Modelling Consumer Choice using Compositional Data Analysis

Evans Yeboah (evansyeboah@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Associate Professor Ian Durbach
University of Cape Town, South Africa

18 May 2017

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

In this essay we examine the problem of modelling the combination of brands that a customer will buy in a category. In most product categories (e.g. chocolates), consumers like and buy several brands rather than just one, and share their spending between brands in proportion to the degree to which they like the brand. This gives rise to so-called "compositional" data, which can be modelled using compositional data analysis. In this essay we apply compositional data analysis to some problems in consumer behaviour: clustering together consumers with similar profiles ("customer segmentation"). Using panel data, we build a predictive model to explain how a customer's profile affects his/her choice of brands.

Keywords: Compositional data, logratio transformation, cluster analysis, multivariate linear regression, logistic regression, consumer purchasing behaviour.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Evans Yeboah, 18 May 2017

Contents

1. Introduction

Consumers exhibit different buying behaviours, which make prediction of their preference for a particular brand difficult, but not impossible. A customer's change in preference for a brand may be induced by external factors such as brand promotions or internal factors as satisfaction. Quite a good number of customers would prefer to buy a cheaper brand, neglecting its quality or fame. However, a switch of brands may also occur naturally, by a customer's decision to try other brands.

In particular, instead of buying a single brand, consumers spread their spending proportionally between different brands. We could conduct a study on consumers' choice for, say a brand A. But there is an obvious between-brand relationship. A customer that buys for instance, a soft drink A, clearly will not buy soft drink B unless for occasional purposes. What is often of interest to us is to explain this multivariate nature of choosing between brands – that is, buying different brands simultaneously. One way of doing this is to treat these brand choices as proportions; which can be modelled using compositional data analysis.

We say data is compositional if it consists of vectors whose components describe proportions of some whole feature (Aitchison, 2003). Compositional data contains relative information derived from the ratio of components. For example, in a household survey, if we are interested in studying the proportions of expenditure on accommodation, food, electricity, and transport then our data is compositional. The questions we might want to answer are: to what extent will a man's pattern of budget-share depend on his total amount spent? Or, are the expenditure patterns for men and women different? If we are interested in the total expenditure, then the data isn't compositional any more. It's looking at expenditure shares that gives the data its compositional nature.

Compositional data can be presented as vectors of proportions, concentrations, frequencies, percentages, or in parts per million. Because these proportions are real numbers, a cursory look is to analyse them as multivariate data (Pawlowsky-Glahn et al., 2015). However, compositional data are constrained; hence it is inappropriate to apply standard multivariate methodology directly to them. Compositional data analysis uses a transformation from the constrained $\mathbb{R}^n$ space to an unconstrained $\mathbb{R}^{n-1}$ space to solve this problem and provide a flexible way of modelling such data. The basic principle of compositional data analysis according to Aitchison (2003) is that: "any meaningful function of a composition can be expressed in terms of ratios of the components of the composition."

Several models have attempted to describe and predict the patterns of consumer choice. Such predictions would require a venerable technique that would yield reliable results. The most prominent is the comprehensive Dirichlet model, notably known to be one of the best empirical generalisations in marketing. As we see in Chapter 2, the model and its related distributions have some drawbacks, which suggest a look at other plausible methods.

This essay considers a descriptive and predictive modelling approach: cluster analysis and compositional data analysis. We use cluster analysis to ascertain whether or not the buying patterns of consumers form similar groups; and assess the effect of some demographics on these cluster groups. Then we employ the techniques in compositional analysis: particularly the additive logratio approach to investigate whether a customer's choice of brands in say year 2 depends on his/her choices in year 1.

1.1 Research Questions

We use a two-year panel dataset containing purchases of the four most consumed chocolate brands: Cadbury, Kit Kat, Lindt and M&M. The dataset is a sample of 1617 consumers in Australia; and we attempt to answer the following questions:

(a) Can we group customers with similar spending patterns into clusters? If so, will a customer assigned to cluster i in year 1 be assigned to the same cluster in year 2?

(b) Can we predict a customer's assignment to an ith cluster based on some demographics?

