Latent class analysis with grocery store's shoppers data

110077443_Brendon Pedro

3/10/2022

Please note that all code in this document is presented in a grey box and the output reflected below each box

- The below code allows lengthy lines of code to display neatly within the grey box (wrapping it)

```
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

Import Data

```
Grocery <- read.csv("Grocery.csv", header = TRUE)
```

Get an overview of the variables

```
names(Grocery)
```

```
##  [1] "ï..Beverage"       "Frozen.Pizza"      "Facial.Tissue"
##  [4] "Laundry.Detergent" "Shampoo"           "Soup"
##  [7] "Spaghetti.sauce"   "Sugar"             "Peanut.Butter"
## [10] "Beer"              "Milk"              "Yogurt"
## [13] "Income"            "age"               "trans"
```

```
# From 'Beverage' to 'Yogurt' are dummies that indicate
# whether or not a customer purchases such a grocery
# store's product Income: Customer income Age: Customer's
# age Trans: we don't have information what 'trans' is
# about so we won't use it

names(Grocery)[1] <- "Beverage"
names(Grocery)
```

```
##  [1] "Beverage"          "Frozen.Pizza"      "Facial.Tissue"
##  [4] "Laundry.Detergent" "Shampoo"           "Soup"
##  [7] "Spaghetti.sauce"   "Sugar"             "Peanut.Butter"
## [10] "Beer"              "Milk"              "Yogurt"
## [13] "Income"            "age"               "trans"
```

Step 1: Determine research goal:

- We want to identify different customer segments that are useful to customize offerings for grocery store's products

Step 2: we choose variables that align with the research goal (see lecture) and prepare data

```
# check variable type
str(Grocery)
```

```
## 'data.frame':    9800 obs. of  15 variables:
##  $ Beverage         : int  1 1 1 1 0 1 0 1 1 1 ...
##  $ Frozen.Pizza     : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ Facial.Tissue    : int  1 0 1 0 1 1 1 0 0 0 ...
##  $ Laundry.Detergent: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Shampoo          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Soup             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Spaghetti.sauce  : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ Sugar            : int  0 1 0 0 0 0 0 0 1 0 ...
##  $ Peanut.Butter    : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Beer             : int  0 0 1 0 0 1 0 0 0 0 ...
##  $ Milk             : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ Yogurt           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Income           : int  1000 0 425 1302 0 2365 0 0 0 0 ...
##  $ age              : int  65 31 21 62 36 41 46 27 34 67 ...
##  $ trans            : int  83 55 12 0 0 17 0 0 0 50 ...
```

```
# We see that from Beverage to Yogurt are integer, but they
# should be factor (dummy variable)

