Latent Class Analysis with Starbucks Survey Data

Emma Bright and Krista Miller

March 13th, 2022

# Executive Summary

Survey data is a popular method researchers use to gather information about specific topics from targeted populations. While the literal responses collected from individuals in these surveys are important to analyze and understand, it can also be helpful to see what latent, underlying characteristics may have resulted in an individual’s survey response. Latent class analysis (LCA) is a statistical technique that can be used to provide such insight into survey data by identifying latent classes[[1]](#footnote-2) of customers using independent observations and latent variables[[2]](#footnote-3).

The objective of this research project is to explore LCA and its application by attempting to categorize Starbucks customer survey respondents into distinct classes of customers based on latent variables present within their responses to survey questions.

To achieve this objective, we performed appropriate data preparation methods, such as exploratory analysis, variable selection and encoding of our data, and applied the poLCA R package to build an LCA model. Our “best fit” LCA model was determined by iteratively adding latent classes to our parameters until we had a model that satisfied statistical measures such as minimizing the AIC and BIC and increasing model parsimony.

Our final model results showed that there were two distinct latent classes present within our Starbucks customer survey data: the “Poor, Elder Millennials” and the “Comfortable Gen X’ers”. These results could be used to create targeted recommendations or campaigns and drive other business decisions that would increase Starbucks revenue.

# Introduction

Survey data provides researchers with an opportunity to gather information regarding specific topics from target respondents. Surveys are a cost effective, reliable, and versatile method of data collection which has led to their popularity among researchers and corporations. However, it can be difficult to craft a survey question that adequately captures the complexities that underly a person’s responses; especially for questions that have canned, categorical responses such as “yes” and “no”.

To address this, there are statistical methods available that allow a researcher to gain insight into these underlying complexities, such as Latent Class Analysis. LCA is a statistical analysis technique that can be used to provide insight into survey data by identifying latent classes of customers using independent observations and latent variables (Penn State University, 2019).

# Objective

The goal of this research project is to explore the Latent Class Analysis statistical technique and its application with survey data. Throughout our exploration we will be seeking to determine if LCA can be used to classify Starbucks customer survey respondents into distinct latent subclasses based on the latent variables present within responses to questions covering consumption behaviors, demographics, and perceptions of retail experiences.

# Methods

Latent Class Analysis is a clustering method used to group data into subclasses and is often utilized as a method for profiling at-risk populations in the healthcare field or in conjunction with other statistical methods, such as regression, for predictive analysis (Weller et al., 2020). In LCA, classes are identified and created from latent, non-observable, variables within individual responses from survey data.  These classes are then reviewed by the researcher performing the LCA and labeled according to the probabilities of response to each manifest variable, given the respondent’s class membership.

## Assumptions

Latent Class Analysis does not rely on any assumptions regarding linearity, distributions of the data, or homogeneity of the data (Penn State University, 2019). The major considerations of this method are that the observations within the dataset must be independent and consist solely of categorical or ordinal features. It is also assumed that the groups derived from LCA are both mutually exclusive and exhaustive. In other words, a class member can belong to only one class, and the data encompasses all possible classes.

# Data Preparation

Data preparation is an important part of performing any statistical analysis and though there were not many assumptions we needed to take into consideration with our statistical technique, we still needed to perform exploratory analysis, variable selection and encode our data. The Starbucks Customer Survey dataset we prepared for our LCA was collected from Kaggle and consisted of 112 independent customer responses to 21 survey questions (Hamzah, 2020).

## Exploratory Analysis

Given that it’s the responsibility of the researcher to label the latent classes output from an LCA model, it’s particularly important to perform exploratory analysis. This analysis provides additional context for labeling and allows a researcher to use reasonable judgment in determining which variables to include within an LCA Analysis.

As we dove into our exploratory analysis, we found a natural divide in the types of survey questions that were asked. There appeared to be three groups of questions: 1) demographic questions encompassing variables such as age, income, employment status and gender, 2) buyer behavior questions encompassing variables such as how much time the respondent typically spends in the store, how they typically make their purchases, what they buy and how much money they usually spend and 3) experience variables consisting of questions regarding perception of ambiance, service, feelings about sales and promotions, and overall quality of the Starbucks product.

We performed univariate, bivariate, and multivariate analysis to find the frequency of responses and any highly correlated manifest variables. Then, the outputs of our exploratory analysis were used to determine which variables to select for our LCA model.

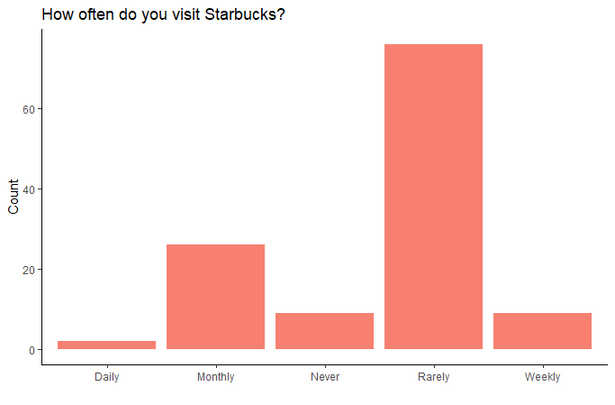


Figure 1.0 This chart was an outcome of the univariate analysis performed showing the frequency of responses to the Buy Behavior question "How often do you visit Starbucks?"



Figure 2.0 This correlation plot was an outcome of the bivariate analysis performed to show correlations between the 'Customer Experience' variables.

## Variable Selection

It is up for debate whether and how useful variable selection is for LCA. The consensus among researchers is that the more observations and the more variables the better, however, it is also important to consider that the researcher performing the LCA must interpret the probability of each of these variables in the LCA model output (Weller et al., 2020). Therefore, if you have an excessive number of variables or variables you don’t thoroughly understand it could affect the model parsimony and class labeling.