(c) If we consider either the number of purchases made or the amounts paid on products bought, do we obtain the same cluster assignments?

(d) Is there a relationship between customers' profile and their choice of brands? How strong is this relationship? Is it linear or non-linear?

(e) Can we develop a predictive model for consumers' choice of the four chocolate brands under study?

1.2 Objectives

The main objectives of the essay are to:

(a) use the K-means clustering algorithm to group similar customers into segments and assess the stability of cluster assignments;

(b) use logistic regression to determine the relationship between some customer demographics and their assigned clusters;

(c) apply the techniques of compositional data analysis to model the dynamics of consumers' choice of the four chocolate brands; and

(d) use the developed model to predict a customer's likelihood of purchasing a chocolate brand in the future.

1.3 Organisation of the Essay

Having established the purpose of the essay in Chapter 1, in Chapter 2, we review related literature on modelling approaches of consumer behaviour. In addition, we review the theories of multiple and multivariate linear regression and logistic regression. Chapter 3 discusses the main methods employed for analyses; namely, cluster analyses and compositional data analyses. The main analyses and discussions are performed in Chapter 4. Finally, Chapter 5 concludes the essay based on the results and findings.

2. Review of Relevant Literature

This chapter reviews related studies already conducted. The dataset used for this essay is a panel data, hence we give a brief description of it. Next we take a look at some common approaches to modelling purchase behaviour. Furthermore, we give some introductory concepts of multiple and multivariate linear regression as these are the building blocks for the later compositional approaches we use; and a brief look at logistic regression.

2.1 Panel Data

A time series data that provides multiple observations on each individual in a given sample is termed a panel or cross-sectional data (Hurlin, 2010). Panel data provides two sets of information. While the differences between subjects provide cross-sectional information, we obtain within-subject information from the changes within subjects over time (Hurlin, 2010).

Consider Table 2.1 below, which shows a panel dataset collected (characteristics of age, stipends and grade point average (GPA)) over the course of four years for two undergraduate students. From this dataset, we could focus on the differences between each student or assess the changes in observed phenomena for one student over the course of the study (e.g., the changes in stipend over time of student 1).

Student ID	Year	Age (years)	Stipend ($)	GPA (out of 5)
1	2011	19	400	4.22
1	2012	20	400	4.77
1	2013	21	450	4.90
1	2014	22	460	4.90
2	2011	18	350	4.22
2	2012	19	390	4.41
2	2013	20	430	4.54
2	2014	21	450	4.65

Table 2.1: A Panel Data Set of Two Students

2.1.1 Types and Characteristics of Panel Data.

Panel data are of three types: (a) short panels of many individuals but few time periods, (b) long panel of few individuals but many time periods, and (c) both long and short panels with many individuals and many time periods.

They are characterised by large number of data points, which increases the degrees of freedom to explore explanatory variables and relationships. Panel data also assumes correlation (clustering) over time for a given independent panellist. For instance, the stipend for the same student is correlated over time but it is independent across the observed students. As such, we need some way of modelling this clustering.

The use of panel data allows us to control some missing values by observing changes in the response variable. This is possible if these missing values vary over time, but are constant between subjects, or vary between subjects but are constant over time (Torres-Reyna, 2007).

2.2 Common Approaches to Modelling Purchase Behaviour

2.2.1 Nature of Consumer Brand Choice.

The behaviour of a consumer may be subject to change. Consider a consumer's behaviour to be defined by purchasing the same brand. This consumer may change behaviour, and purchase a different brand, based on a change in preference or attitude. However, identifying a change of behaviour is not straightforward when the sequences of brand choices of consumers looks like a stochastic process.

According to Bass et al. (1984), frequent switching among brands is the usual nature of consumer brand choice behaviour; and that low-priced products account for frequent purchasing. The brand choice process, they say, takes the form of a stochastic process, and that it is mostly not characterised by consumers' loyalty to a brand.

Bass et al. (1984) investigated the nature of the brand choice process at the individual family level. They employed four tests: t, likelihood ratio, binomial runs, and multinomial runs tests. They assumed stationarity of the order of brand process, and having passed stationarity tests, performed separate analyses for purchase sequences of stationary and nonstationary behaviour.