# correct variable type
Grocery[1:12] <- lapply(Grocery[1:12], as.factor)
str(Grocery)
```

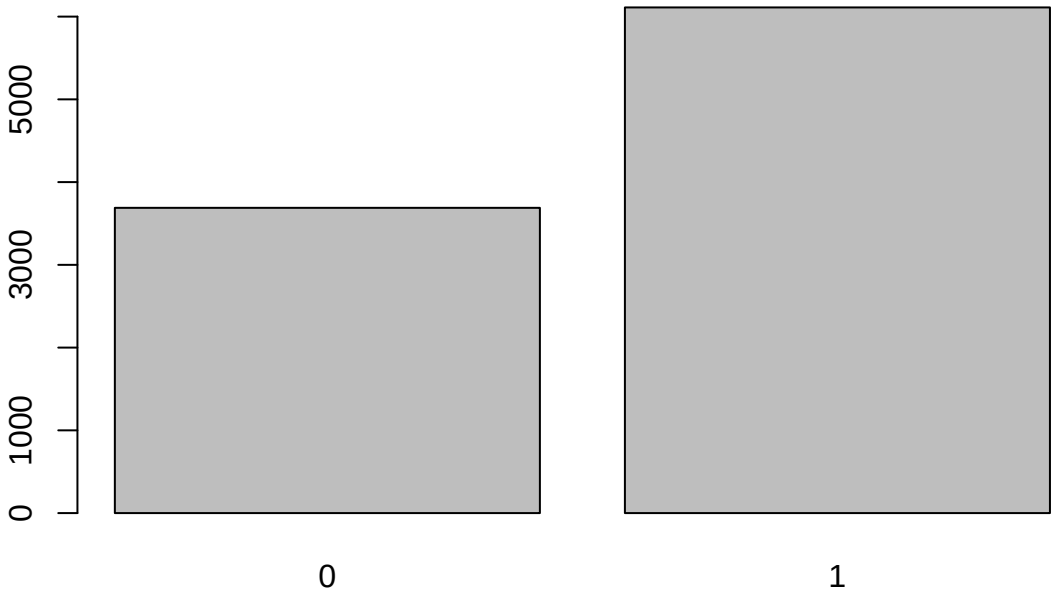
```
## 'data.frame':    9800 obs. of  15 variables:
##  $ Beverage         : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 1 2 2 2 ...
##  $ Frozen.Pizza     : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
##  $ Facial.Tissue    : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 1 1 1 ...
##  $ Laundry.Detergent: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Shampoo          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Soup             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Spaghetti.sauce  : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
##  $ Sugar            : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 2 1 ...
##  $ Peanut.Butter    : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
##  $ Beer             : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 1 1 1 ...
##  $ Milk             : Factor w/ 2 levels "0","1": 1 1 2 1 1 1 1 1 1 1 ...
##  $ Yogurt           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Income           : int  1000 0 425 1302 0 2365 0 0 0 0 ...
##  $ age              : int  65 31 21 62 36 41 46 27 34 67 ...
##  $ trans            : int  83 55 12 0 0 17 0 0 0 50 ...
```

- Overwiew Variables

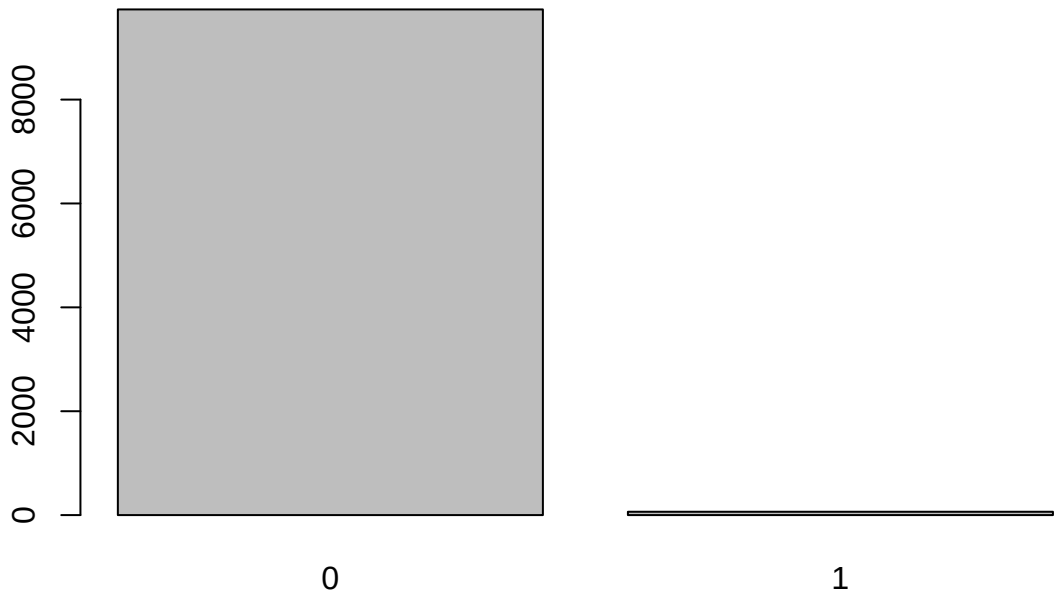
```
summary(Grocery)
```

