Modeling demographic and health survey (DHS) data by latent class models: An application

JOSÉ G. DIAS
ISCTE - IUL Business School
Dep. of Quantitative Methods & UNIDE
Av. das Forças Armadas, Lisboa
PORTUGAL
jose.dias@iscte.pt

Abstract: This paper models a battery of indicators using a latent class model. DHS (Demographic and Health Survey) data from Mozambique measuring the sources of information on how to avoid HIV/AIDS transmission are considered. The clustering of the women's sample shows four clusters with different profiles. A set of variables is used to describe these segments.

Key–Words: Latent class models, DHS data, AIDS, Mozambique, Segmentation methods

1 Introduction

Fast computing technology has provided the possibility for more complex models, which better describe and take into account specific characteristics of data.

Latent class analysis has become a very popular modeling technique as it enables the identification of discrete heterogeneity in populations, *i.e.*, this clustering technique finds homogeneous subpopulations in a given population (McLachlan and Peel, 2000; Dias, 2004; Skrondal and Rabe-Hesketh, 2004; Dias and Willekens, 2005). This model-based clustering approach is applied to data from the Mozambique DHS 2003. In particular, we model twelve binary indicators of access to information on how to avoid HIV/AIDS from a sample of Mozambican women.

The structure of the paper is the following: Section 2 describes the latent class model. Section 3 presents the application to DHS data and Section 4 discusses the implications of these results.

2 The proposed model

In latent class models individuals are assumed to belong to a given latent class or cluster s ($s = 1, \ldots, S$) with unknown class membership. Consider a sample of n observations. An observation is denoted by i ($i = 1, \ldots, n$) and is characterized by J indicators. The vector of indicators is defined by $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})$. The latent class model with S latent classes for $\mathbf{y}_i$ is defined by the composite density

$$f(\mathbf{y}_i; \boldsymbol{\varphi}) = \sum_{s=1}^{S} \pi_s f_s(\mathbf{y}_i; \boldsymbol{\theta}_s), \qquad (1)$$

where the underlying discrete latent variable, z_i, has a multinominal distribution, such that $z_i \sim Multi_{S-1}(\boldsymbol{\pi})$, with $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_{S-1})$, $\pi_s > 0$ and $\sum_{s=1}^{S} \pi_s = 1$. The conditional probability function (of latent class s) is $f_s(\mathbf{y}_i; \boldsymbol{\theta}_s)$, and the vector of parameters of the model is $\boldsymbol{\varphi} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S)$, where $\boldsymbol{\theta}_s$ is the vector of parameters in the latent class s.

An important issue in latent class modeling is estimation. In this setting, the maximum likelihood estimate of a set of independent observations can only be obtained by iterative procedures, such as the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977). This is an algorithm for maximum likelihood estimation with incomplete data using data augmentation. It is divided into two steps: the E-step consists in associating each individual observation with its conditional expectation of class membership, given the observed data; the M-step consists in maximizing the full data log-likelihood function ($\ell_C(\hat{\boldsymbol{\varphi}}; \mathbf{y})$) using the complete data as the observed data.

We use the *BIC – Bayesian Information Criterion* (Schwarz, 1978; Dias, 2006) to select the optimal number of classes in the model, which corresponds to the model that minimizes

$$C_S = -2\ell_S(\hat{\boldsymbol{\varphi}}; \mathbf{y}) + N_S \cdot \log n, \qquad (2)$$

where $\ell_S(\hat{\varphi};\mathbf{y})$ is the log-likelihood function of the latent class model with S latent classes and N_S represents the number of parameters in the model.

3 Application

The Demographic and Health Surveys (DHS) collect and disseminate accurate and nationally representative data on population characteristics, fertility, reproductive health, maternal and child health including antenatal care, immunisation and nutritional indicators, infant and child mortality, morbidity including HIV and Malaria and a range of socioeconomic variables from over 75 countries.

