English Longitudinal Study Data : CLUSTERING OF HEATH BEHAVIOURS

1. NUTRITION (FRUIT AND VEGETABLE INTAKE)

# SUMMARY

* **Data Extraction and Cleaning in R :** The English Longitudinal Study dataset was accessed using the UK Data Archive. From this, data on fruit and vegetable (FV) intake was extracted and cleaned using R. First, FV intake data was extracted from ELSA Waves 3-9 as no FV data was collected in Waves 1 and 2. Next, the data collected in different waves was re-coded to ensure cross-wave symmetry. Since coding for Waves 3 and 4 was different from Waves 5-9, data from the former Waves was recoded to standardise FV intake data across waves (details provided in Section 2.1) . Finally, FV intake data in each Wave was categorised to five levels : (≥0–2.5, ≥2.5–5, ≥5–7.5, ≥7.5–10, and >10 portions of 80 g per day). For a summary of the FV intake data, refer to the file **‘Descriptives\_Fruit\_Veg.doc’**.
* **Latent Class Analysis in MPlus :** Latent class analysis (LCA) was used to estimate latent classes/subgroups based on participants’ FV trajectories between Waves 3 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. This also holds true for cases where missingness is a function of observed outcomes 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\_Data’**

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

All data on fruit and vegetable intake is stored in the folder **“Fruit\_Veg\_Intake”** and all R scripts for processing data on fruit and vegetable intake are stored in the **‘Scripts’** folder.

## Load Data files from ELSA Waves 3-8

Files accessed:

* Wave 3 : **‘wave\_3\_elsa\_data\_v4.tab’**
* 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**

Fruit and vegetable intake data was extracted from the aforementioned files.

# Recoding extracted data in R

At Waves 3 and 4, diet was assessed with items such as ‘How much of the following did you eat yesterday’ with answers including ‘Average slices of a very large fruit, such as melon’ and ‘Salad (cereal bowlfuls)’. In Waves 5–8 participants were asked ‘how many portions of fruit/vegetables do you eat on a given day’. In Wave 9, a separate and detailed module was conducted on nutrition, from which variables describing portions of fruit/vegetables eaten on a given day had been derived and made available. Code for importing and cleaning data on fruit and vegetable intake from Waves 3-9 of the English Longitudinal Study can be found in the RScript file **‘1. Importing\_Recoding\_FV\_Script.R’**.

## MAKING FRUIT/VEG INTAKE DATA CONSISTENT ACROSS WAVES

To make fruit and vegetable intake more consistent across all waves, answers from Waves 3 and 4 were recoded to correspond with the instructions from Waves 5 and 6 such that one portion approximates one serving size. Conversion rates are shown in figure 1. This conversion method has been previously used by Hackett et .al (2018) and Kojima et. al (2020) to conduct longitudinal analyses on fruit and vegetable intake from the ELSA study.

