English Longitudinal Study Data : CLUSTERING OF HEATH BEHAVIOURS

4. SMOKING

# SUMMARY

* **Data Extraction and Cleaning in R :** The English Longitudinal Study dataset was accessed using the UK Data Archive. From this, data smoking was extracted and cleaned using R. Smoking data was categorised into two levels as per current smoking status : smoker, non-smoker (for details, see Section 1.2).
* **Latent Class Analysis in MPlus :** Latent class analysis (LCA) was used to estimate latent classes/subgroups based on participants’ smoking trajectories between Waves 4 to 9. LCA uses maximum likelihood estimation for modelling the data. This method does not involve any imputations for missing data and treats all missing values as part of the model, under the assumptions of missing completely at random and missing at random. This also holds true for cases where missingness is a function of other observed variable or covariates.
* **Latent Class Analysis in MPlus using weighted data :** LCA was conducted using wave 9 weights, that calculated weights for 4,848 core members who responded to all six waves since Wave 4.

**Note : In the ELSA dataset, questions on health behaviours were asked to all participants and no exclusions were made based on any given criteria.**

# 1. Data Cleaning in R

The root folder containing all the data in the document is **‘ELSA\_Project’**

The main file directory for the data processed in this project is **‘ELSA\_Project.Proj’**. When using R, start here.

All data on smoking is stored in the folder **‘Data/Processed\_Data/R\_Data/Smoking’** and all R scripts for processing data on smoking are stored in the **‘Scripts/Smoking’** folder.

## Load Data files from ELSA Waves 3-9

Files accessed:

* Wave 4 : **‘wave\_4\_elsa\_data\_v3.tab’**
* Wave 5 : **‘wave\_5\_elsa\_data\_v4.tab’**
* Wave 6 : **‘wave\_6\_elsa\_data\_v2.tab’**
* Wave 7 : **‘wave\_7\_elsa\_data.tab’**
* Wave 8 : **‘wave\_8\_elsa\_data\_eul\_v2.tab’**
* Wave 9 : **‘wnutrition\_elsa\_wave9\_id.tab**

Smoking data was extracted from the aforementioned files.

Code for importing and cleaning smoking data from Waves 3-9 of the English Longitudinal Study can be found in the RScript file **‘01\_ Importing\_Smoking\_Data.R’**

## CATEGORISING extracted data in R

No recoding was required since the relevant survey questions on smoking were the same across all included Waves.

The final categorisation was based on the answers to one survey question :

1. Do you smoke cigarettes nowadays?

The final categorisation (based on current smoking status):

o Non-smoker

o Current smoker

o Missing

Code for categorising smoking data from Waves 4-9 of the English Longitudinal Study can be found in the RScript file **‘02\_ Categorising\_Smoking.R’**.

.

.

# Smoking- descriptive data

Summary of smoking data can be found in the folder named **‘Figures/Smoking’.**

The R script used to produce these results, namely, **“03\_Descriptives\_Smoking.R”** can be found in the **“Scripts”** folder.

# Weighted latent class analysis

Longitudinal weights were included to perform a weighted latent class analysis. These weights are defined for the subset of cases who have taken part in all waves, up to and including the wave in question. At each wave, the fully responding core members were re-weighted to take account of respondents at the previous wave that were lost through refusal at the current wave or through some other form of sample attrition. Core members from the original sample who returned to the study having missed a wave therefore have no longitudinal weight.

In the present case, the weight **W9W4LWGT** calculated at Wave 9 was chosen. This was calculated for the set of 4,848 core members who had responded to all six waves since Wave 4, and remain living in private households. A alternative option was to include weights for approximately 2200 participants who had responded to all waves (i.e from Wave 1-9), and remain living in private households. However, the second option was dropped in favour of the first since the first would allow us to retain information on nearly double the sample size.