The current study uses data from the Mozambique DHS 2003. Twelve sources of information on AIDS were selected: *Radio*, *TV*, *Magazines*, *Flyers/Posters/Pub Advertising*, *Health workers*, *Church*, *School/Teachers*, *Community meetings*, *Friends/relatives*, *At work*, *Health post*, *Health activist*. Respondents were asked to answer whether they had heard how to avoid AIDS through those media. Thus, the density f_s corresponds to a product of Bernoulli distributions (*i.e.*, within each latent class we assume the responses to the indicators are independent), and θ_{js} is the probability of answering *yes* in indicator j conditional on belonging to segment s.

We use the subsample of all 12398 women. They are aged between 15 and 49 years old (see column *Aggregate* in Table 3). Regarding education, 36.2% of the women did not attended the school, whereas only 0.3% of them have a higher degree. The large majority of them live in a partnership (married, living together, not living together), being 18.2% single. These variables will be used in the profiling of the results.

Before estimating the latent class model, we give a short description of the data used as input. Last column of Table 2 provides the percentages for the full sample. We conclude that the main sources of information for these women are the radio (79.7%) and friends/relatives (52.8%). To a lesser extent for some women, TV (20.8%) and community meetings (15.0%) may play a role as sources of information. First, we estimated the latent class model with these twelve indicators and, then, latent classes are characterized by profiling variables (Table 3).

Latent class models with different number of classes, ranging from $S = 1$ to $S = 8$, were estimated. For the EM algorithm random initialization was considered and the convergence tolerance level equals 10^{-6}. A (global) minimum for BIC was reached when $S = 4$ and, therefore, a solution with four latent classes was chosen (Table 1).

Table 2 reports the characterization of the four latent classes in terms of size and distribution of the twelve indicators.

Latent class 1 is the largest, corresponding to 53.8% of the sample, and includes women with short access to means of information on AIDS. They rely mostly on the radio and are strongly dependent on friends and relatives (69.9%, comparing with 52.8% at aggregate level).

Latent class 2 corresponds to 23.7% of the sample. Comparing with the aggregate profile, this group is quite below the aggregate profile for all the mass media such as radio, TV, Magazines, Flyers, Posters, and Advertising. This group relies on the health structure (*e.g.*, health works, health post, health activist), church, and community meetings. Interestingly, this group is far below the average in terms of friends and relatives, which shows some isolation at the inner circle that is compensated at the community level.

Latent classes 3 and 4 show a rather different profile when compared with latent classes 1 and 2. Latent class 3, with 14.4% of the women, has access to several sources of information. The main difference with the aggregate profile has to do with the access to information broadcast by TV. Indeed, at the aggregate level 20.8% had access to information on AIDS through TV, but in this segment this percentage increases to 85.6%. The same happens with radio at a smaller scale.

Finally, latent class 4 (size 8.0%) contains the women who has access to all mass media communication, in particular in written support such as Magazines and Flyers. Moreover, they also have access to information through school/teachers. Less relevant in this group is the social network provided by community, friends and relatives.

The variables that best discriminate the four latent classes are *Region*, *Place of residence*, *Age group*, *Education*, and *Current marital status*. Table 3 reports the aggregate profile of the sample as well as the profile within each latent class. Larger distances to the aggregate results mean more discriminant power of the variable.

When compared with the overall sample, in latent class 1 the proportions for all regions are above the aggregate pattern with the exception of Maputo; it is strongly concentrated in the countryside (more 14.4% than at the aggregate level); they are less well educated and tend to have a partnership. Latent class 2 shows a profile similar to latent class 1, but less impoverished. For example, in latent class 1, 45.5% of the women have not attended school; in latent class 2, this percentage reduces to 42.0%.

Latent classes 3 and 4 have similar patterns. Both tend to be concentrated in Maputo Cidade (capital city) and Maputo Region (excluding the capital city),

Table 1: Model selection (number of latent classes).