```
##  Beverage Frozen.Pizza Facial.Tissue Laundry.Detergent Shampoo  Soup
##  0:3689   0:9736       0:5375        0:9526            0:9536   0:9798
##  1:6111   1:  64       1:4425        1: 274            1: 264   1:   2
##
##
##
##
##  Spaghetti.sauce Sugar    Peanut.Butter Beer     Milk     Yogurt
##  0:9557          0:9400   0:8980        0:9027   0:9565   0:8452
##  1: 243          1: 400   1: 820        1: 773   1: 235   1:1348
##
##
##
##
##      Income            age             trans
##  Min.   :    0   Min.   : 0.00   Min.   :  0.00
##  1st Qu.:    0   1st Qu.:28.00   1st Qu.:  0.00
##  Median :    0   Median :41.00   Median :  1.00
##  Mean   : 1035   Mean   :42.91   Mean   : 26.42
##  3rd Qu.: 1634   3rd Qu.:57.00   3rd Qu.: 44.00
##  Max.   :47074   Max.   :98.00   Max.   :282.00
```

```
plot(Grocery$Beverage)
```




```
plot(Grocery$Frozen.Pizza)
```



- Create data frame for cluster analysis that contains the variables to be used in LCA

```
cluster.lca.df <- data.frame(Grocery$Beverage, Grocery$Frozen.Pizza,
    Grocery$Facial.Tissue, Grocery$Laundry.Detergent, Grocery$Shampoo,
    Grocery$Soup, Grocery$Sugar, Grocery$Peanut.Butter)
summary(cluster.lca.df)
```

```
##  Grocery.Beverage Grocery.Frozen.Pizza Grocery.Facial.Tissue
##  0:3689           0:9736               0:5375
##  1:6111           1:  64               1:4425
##  Grocery.Laundry.Detergent Grocery.Shampoo Grocery.Soup Grocery.Sugar
##  0:9526                    0:9536          0:9798       0:9400
##  1: 274                    1: 264          1:   2       1: 400
##  Grocery.Peanut.Butter
##  0:8980
##  1: 820
```

```
str(cluster.lca.df)  # LCA required binary variables. All variables are factors/binary variables so goo
```

```
## 'data.frame':    9800 obs. of  8 variables:
##  $ Grocery.Beverage         : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 1 2 2 2 ...
##  $ Grocery.Frozen.Pizza     : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
##  $ Grocery.Facial.Tissue    : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 1 1 1 ...
##  $ Grocery.Laundry.Detergent: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ Grocery.Shampoo          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Grocery.Soup             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Grocery.Sugar            : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 2 1 ...
##  $ Grocery.Peanut.Butter    : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
```

Step 3: Latent Class Analysis

```
# Call required packages to perform LCA
library(poLCA)
```

```
## Loading required package: scatterplot3d
```

```
## Loading required package: MASS
```

```
library(scatterplot3d)
library(MASS)
```

- Define the underlying model:

```
model <- with(cluster.lca.df, cbind(Grocery.Beverage, Grocery.Frozen.Pizza,
    Grocery.Facial.Tissue, Grocery.Laundry.Detergent, Grocery.Shampoo,
    Grocery.Soup, Grocery.Sugar, Grocery.Peanut.Butter)) ~ 1
```

```
# we used the with function, so we didn't have to call the
# data before using the variables Basically, the function
# defines that the variables in cbind () are explained by
# an intercept only. We vary the number of factors next.