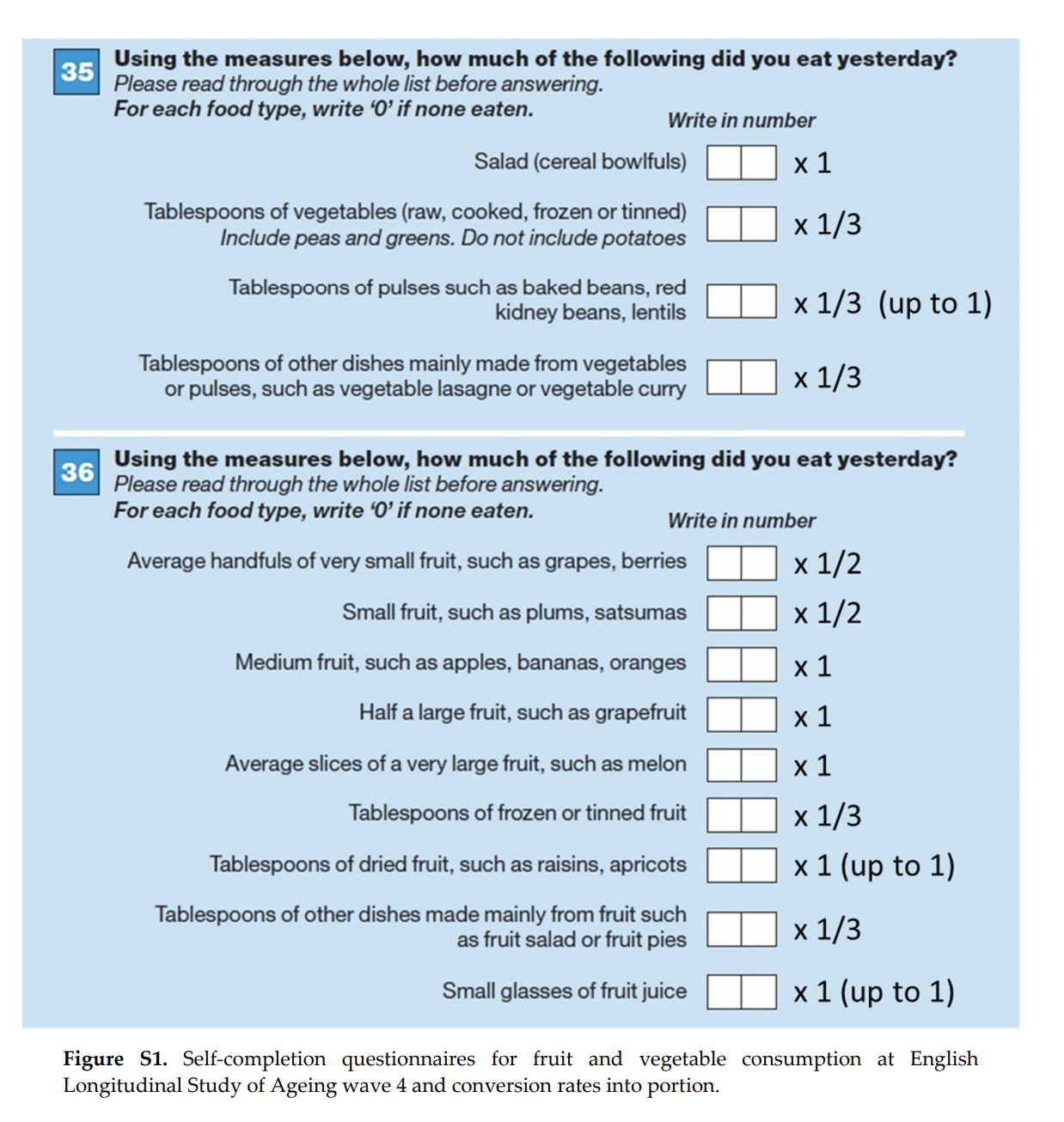
“*The amount of consumed fruit and vegetables were converted into portions (1 portion = 80 g for both vegetables and fruit) in accordance with the Welsh Health Survey methodology and the “5 A Day” campaign portion size from the National Health Service (NHS).*” Kojima et. al (2020, p.3)

Figure 1. Self-completion questionnaires for fruit and vegetable intake at ELSA Wave 4 and conversion rates into portion, as seen in Kojima et. al (2020)

## removing outliers

The number of portions of fruits and vegetables reported during each wave were added together to create one variable that represented the total number of fruit and vegetable portions (i.e., servings) eaten during one day. Because there were some outliers for the total number of daily fruit and vegetables, attempts were made to reduce the skew. Outliers were defined as values more than mean value plus 3 standard deviations for each fruit and vegetable consumption, and those with at least one implausibly high value were excluded from analysis **(n = 276).** Code for this step can be found in R script ‘**2. Removing\_Outliers\_FV\_Script.R’**.

## CATEGORISING FRUIT/VEG DATA

The number of portions of fruits and vegetables reported during each wave were added together to create one overall variable that represented the total number of fruit and vegetable portions (i.e., servings) eaten in a day. Subsequently, this data was divided into five groups (≥0–2.5, ≥2.5–5, ≥5–7.5, ≥7.5–10, and >10 portions of 80 g per day). The cut-points were chosen based on the “5 A Day” campaign and the recent findings on beneficial effects of consuming more than 5 portions (7.5 and 10 portions) of fruit and vegetables (Kojima et. al, 2018). Code for this step can be found in R script **‘3. Categorising\_FV\_Script.R’**.

# fruit/veg intake descriptive data

Summary of fruit and vegetable intake data can be found in the document named **‘Descriptives\_Fruit\_Veg.doc’.**

The R script used to produce these results, namely, **“4. Descriptive\_Graphs\_Script.R”** can be found in the **“Scripts”** folder.

# LCA Analysis Using MPlus

To prepare the data for analysis, the headers from **‘3. Categorised\_FruitVeg\_3to9.csv’** were removed using Excel. This file was saved as **‘Weighted\_Categorised\_FV\_3to9\_Mplus.csv’** and used for analysis in MPlus.

## 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 for all models except a 6-class model and does not serve as a clear indicator for model selection.

**2. Bootstrap Likelihood ratio test :** This test is interpreted the same way as Lo-Mendell Rubin Adjusted likelihood ratio test. However, in the present case, this test does not serve as a clear indicator for the best-fitting model among those examined.