We explored R packages that were designed for dimensionality reduction, however, after performing additional research on LCA it made more sense to choose variables based on the outputs of our exploratory analysis and our own reasonable judgment as we must interpret the output of the LCA model by variable.

In the end, we decided to select two variables from each of the three question categories: Demographic, buyer behavior, and customer experience. For the demographic variables, we chose to include Income and Age as both the Gender and Employment status variables were highly correlated with Income. For the buyer behavior variables, we chose to select ‘How much time do you usually spend at Starbucks?’ and ‘How do you usually consume Starbucks?’. For the customer experience variables, we chose ‘How would you rate the quality of Starbucks compared to other brands (Coffee Bean, Old Town White Coffee..) to be:’ and ‘How would you rate the service at Starbucks? (Promptness, friendliness, etc..)’.

## Encoding Data

To use the poLCA R package for Latent Class Analysis, we needed to transform our categorical data into numerical data. To do this, we performed two types of categorical encoding to transform our data: one-hot encoding and label-encoding.

One-hot encoding is a technique that transforms all the elements within a categorical column into new columns represented by binary values (0 or 1) to signify the presence of the category value. This technique was utilized where there were many data points present within a single column; for example, a list of items, such as coffee or pastries, that they frequently purchase.

Label encoding was performed to transform the remaining categorical elements into numerical values without any additional column. For example, this was used to convert columns with values such as “Agree” and “Strongly Agree” to values such as 4 and 5.

Finally, we needed to transform our data to reflect the presence of manifest variables. In latent variable models, a manifest variable is an observable variable, meaning a variable that can be measured directly. A manifest variable can be continuous or categorical, however, it cannot be 0, so we had to add 1 to each of our newly transformed numerical variables.

# Results

After performing the data preparation steps on our Starbucks customer survey data, we were ready to run our LCA model. The number of latent classes are chosen by iteratively adding potential classes to the LCA model parameters and evaluating the output to determine which model is best fit to the data.

Judging “best fit” requires consideration of several conditions including evaluation of the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and overall model parsimony (Penn State University, 2019). While most model evaluations focus primarily on minimizing AIC and BIC, with an LCA model it is also very important to consider the model interpretability as the researcher is who is responsible for interpreting the output must apply the final class labels.

Our “best fit” LCA model consisted of 2 classes. This model was chosen for several reasons, specifically, this model has the lowest AIC, a lower BIC than higher class models, and the highest model interpretability.

Figure 3 The statistical outputs, AIC and BIC, of our LCA Models by number of classes.

After selecting our 2-class model, we can begin to interpret the output. With an LCA model, probabilities are generated based on the assumption of underlying latent variables, not the direct response of the survey question.  For example, if you were a class 1 member, the probability that you responded that your age was 20-29 was 80%. In other words, this does not mean 80% of survey respondents are 20-29, this means that given you are in class 1, you are likely to respond that you are 20-29 regardless of accuracy or truthfulness.

In general, we found that if you were a member of class 1, you were likely to respond that you were younger, poorer, on-the-go, and were spending less than 30 minutes at Starbucks.  This group was also more likely to highly rate the service and rate the product as mid-tier.  If you were a member in class 2, you were likely to respond that you were over 30, have a higher income, and spend more time drinking coffee at a Starbucks. This group was more likely to rate the service as mid-tier and highly rate the product.



Figure 4 The output of our two-class LCA model.

# Conclusion

Our objective was to determine if LCA can be used to classify Starbucks customer survey respondents into distinct latent subclasses based on the latent variables present within the responses to questions regarding their consumption behaviors, demographics, and perceptions of their retail experiences. We were able to achieve this objective by performing data preparation and running iterative LCA models which found two distinct latent classes present within our data.

The final step for LCA is for the researcher to assign class labels to the output classes. We decided to label our two classes the “Poor, Elder Millennials” and the “Comfortable, Gen X’ers”. 81.25% of the surveyed population were predicted to be within our “Poor, Elder Millennial” class, while 18.75% of the survey population were predicted to be in the “Comfortable, Gen X’ers” class. Given that one of our assumptions of LCA is mutual exclusive and exhaustive groups, a researcher could generalize this to make assumptions about the general population that visits this Starbucks location.

In summary, LCA is a method that can be applied to complex datasets to organize observed variables that represent unobservable constructs into two or more potentially meaningful, homogenous subgroups. LCA can be particularly useful for identifying subgroups who could benefit from a common intervention based on their shared characteristics and similar survey patterns. In terms of our dataset, Starbucks could use this information for targeted recommendations or campaigns and drive other business decisions to increase revenue.

**References**

Hamzah, M. (2020, May). Starbucks customer survey, Version 2. Retrieved February 20, 2022, from https://www.kaggle.com/mahirahmzh/starbucks-customer-retention-malaysia-survey

Penn State University. (2019, March 7). *1&1 webinar on Latent Class Analysis with Bethany Bray* [Video]. YouTube. [1&1 Webinar on Latent Class Analysis (LCA) with Bethany Bray](https://www.youtube.com/watch?v=BGdvoEzLCYA)

Weller, B.E., Bowen, N.K., & Faubert, S.J. (2020). Latent Class Analysis: A guide to best practice. *Journal of*  *Black Psychology*. [https://doi.org/10.1177/0095798420930932](https://doi.org/10.1177%2F0095798420930932)

1. Latent Classes: A group that is clustered together based upon latent, unobservable variables within a dataset. [↑](#footnote-ref-2)
2. Latent Variables: Unquantifiable and “unseen” factors present within data. These could include a person’s level of neurosis, conscientiousness, or openness to answering questions. [↑](#footnote-ref-3)