#-----------------------------------------------------------
# The poLCA() function uses a formula interface to
# determine which items are included in the model (see
# package description).  The PoLCA function runs the LCA
# for predetermined numbers of classes.
```

```
seg.lca.3 <- poLCA(model, data = cluster.lca.df, nclass = 3,
    na.rm = TRUE)
```

```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $Grocery.Beverage
##                 0      1
## class 1:  0.6085 0.3915
## class 2:  0.1577 0.8423
## class 3:  0.0000 1.0000
##
## $Grocery.Frozen.Pizza
##                 0      1
```

```
## class 1:  1.0000 0.0000
## class 2:  0.9951 0.0049
## class 3:  0.9728 0.0272
##
## $Grocery.Facial.Tissue
##               0      1
## class 1:  0.4860 0.5140
## class 2:  0.9996 0.0004
## class 3:  0.1421 0.8579
##
## $Grocery.Laundry.Detergent
##               0      1
## class 1:  1.0000 0.0000
## class 2:  0.9945 0.0055
## class 3:  0.8635 0.1365
##
## $Grocery.Shampoo
##               0      1
## class 1:  0.9762 0.0238
## class 2:  1.0000 0.0000
## class 3:  0.9293 0.0707
##
## $Grocery.Soup
##               0     1
## class 1:  1.0000 0e+00
## class 2:  0.9996 4e-04
## class 3:  0.9994 6e-04
##
## $Grocery.Sugar
##               0      1
## class 1:  1.0000 0.0000
## class 2:  0.9635 0.0365
## class 3:  0.8376 0.1624
##
## $Grocery.Peanut.Butter
##               0      1
## class 1:  0.9951 0.0049
## class 2:  0.9150 0.0850
## class 3:  0.6941 0.3059
##
## Estimated class population shares
##  0.5533 0.252 0.1947
##
## Predicted class memberships (by modal posterior prob.)
##  0.556 0.3393 0.1047
##
## =========================================================
## Fit for 3 latent classes:
## =========================================================
## number of observations: 9800
## number of estimated parameters: 26
## residual degrees of freedom: 229
## maximum log-likelihood: -19705.47
##
```

```
## AIC(3): 39462.94
## BIC(3): 39649.88
## G^2(3): 217.2594 (Likelihood ratio/deviance statistic)
## X^2(3): 256.1973 (Chi-square goodness of fit)
##
## ALERT: iterations finished, MAXIMUM LIKELIHOOD NOT FOUND
##
```

```
seg.lca.4 <- poLCA(model, data = cluster.lca.df, nclass = 4,
    na.rm = TRUE)
```

```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $Grocery.Beverage
##                0      1
## class 1:  0.0000 1.0000
## class 2:  0.2868 0.7132
## class 3:  0.0000 1.0000
## class 4:  0.5924 0.4076
##
## $Grocery.Frozen.Pizza
##                0      1
## class 1:  0.9888 0.0112
## class 2:  0.9966 0.0034
## class 3:  0.9410 0.0590
## class 4:  1.0000 0.0000
##
## $Grocery.Facial.Tissue
##                0      1
## class 1:  0.3270 0.6730
## class 2:  0.9876 0.0124
## class 3:  0.1018 0.8982
## class 4:  0.4115 0.5885
##
## $Grocery.Laundry.Detergent
##                0      1
## class 1:  0.9581 0.0419
## class 2:  0.9969 0.0031
## class 3:  0.6820 0.3180
## class 4:  1.0000 0.0000
##
## $Grocery.Shampoo
##                0      1
## class 1:  1.0000 0.0000
## class 2:  1.0000 0.0000
## class 3:  0.8016 0.1984
## class 4:  0.9718 0.0282
##
## $Grocery.Soup
##                0     1
## class 1:  0.9991 9e-04
## class 2:  0.9998 2e-04
## class 3:  1.0000 0e+00
```

```
## class 4:  1.0000 0e+00
##
## $Grocery.Sugar
##                0      1
## class 1:  0.7502 0.2498
## class 2:  0.9969 0.0031
## class 3:  0.9512 0.0488
## class 4:  1.0000 0.0000
##
## $Grocery.Peanut.Butter
##                0      1
## class 1:  0.6393 0.3607
## class 2:  0.9507 0.0493
## class 3:  0.7798 0.2202
## class 4:  0.9966 0.0034
##
## Estimated class population shares
##  0.1469 0.2946 0.0657 0.4928
##
## Predicted class memberships (by modal posterior prob.)
##  0.0952 0.3093 0.0395 0.556
##
## =========================================================
## Fit for 4 latent classes:
## =========================================================
## number of observations: 9800
## number of estimated parameters: 35
## residual degrees of freedom: 220
## maximum log-likelihood: -19648.77
##
## AIC(4): 39367.53
## BIC(4): 39619.19
## G^2(4): 103.8543 (Likelihood ratio/deviance statistic)
## X^2(4): 113.827 (Chi-square goodness of fit)
##
## ALERT: iterations finished, MAXIMUM LIKELIHOOD NOT FOUND
##
```