**3. 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 6-class model.

**4. Bayesian information criterion (BIC) :** A smaller BIC value indicates better model estimation. In the present case, BIC is clearly the smallest for a 4-class model.

*Table 1.* Criteria used for selection of Latent Class Model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **LCA Models** | **Lo-Mendell Rubin Adjusted likelihood ratio test** | **Bootstrap Parametric Likelihood ratio test** | **AIC** | **BIC**  **(Sample size Adjusted BIC)** | **Entropy** |
| 2-class | **p<.00001\*** | **p<.00001\*** | 32999.695 | 33319.823 (33138.730) | 0.792 |
| 3-class | **p<.00001\*** | **p<.00001\*** | 32348.834 | 32831.834 (32558.607) | 0.758 |
| 4-class | **p<.00001\*** | **p<.00001\*** | 32105.718 | 32751.590 (32386.227) | 0.757 |
| 5-class | **p=.0066\*** | **p<.00001\*** | 32017.405 | 32826.150 (32368.652) | 0.699 |
| 6-class | **p=0.3522** | **p<.00001\*** | 32004.110 | 32975.727 (32426.094) | 0.663 |

Note: \* is used to denote p-values <.05.

## Interpreting the selected 4-class model

In addition to the aforementioned criteria, the selected 4-class model also had the following advantage :

**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. Entropy in the selected 4-class model was 0.757.

**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.

**Observations/cases inputted (n) =** 2031

*Table 2.* Estimated probabilities and sample proportions of the selected latent 4-Class Model

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **%** | **Portions of Fruit/Veg intake per day (1 portion = 80g/day)** | **Wave 3**  **(2006-07)** | **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**  Low-adequate F/V intake | 35.56 | **≥0–2.5** | .124 | .135 | .049 | .052 | .043 | .045 | .332 |
|  | **≥2.5–5** | .418 | .435 | **.642** | **.572** | **.542** | **.566** | .440 |
|  | **≥5–7.5** | .340 | .329 | .295 | .358 | .402 | .364 | .159 |
|  | **≥7.5–10** | .075 | .081 | .013 | .007 | .010 | .020 | .050 |
|  | **>10** | .043 | .020 | .000 | .010 | .003 | .005 | .019 |
| **Latent Class 2**  Very low F/V intake | 12.51 | **≥0–2.5** | .427 | .442 | **.564** | **.580** | **.558** | **.549** | **.666** |
|  | **≥2.5–5** | .353 | .361 | .403 | .367 | .360 | .348 | .260 |
|  | **≥5–7.5** | .153 | .137 | .033 | .050 | .070 | .099 | .058 |
|  | **≥7.5–10** | .062 | .049 | .000 | .003 | .012 | .004 | .016 |
|  | **>10** | .006 | .011 | .000 | .000 | .000 | .000 | .000 |
| **Latent Class 3** | 42.39 | **≥0–2.5** | .048 | .041 | .005 | .003 | .003 | .003 | .150 |
|  |  | **≥2.5–5** | .244 | .247 | .125 | .078 | .095 | .102 | .382 |
| Adequate-high F/V intake |  | **≥5–7.5** | .429 | .409 | **.743** | **.812** | **.769** | **.767** | .317 |
|  |  | **≥7.5–10** | .201 | .233 | .106 | .094 | .107 | .109 | .116 |
|  |  | **>10** | .078 | .070 | .020 | .012 | .027 | .019 | .036 |
| **Latent Class 4**  High-Very High F/V intake | 9.53 | **≥0–2.5** | .026 | .000 | .005 | .010 | .000 | .001 | .000 |
|  | **≥2.5–5** | .132 | .104 | .050 | .026 | .000 | .034 | .249 |
|  | **≥5–7.5** | .348 | .280 | .480 | .327 | .315 | .302 | .275 |
|  | **≥7.5–10** | .339 | .436 | .305 | .403 | .496 | .408 | .261 |
|  | **>10** | .155 | .180 | .160 | .235 | .188 | .256 | .215 |

# Weighted latent class analysis – COMPLETE CASE 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. Moreover, fruit and vegetable intake data relevant to the current analysis was only assessed Wave 3 onwards.

## . 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 for all models except a 5-class model, where the unadjusted LRT test is significant and does not serve as a reliable criterion for model selection.

**2. Bootstrap Likelihood ratio test :** This test is not calculated for weighted data.

**3. 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.

**4. Bayesian information criterion (BIC) :** A smaller BIC value indicates better model estimation. In the present case, BIC is clearly the smallest for a 4-class model.