## . CRITERIa FOR CHOOSING LATENT CLASS MODEL

A 4-class latent model was chosen based on the following criteria (in no order of ranking) :

**1. Lo-Mendell Rubin Adjusted likelihood ratio test :** A p value <.05 on this test suggests that a simpler model with k – 1 classes can be rejected in favour of the estimated model with k-classes. As seen in Table 1, this criterion is significant till the fifth any models and would suggest a four-class model as best fit.

**2. Akaike information criterion (AIC) :** A smaller AIC value indicates better model estimation. However, as can be seen in Table 1, AIC continues to decrease up until the 5-class model, however the decrease between 5-class and 4-class models is very subtle, indicating both models are likely to fit better than a 3-class model.

**3. Bayesian information criterion (BIC) :** A smaller BIC value indicates better model estimation. However, as can be seen in Table 1, BIC continues to decrease up until the 4-class model, indicating that the 4-class model is likely to fit better than a 5-class model.

Based on these criteria, the 4-class appear to be superior models compared to other models. Equally, a 4-class was chosen was easily interpretable.

*Table 3.* Criteria used for selection of Latent Class Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **LCA Models** | **Lo-Mendell Rubin Adjusted likelihood ratio test** | **AIC** | **BIC**  **(Sample size Adjusted BIC)** | **Entropy** |
| 2-class | **p<.0001** | 6529.639 | 6611.432 (6570.124) | 0.961 |
| 3-class | **p<.0001** | 6160.979 | 6286.815 (6223.264) | 0.950 |
| 4-class | **p = 0.0029** | 6088.619 | 6258.497 (6172.703) | 0.962 |
| 5-class | p = 0.3511 6073.304 6287.225 (6179.188) 0.963 | | | |

## Interpreting the selected 3-class model

In addition to the aforementioned criteria, the selected 3-class model also had the following advantages :

**1. Entropy :** Entropy is used to assess model quality. A high entropy value (> 0.8) suggests that there is a high probability that a subject (e.g., person) will fall in only one of the classes. In the present case, entropy was very high (0.984).

**2. Homogeneity and Class separation :** Once the aforementioned criteria were assessed, the 4-class model was also examined for ease of interpretability. Models with high class separation (i.e. the degree to which latent classes can clearly be distinguished from each other ) and high homogeneity (i.e. how similar individuals in a class are to each other or the degree to which response probabilities are close to 0 and 1 in each class) are generally preferred.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sample proportion (%)** | **Avg. no. of cigarettes smoked per day** | **Wave 4**  **(2008-09)** | **Wave 5**  **(2010-11)** | **Wave 6**  **(2012-13)** | **Wave 7**  **(2014-15)** | **Wave 8**  **(2016-17)** | **Wave 9**  **(2018-19)** |
| **Latent Class 1**  Gradual quitters | 3.2 | **Non-Smoker** | .098  .156 | .055  .156 | .389  .156 | **.622**  **.156** | **.953**  **.156** | **.966**  **.156** |
|  | **Smoker** | **.902** | **.945** | **.611** | .378 | .047 | .034 |
| **Latent Class 2**  Gradual initiators | .7 | **Non-Smoker** | **.716**  **.156** | **.517**  **.156** | **.706**  **.156** | **.598**  **.156** | .306  .156 | .276  .156 |
|  | **Smoker** | .284 | .483 | .294 | .402 | **.694** | **.724** |
| **Latent Class 3**  Smokers | 8.3 | **Non-Smoker** | .000  .156 | .068  .156 | .043  .156 | .022  .156 | .056  .156 | .150  .156 |
|  | **Smoker** | **1.000** | **.932** | **.957** | **.978** | **.944** | **.850** |
| **Latent Class 4**  Non-smoker | 87.7 | **Non-smoker** | **.987** | **.995** | **.996** | **.999** | **.998** | **.999** |
|  |  | **Smoker** | 0.13 | .005 | .004 | .001 | .002 | .001 |

*Table 4.* Estimated probabilities and sample proportions of the selected latent 3-Class Mode