```
seg.lca.5 <- poLCA(model, data = cluster.lca.df, nclass = 5,
    na.rm = TRUE)
```

```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $Grocery.Beverage
##                0      1
## class 1:  0.7771 0.2229
## class 2:  0.0000 1.0000
## class 3:  0.0000 1.0000
## class 4:  0.9304 0.0696
## class 5:  0.0000 1.0000
##
## $Grocery.Frozen.Pizza
##                0      1
```

```
## class 1:  1.0000 0.0000
## class 2:  0.9841 0.0159
## class 3:  0.9481 0.0519
## class 4:  0.9995 0.0005
## class 5:  0.9964 0.0036
##
## $Grocery.Facial.Tissue
##                0      1
## class 1:  0.1404 0.8596
## class 2:  0.2658 0.7342
## class 3:  0.0820 0.9180
## class 4:  0.9316 0.0684
## class 5:  0.7606 0.2394
##
## $Grocery.Laundry.Detergent
##                0      1
## class 1:  1.0000 0.0000
## class 2:  0.9274 0.0726
## class 3:  0.7257 0.2743
## class 4:  0.9995 0.0005
## class 5:  0.9938 0.0062
##
## $Grocery.Shampoo
##                0      1
## class 1:  0.9808 0.0192
## class 2:  1.0000 0.0000
## class 3:  0.7443 0.2557
## class 4:  0.9728 0.0272
## class 5:  0.9978 0.0022
##
## $Grocery.Soup
##                0     1
## class 1:  1.0000 0e+00
## class 2:  0.9990 1e-03
## class 3:  1.0000 0e+00
## class 4:  1.0000 0e+00
## class 5:  0.9998 2e-04
##
## $Grocery.Sugar
##                0      1
## class 1:  1.0000 0.0000
## class 2:  0.7466 0.2534
## class 3:  0.9600 0.0400
## class 4:  0.9995 0.0005
## class 5:  0.9763 0.0237
##
## $Grocery.Peanut.Butter
##                0      1
## class 1:  0.9997 0.0003
## class 2:  0.5302 0.4698
## class 3:  0.8456 0.1544
## class 4:  0.9651 0.0349
## class 5:  0.9669 0.0331
##
```

```
## Estimated class population shares
##  0.2383 0.1155 0.0624 0.2055 0.3783
##
## Predicted class memberships (by modal posterior prob.)
##  0.1717 0.0901 0.0327 0.2047 0.5008
##
## =========================================================
## Fit for 5 latent classes:
## =========================================================
## number of observations: 9800
## number of estimated parameters: 44
## residual degrees of freedom: 211
## maximum log-likelihood: -19626.97
##
## AIC(5): 39341.94
## BIC(5): 39658.31
## G^2(5): 60.26288 (Likelihood ratio/deviance statistic)
## X^2(5): 76.30364 (Chi-square goodness of fit)
##
## ALERT: iterations finished, MAXIMUM LIKELIHOOD NOT FOUND
##
```

```
seg.lca.6 <- poLCA(model, data = cluster.lca.df, nclass = 6,
    na.rm = TRUE)
```