On the basis of the study of frequently-purchased and low-priced products, Bass et al. (1984) concluded that there are four types of choice behaviour: (a) nonstationary behaviour, (b) stationary and zero-order behaviour, (c) stationary and nearly zero-order behaviour, and (d) only stationary behaviour. Using panel and experimental data, their analyses indicated that the purchase sequences of most stationary consumers consistently follow a zero-order process[1].

2.2.2 The Negative Binomial Distribution (NBD) Model for Customer Purchases.

In many product categories, some customers purchase more regularly and some others never purchase. In marketing, is a common to describe the purchasing rates of individuals purchasing in a Poisson manner with a gamma distribution across the population of customers. The negative binomial distribution (NBD) is a mixture of Poisson distributions, where the mixture parameter is given by a gamma distribution.

Morrison and Schmittlein (1988) have assessed the usefulness and implications of the negative binomial distribution (NBD) model. They viewed the NBD model as a reasonable representation of observed buying patterns of consumers and noted that, it is certain there will be some randomness in the purchasing pattern of each customer. In cases where multiple deviations occurred together, they made some conjectures about the robustness of the NBD model. They also stressed that though the NBD model may function, it still requires working on variations of the negative binomial distribution.

The NBD has three characteristics: Poisson purchasing, gamma heterogeneity and stationarity defined as follows:

(a) *Poisson Purchasing*: A Poisson process with a rate λ generates each ith customer's purchasing occasion.

(b) *Gamma Heterogeneity*: Across the population of consumers, these λ purchasing rates are fitted to a gamma distribution. The distribution on λ combines very well with the individual-level Poisson and exponential distributions. This helps in making useful predictions in a simple mathematical form, about the purchase patterns of distinct consumer groups.

[1] By a zero order process, we mean that the brands you purchase next do not depend on the brands purchased before.

(c) *Stationarity*: Since individual consumers remain as Poisson purchasers, their purchasing rates, λ also remain unchanged over time. This makes the NBD a stationary model and makes future purchase predictions on purchasing history very easy.

After an extensive review, Morrison and Schmittlein (1988) made two important findings. Firstly, they established that the NBD can be a reasonable model for either units bought or when and how often consumers purchase a product. Secondly, the NBD's conditional expectations can be used as guidelines for analysing sales growth and decline. They further affirm that, the identification of a stable baseline period of any duration is an issue, but do not see these as limitations to using the NBD.

2.2.3 The Dirichlet Buying Behaviour Model.

Where sales of different brands show no special patterns and each brand shows slight variation over the time-periods observed, we may consider the Dirichlet model. The comprehensive Dirichlet model was developed by Goodhardt et al. (1984). It describes how frequently-purchased consumer products such as soaps and soft drinks are bought during stationary and unsegmented market conditions.

Given N consumers buying in a product-class of b brands, the model probabilistically specifies the number of purchases by a consumer, and the specific brand that consumer buys on each purchase occasion in a given period. The model has a crucial property of combining different brands into what is termed: "super-brand."

According to Goodhardt et al. (1984), the Dirichlet model assumes a mixture of distributions at four levels :

(a) a Poisson process that describes purchasing of the product-class for each consumer,

(b) a Gamma distribution that describes the purchasing rates of distinct consumers,

(c) a multinomial distribution describing each consumer's choice of brand among available ones, and

(d) a multivariate Beta or "Dirichlet" distribution describing these choices across distinct consumers.

The required inputs for the model are the sales level of each brand and the two aspects of consumer diversity as parameters. These two aspects are how much consumers differ from each other based on their purchasing rates, and choice of brands. Goodhardt et al. (1984) used these inputs from a large consumer panel data, taking the buying occasion for a given brand as the unit of analysis and separately analysing the price paid on each purchase occasion.