# Latent classes	Log-likelihood value	# Parameters	BIC
1	-48186.4	12	96485.9
2	-46199.0	25	92633.5
3	-46026.3	38	92410.9
4	**-45918.6**	**51**	**92317.9**
5	-45863.0	64	92329.3
6	-45834.2	77	92394.1

Table 2: HIV/AIDS information source.

	Latent classes				Aggregate
	1	2	3	4	
Cluster Size	0.538	0.237	0.144	0.080	1.000
Indicators					
Radio	0.755	0.723	0.984	0.961	0.797
TV	0.013	0.027	0.856	0.888	0.208
Magazines	0.002	0.006	0.101	0.645	0.069
Flyers/Posters/Advertising	0.024	0.044	0.033	0.394	0.060
Health workers	0.018	0.082	0.021	0.064	0.037
Church	0.060	0.154	0.045	0.075	0.081
School/Teachers	0.046	0.081	0.174	0.277	0.092
Community meetings	0.125	0.302	0.050	0.049	0.150
Friends/relatives	0.699	0.223	0.470	0.392	0.528
At work	0.006	0.017	0.015	0.049	0.013
Health post	0.054	0.126	0.074	0.034	0.072
Health activisties	0.018	0.103	0.069	0.096	0.052

defining a urban setting. Indeed, the capital city contains 11.2% of the respondents, but in latent class 3 and 4 they are 30.3% and 45.6%, respectively. They tend to be young, below 25 years old. The main distinction between latent classes 3 and 4 has to do with level of education attained. In latent class 4, 46.5% of the women reach the secondary school, whereas in latent class 3 only 23.3% go for that level of education. The fact that they are studying tends to reduce the proportion of women with a partnership in latent class 4. For example, at the aggregate level 18.2% women reported they are single at the time of the survey. However, in latent classes 3 and 4 that percentage reaches 33.5% and 44.1%, respectively.

Combining both analysis we obtain a clear picture of each segment or latent class. Latent classes 1 and 2 tend to be rural from outside of the region of Maputo. Lacking of schooling and access to mass media (*e.g.*, lack of electricity at the household level) may explain the difficulty of access to information in these two segments. Segment 2 has more support from heath structures than segment 1 that is more dependent on (informal) sources provided by friends and relatives. On the other hand, segments 3 and 4 (22.4% of the sample) are mainly from urban areas of Maputo region. They tend to be young, educated, and single. Despite both segments have access to TV, segment 4 shows a broader range of media access, in particular written material such as Magazines and Flyers, whereas segment 3 is more centered on radio and TV.

4 Discussion

Historically, the AIDS pandemic has been characterized through two distinct patterns: pattern I and pattern II (World Bank, 1997; Piot *et al.*, 1988). Pattern I occurs mostly among men in developed countries, mainly among homosexuals and intravenous drug users. Pattern II occurs in certain developing countries (*e.g.*, some Sub-Saharan countries), being spread mainly through heterosexual contact, with strong concentration in impoverished areas of large cities and along transportation routes. The mobility of the population along the transport corridors that link Mozambique and the port of Beira to Zimbabwe and Malawi has been identified as an important factor for higher prevalence of AIDS rates in the central region of Mozambique, in which the poor women have been identified as the most vulnerable segment

Table 3: Profiling of the latent classes.