```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
## $Grocery.Beverage
##                0      1
## class 1:  0.7676 0.2324
## class 2:  0.3501 0.6499
## class 3:  0.0000 1.0000
## class 4:  0.0000 1.0000
## class 5:  0.0000 1.0000
## class 6:  0.5799 0.4201
##
## $Grocery.Frozen.Pizza
##                0      1
## class 1:  1.0000 0.0000
## class 2:  0.9980 0.0020
## class 3:  0.9689 0.0311
## class 4:  0.9843 0.0157
## class 5:  0.9131 0.0869
## class 6:  1.0000 0.0000
##
## $Grocery.Facial.Tissue
##                0      1
## class 1:  0.6236 0.3764
## class 2:  0.9873 0.0127
## class 3:  0.0000 1.0000
## class 4:  0.3332 0.6668
## class 5:  0.2259 0.7741
## class 6:  0.0090 0.9910
```

```
##
## $Grocery.Laundry.Detergent
##                0      1
## class 1:  1.0000 0.0000
## class 2:  0.9987 0.0013
## class 3:  0.8204 0.1796
## class 4:  1.0000 0.0000
## class 5:  0.0266 0.9734
## class 6:  1.0000 0.0000
##
## $Grocery.Shampoo
##                0      1
## class 1:  0.9118 0.0882
## class 2:  1.0000 0.0000
## class 3:  0.6500 0.3500
## class 4:  0.9946 0.0054
## class 5:  0.9283 0.0717
## class 6:  1.0000 0.0000
##
## $Grocery.Soup
##                0     1
## class 1:  1.0000 0e+00
## class 2:  0.9998 2e-04
## class 3:  1.0000 0e+00
## class 4:  0.9992 8e-04
## class 5:  1.0000 0e+00
## class 6:  1.0000 0e+00
##
## $Grocery.Sugar
##                0      1
## class 1:  1.0000 0.0000
## class 2:  0.9988 0.0012
## class 3:  0.9794 0.0206
## class 4:  0.7815 0.2185
## class 5:  0.8926 0.1074
## class 6:  1.0000 0.0000
##
## $Grocery.Peanut.Butter
##                0      1
## class 1:  0.9847 0.0153
## class 2:  0.9706 0.0294
## class 3:  0.8760 0.1240
## class 4:  0.6595 0.3405
## class 5:  0.6746 0.3254
## class 6:  1.0000 0.0000
##
## Estimated class population shares
##  0.1276 0.4104 0.0378 0.1706 0.0212 0.2324
##
## Predicted class memberships (by modal posterior prob.)
##  0.0109 0.5015 0.0153 0.0945 0.0249 0.3529
##
## =========================================================
## Fit for 6 latent classes:
```

```
## ============================================================
## number of observations: 9800
## number of estimated parameters: 53
## residual degrees of freedom: 202
## maximum log-likelihood: -19627.54
##
## AIC(6): 39361.07
## BIC(6): 39742.15
## G^2(6): 61.39584 (Likelihood ratio/deviance statistic)
## X^2(6): 58.4503 (Chi-square goodness of fit)
##
## ALERT: iterations finished, MAXIMUM LIKELIHOOD NOT FOUND
##
```

Step 4: CHoose the best number of clusters according to BIC