*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<.00001\*** | 27929.554 | 28204.752 (28049.076) | .800 |
| 3-class | **p<.00001\*** | 27332.145 | 27747.750 (27512.647) | 0.764 |
| 4-class | **p<.00001\*** | 27080.554 | 27636.566 (27322.036) | 0.755 |
| 5-class | p= 0.8602 (unadjusted value used since adjusted test did not return a value) | 27062.546 | 27758.965 (27365.009) | 0.767 |

Note: \* is used to denote p-values <.05.

## Interpreting the selected 4-class model

In addition to the aforementioned criteria, the selected 4-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 not used as a prime indicator for model-fit.

**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.

**Observations/cases inputted (n) =** 2031

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **%** | **Portions of Fruit/Veg intake per day (1 portion = 80g/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**  Low-adequate F/V intake | 35.29 | **≥0–2.5** | .156 | .066 | .076 | .067 | .068 | .369 |
|  | **≥2.5–5** | .432 | **.694** | **.602** | **.561** | **.589** | .423 |
|  | **≥5–7.5** | .311 | .232 | .302 | .360 | .322 | .135 |
|  | **≥7.5–10** | .078 | .008 | .009 | .010 | .018 | .054 |
|  | **>10** | .023 | .000 | .011 | .002 | .004 | .019 |
| **Latent Class 2**  Very low F/V intake | 11.42 | **≥0–2.5** | .440 | **.653** | **.681** | **.625** | **.638** | **.709** |
|  | **≥2.5–5** | .377 | .330 | .284 | .341 | .300 | .231 |
|  | **≥5–7.5** | .136 | .017 | .034 | .029 | .062 | .051 |
|  | **≥7.5–10** | .037 | .000 | .000 | .005 | .000 | .009 |
|  | **>10** | .011 | .000 | .000 | .000 | .000 | .000 |
| **Latent Class 3**  Adequate-high F/V | 41.34 | **≥0–2.5** | .059 | .006 | .000 | .000 | .002 | .157 |
|  | **≥2.5–5** | .257 | .147 | .097 | .129 | .129 | .407 |
|  | **≥5–7.5** | .389 | **.735** | **.827** | **.766** | **.745** | .311 |
|  | **≥7.5–10** | .226 | .093 | .075 | .084 | .109 | .097 |
|  | **>10** | .069 | .069 | .001 | .020 | .015 | .028 |
| **Latent Class 4**  High-Very High F/V intake | 11.93 | **≥0–2.5** | .001 | .008 | .030 | .000 | .000 | .048 |
|  | **≥2.5–5** | .160 | .039 | .030 | .000 | 0.65 | .260 |
|  | **≥5–7.5** | .339 | .534 | .323 | .355 | .407 | .231 |
|  | **≥7.5–10** | .361 | .288 | .399 | .482 | .348 | .247 |
|  | **>10** | .140 | .131 | .219 | .163 | .180 | .215 |

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

**References**

Hackett, R. A., Moore, C., Steptoe, A., & Lassale, C. (2018). Health behaviour changes after type 2 diabetes diagnosis: Findings from the English Longitudinal Study of Ageing. *Scientific reports*, *8*(1), 1-8.

Kojima, G., Avgerinou, C., Iliffe, S., Jivraj, S., Sekiguchi, K., & Walters, K. (2018). Fruit and vegetable consumption and frailty: a systematic review. *The Journal of Nutrition, Health & Aging*, *22*(8), 1010-1017.

Kojima, G., Iliffe, S., Jivraj, S., & Walters, K. (2020). Fruit and Vegetable Consumption and Incident Prefrailty and Frailty in Community-Dwelling Older People: The English Longitudinal Study of Ageing. *Nutrients*, *12*(12), 3882.

# Weighted latent class analysis – ALL CASE 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. Moreover, fruit and vegetable intake data relevant to the current analysis was only assessed Wave 3 onwards.

## . 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 for all models except a 5-class model, where the unadjusted LRT test is significant and does not serve as a reliable criterion for model selection.

**2. Bootstrap Likelihood ratio test :** This test is not calculated for weighted data.

**3. 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.

**4. Bayesian information criterion (BIC) :** A smaller BIC value indicates better model estimation. In the present case, BIC is clearly the smallest for a 4-class model.

*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<.00001\*** | 183740.472 | 184201.079 (184007.227) | 1 |
| 3-class | **p<.00001\*** | 169730.251 | 170424.938 (170132.570) | 0.860 |
| 4-class | **p<.00001\*** | 166282.045 | 167210.811 (166819.928) | 0.801 |
| 5-class | p= 0.0020 (unadjusted value used since adjusted test did not return a value) | 164670.408 | 165833.254 (165343.856) | 0.782 |

Note: \* is used to denote p-values <.05.

## Interpreting the selected 4-class model

In addition to the aforementioned criteria, the selected 4-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 not used as a prime indicator for model-fit.

**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.