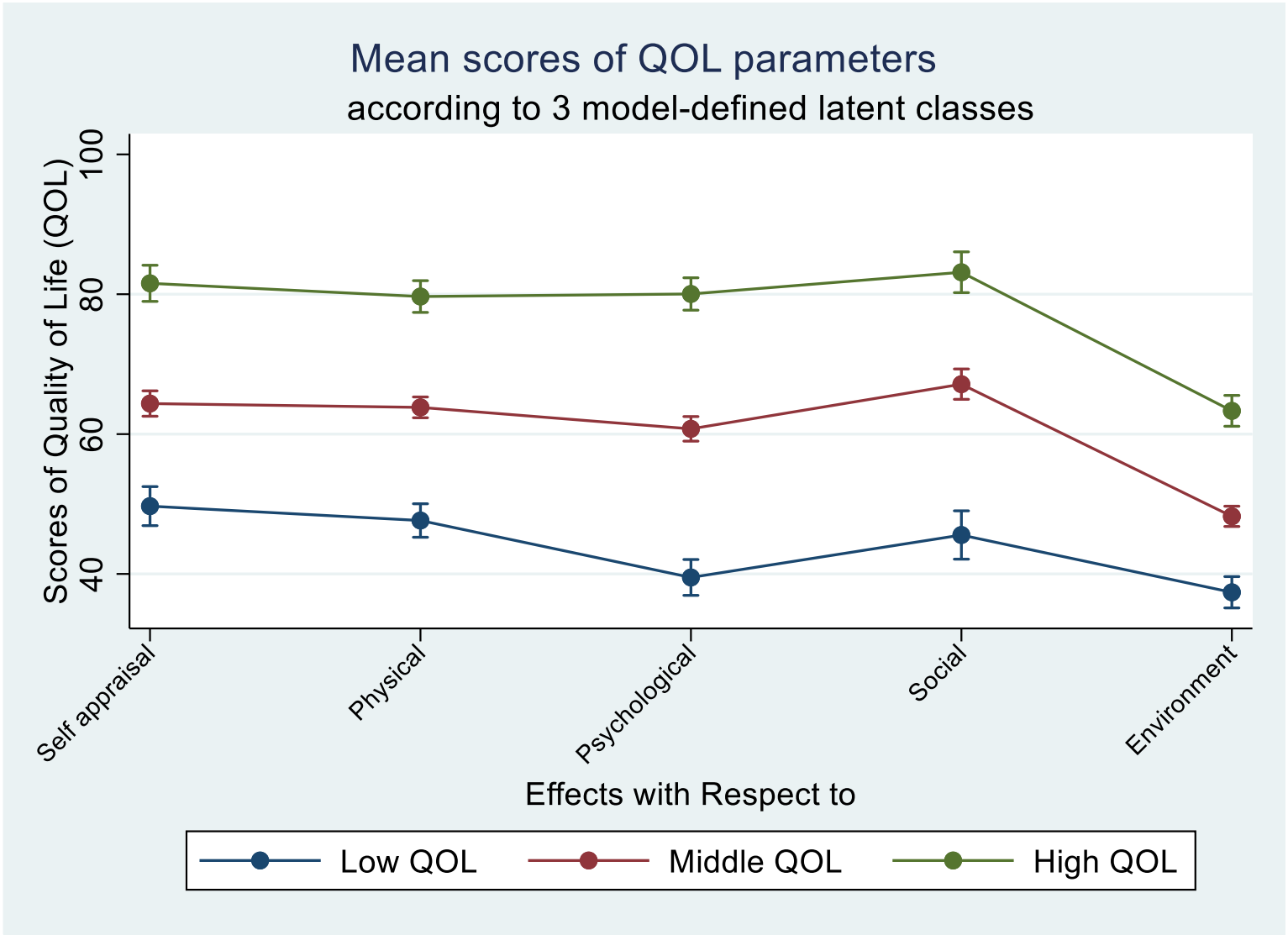
```
. ereturn list

scalars:
                  e(N) =  599

macros:
             e(title) : "Latent class marginal means"
        e(properties) : "b V"

matrices:
                  e(b) :  1 x 15
                  e(V) :  15 x 15
```

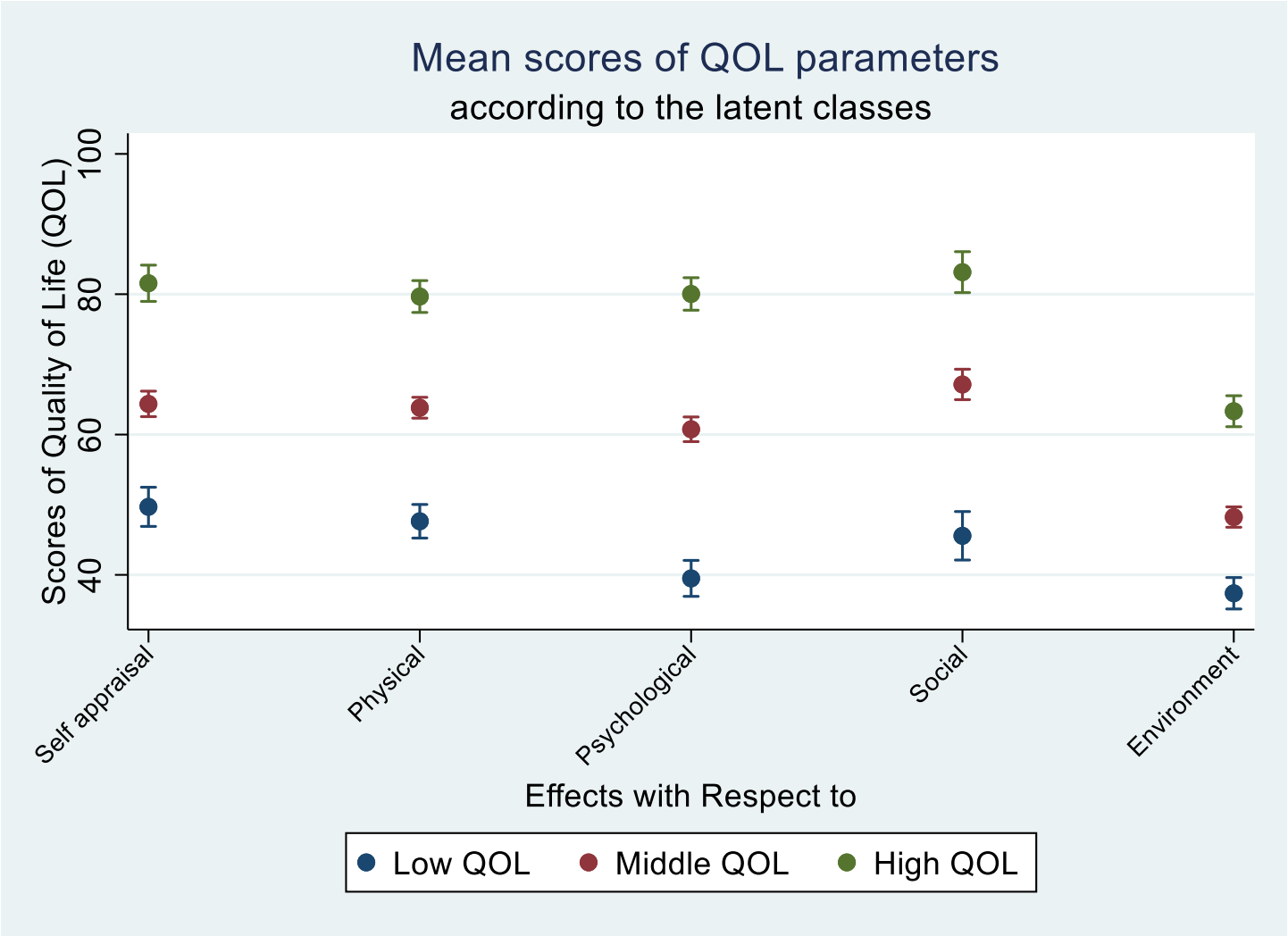
```
D[1,15]
            1:          1:          1:          1:          1:          2:          2:          2:          2:          2:
     autoav100    phys100   psych100  social100   envir100  autoav100    phys100   psych100  social100   envir100
y1   49.698243  47.640352  39.497187  45.568586   37.37635  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 latent classes
Scores of Quality of Life (QOL)
40
60
80
100
Self appraisal
Physical
Psychological
Social
Environment
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)
40
60
80
100
Self appraisal
Physical
Psychological
Social
Environment
Effects with Respect to
Low QOL
Middle QOL
High QOL

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.

Closing remarks – I

- 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.

Closing remarks – II

- 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.

Closing remarks – III

- 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

References

- Acock, Alan C. 2013. Discovering Structural Equation Modeling Using Stata. Revised edition. StataPress.
- Hallquist MN, Wright AGC. Mixture modeling methods for the assessment of normal and abnormal personality I: Cross-sectional models. J Pers Assess. 2014; 96(3): 256-268.
- Kline, Rex B. 2016. Principles and Practice of Structural Equation Modeling. Fourth edition. Guilford.
- StataCorp. Structural Equation Modeling Reference Manual. Downloadable at: https://www.stata.com/bookstore/structural-equation-modeling-reference-manual/
- StataCorp. Finite Mixture Models Reference Manual. Downloadable at: https://www.stata.com/bookstore/finite-mixture-models-reference-manual/
- The WHOQOL Group. World Health Organization. WHOQOL: measuring quality of life. Geneva: WHO; 1997 (MAS/MNH/PSF/97.4). Also in: http://www.who.int/substance_abuse/research_tools/whoqolbref/en/

Thank you!

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