```
seg.lca.3$bic
```

```
## [1] 39649.88
```

```
seg.lca.4$bic
```

```
## [1] 39619.19
```

```
seg.lca.5$bic
```

```
## [1] 39658.31
```

```
seg.lca.6$bic
```

```
## [1] 39742.15
```

```
# The output of the poLCA function includes BIC values that
# are called bic We request them by referring to the output
# data and asking for the bic in that data (thus $bic) We
# find that the lowest BIC is for the 3 cluster solution so
# we examine the 3 cluster solution further.
```

Step 5: Interpret the clusters

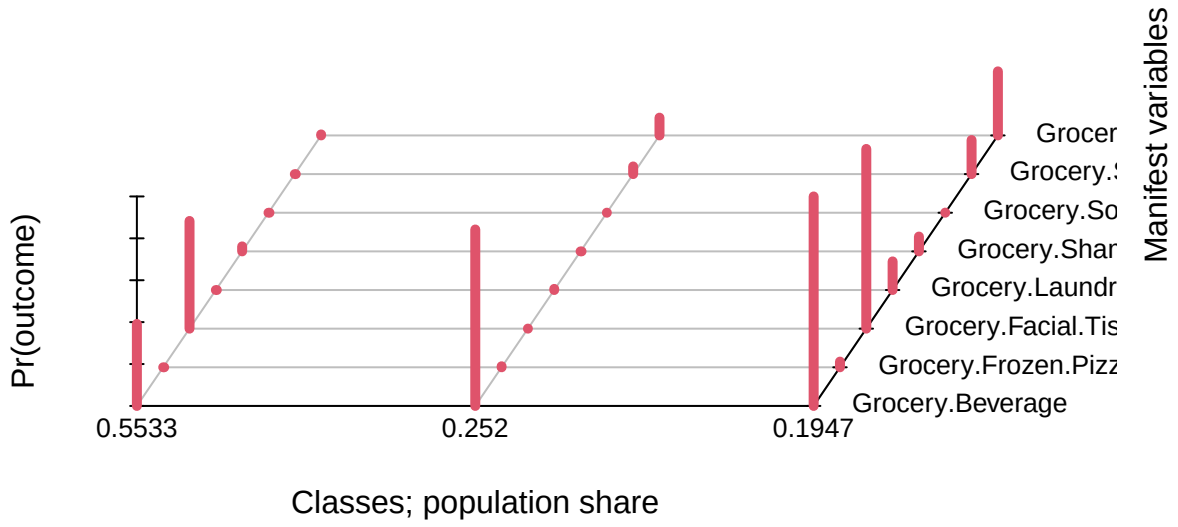
```
seg.lca.3  # Lowest BIC
```

```
## Conditional item response (column) probabilities,
##  by outcome variable, for each class (row)
##
```

```
## $Grocery.Beverage
##                0      1
## class 1:  0.6085 0.3915
## class 2:  0.1577 0.8423
## class 3:  0.0000 1.0000
##
## $Grocery.Frozen.Pizza
##                0      1
## class 1:  1.0000 0.0000
## class 2:  0.9951 0.0049
## class 3:  0.9728 0.0272
##
## $Grocery.Facial.Tissue
##                0      1
## class 1:  0.4860 0.5140
## class 2:  0.9996 0.0004
## class 3:  0.1421 0.8579
##
## $Grocery.Laundry.Detergent
##                0      1
## class 1:  1.0000 0.0000
## class 2:  0.9945 0.0055
## class 3:  0.8635 0.1365
##
## $Grocery.Shampoo
##                0      1
## class 1:  0.9762 0.0238
## class 2:  1.0000 0.0000
## class 3:  0.9293 0.0707
##
## $Grocery.Soup
##                0     1
## class 1:  1.0000 0e+00
## class 2:  0.9996 4e-04
## class 3:  0.9994 6e-04
##
## $Grocery.Sugar
##                0      1
## class 1:  1.0000 0.0000
## class 2:  0.9635 0.0365
## class 3:  0.8376 0.1624
##
## $Grocery.Peanut.Butter
##                0      1
## class 1:  0.9951 0.0049
## class 2:  0.9150 0.0850
## class 3:  0.6941 0.3059
##
## Estimated class population shares
##  0.5533 0.252 0.1947
##
## Predicted class memberships (by modal posterior prob.)
##  0.556 0.3393 0.1047
##
```

```
## ============================================================
## Fit for 3 latent classes:
## ============================================================
## number of observations: 9800
## number of estimated parameters: 26
## residual degrees of freedom: 229
## maximum log-likelihood: -19705.47
##
## AIC(3): 39462.94
## BIC(3): 39649.88
## G^2(3): 217.2594 (Likelihood ratio/deviance statistic)
## X^2(3): 256.1973 (Chi-square goodness of fit)
##
## ALERT: iterations finished, MAXIMUM LIKELIHOOD NOT FOUND
##
```

```
plot(seg.lca.3)
```




```
# Please note that the specific results, the order of the
# classes found, and the plots will look slightly different
# each time you run a new LCA.
```