Variables	Latent classes				Aggregate
	1	2	3	4	
Region					
Niassa	0.075	0.065	0.046	0.044	0.066
Cabo Delgado	0.085	0.091	0.026	0.017	0.072
Nampula	0.109	0.126	0.047	0.032	0.098
Zambezia	0.096	0.143	0.024	0.029	0.091
Tete	0.109	0.086	0.051	0.047	0.090
Manica	0.104	0.084	0.065	0.034	0.088
Sofala	0.121	0.082	0.061	0.064	0.098
Inhambane	0.097	0.094	0.080	0.058	0.091
Gaza	0.100	0.122	0.098	0.073	0.103
Maputo Província	0.063	0.069	0.199	0.147	0.091
Maputo Cidade	0.042	0.038	0.303	0.456	0.112
Place of residence					
Capital, large city	0.042	0.038	0.303	0.456	0.112
Small city	0.187	0.205	0.422	0.380	0.241
Town	0.079	0.082	0.088	0.065	0.080
Countryside	0.691	0.674	0.187	0.099	0.567
Age group					
15-19	0.186	0.184	0.305	0.316	0.213
20-24	0.195	0.188	0.222	0.238	0.201
25-29	0.181	0.179	0.155	0.150	0.174
30-34	0.141	0.140	0.110	0.109	0.134
35-39	0.117	0.122	0.085	0.092	0.112
40-44	0.098	0.102	0.078	0.060	0.093
45-49	0.081	0.087	0.047	0.035	0.074
Highest educational level					
No education	0.455	0.420	0.102	0.032	0.362
Primary	0.519	0.541	0.661	0.469	0.541
Secondary	0.026	0.039	0.233	0.465	0.094
Higher	0.000	0.000	0.004	0.034	0.003
Current marital status					
Single	0.123	0.137	0.335	0.441	0.182
Partnership	0.865	0.850	0.657	0.546	0.806
Other	0.012	0.014	0.008	0.013	0.012

(Agadjanian, 2005). By applying latent class modeling to Mozambique DHS data, we show that there are four segments or groups of women with quite heterogeneous levels of access to information on how to avoid being contaminated by HIV/AIDS.

Our findings clearly point out to the importance of tailoring the means of communication according with the available media. For example, in rural areas radio is more likely to be the most effective way of disseminating messages on how to avoid HIV contamination. In synthesis, latent class models identify different subpopulations with different characteristics which should be reached using different strategies in a social marketing approach (Kotler and Zaltman, 1971; Wedel and Kamakura, 2000).

Acknowledgements: This research was supported by the Fundação para a Ciência e a Tecnologia, grant PTDC/CS-DEM/108033/2008.

References:

[1] V. Agadjanian, Gender, religious involvement, and HIV/AIDS prevention in Mozambique, *Social Science & Medicine* 61, pp. 1529-1539, 2005

[2] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *Journal of the Royal Statistical Society B* 39, pp. 1-38, 1977

[3] J.G. Dias, *Finite Mixture Models. Review, Applications, and Computer-intensive Methods*, PhD Thesis, University of Groningen, The Netherlands, 2004

[4] J.G. Dias, F. Willekens, Model-based clustering of sequential data with an application to contraceptive use dynamics, *Mathematical Population Studies* 12, 135-157, 2005

[5] J.G. Dias, Model selection for the binary latent class model. A Monte Carlo simulation, In: *Data Science and Classification*, V. Batagelj, H.-H. Bock, A. Ferligoj, A. Ziberna (eds.), Springer, Berlin, pp. 91-99, 2006

[6] P. Kotler, G. Zaltman, Social marketing: An approach to planned social change, *Journal of Marketing* 35 (July), pp. 3-12, 1971

[7] G.J. McLachlan, D. Peel, *Finite mixture models*, John Wiley and Sons, New York, 2000

[8] P. Piot, F. Plummer, F. Mhalu, J.L. Lamboray, J. Chin, J. Mann, AIDS: An international perspective, *Science* 239, pp. 573-579, 1988

[9] A. Skrondal and S. Rabe-Hesketh, *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Chapman & Hall/CRC, Boca Raton, FL, 2004

[10] G. Schwarz, Estimating the dimension of a model, *Annals of Statistics* 6, 461-464, 1978

[11] M. Wedel, W. Kamakura, *Market segmentation. Conceptual and methodological foundations*, Kluwer Academic Publishers, Second Edition, Boston, 2000

[12] World Bank, *Confronting AIDS: Public priorities in a global epidemic*, A World Bank Policy Research Report, Oxford University Press, 1997