2.2.4 Regularities in Patterns of Buyer Behaviour.

Many regularities have been observed in the buying behaviour of consumers. These regularities consist of a number of brand performance measures such as: the percentage of buyers, how many bought, loyalty rate, and which other brands they bought. One of the most well-known "regularities" is known as "double jeopardy": users of big brands buy that brand relatively more often than users who buy smaller brands. Thus small brands are doubly 'cursed': they have fewer users, and their users are less (behaviourally) loyal.

Uncles et al. (1995) have studied the Dirichlet model, assessing how the regularities are intertwined and how this modelling approach assist marketing analysts. In their article, they emphasise that a well established model has an important feature of extending it to novel situations. The role of theory in this they say: (a) helps in the mechanics of analysis (b) gives theoretical norms, or benchmarks; and (c) provide insights about buyer behaviour, for instance a typical customer's special loyalty to a brand.

Although the Dirichlet modelling assists market analysts to routinely monitor brand performance on a range of different loyalty measures, it has some shortcomings. Uncles et al. (1995) have observed that the Dirichlet is not suitable for very short periods such as daily purchasing. According to them, the model sometimes makes systematic under-predictions and mostly causes discrepancy problems at the margin. They thus suggested further work to study "model failure."

2.2.5 Why Modelling Averages are not Good Enough.

Undoubtedly, the Dirichlet and related distributions have been valuable in modelling typical brand performance measures (BPMs) in markets. However, there has been insightful critiques of attitudinal measures in relation to brand strength, differentiation and persuasion.

Although the Dirichlet provides accurate BPMs for fixed time-periods, Bongers and Hofmeyr (2010) have argued that its assumptions about individual behaviour are wrong. They challenge how theorists have used the law of double jeopardy to generalise brand performance, preference psychology and advertising. The law, they say accurately describes aggregate market behaviour only in a limited sense. They again point out that, it is equally significant to describe brands in terms of the direction caused by market share changes, than only in terms of their size, since the so called "big/strong" brands are vulnerable to losses in market shares.

According to them, modelling averages based on observations alone (such as the Dirichlet) cannot determine purchase propensities. They argue that attitudinal measures are the best way to resolve the empirical flaws of the law of double jeopardy. Hence the only path to the successful understanding of marketing initiatives may be the use of attitudinal surveys (for example awareness, desirability, preferences, branding).

2.3 Multiple Linear Regression (MLR)

In simple linear regression we explain how a response variable, y relates to one explanatory variable x. This relationship between the response and an explanatory variable can be described with a straight line. However there are situations in which the response variable may depend on, or change with, more than one explanatory variable. In such cases, we can fit a a multiple linear regression model which is linear in the coefficients.

The MLR model is given by:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i \qquad i = 1, \ldots, n \quad \text{and} \quad n > k, \tag{2.3.1}$$

where, β_i are the regression coefficients and ϵ_i is the random error associated with the response variable, y_i. The random error is called a residual when it is associated to the fitted value, $\hat{y}$ of y. While a deterministic model describes the mean response, a stochastic model describes the errors.

The MLR model describes the effect (change in mean response) of each explanatory variable and says that (Erni et al., 2016):

(a) the regression coefficient $\beta_i,\ i > 0$ is the change in mean response, when x_{ki} changes by one unit; given that the value of all the other explanatory variables are held constant.

(b) the expected (average) response is linearly related to both explanatory variables.

(c) the deviations of the observed values (response) are normally distributed around the fitted values.

The MLR model can often be an adequate representation of a more complicated structure within certain ranges of the explanatory variables. We can apply similar least squares techniques for estimating the coefficients when the linear model involves, say, powers and products of the explanatory variables (Walpole et al., 2012).

2.3.1 Estimation of Coefficients using Least Squares.

As in simple linear regression, we assume that the error terms ϵ_i are independent and identically distributed with mean, 0 and common variance σ^2. That is, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The least squares method chooses the coefficient values which minimises the sum of squared on errors (residuals), SSE. We obtain the least squares estimators of the parameters $\beta_0, \beta_1, \dots, \beta_k$ by fitting the MLR model to the data points: $x_{1i}, x_{2i}, \dots, x_{ki}, y_i$. We do this by minimising the expression:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2 \qquad (2.3.2)$$

The minimisation process involves solving for β for which $\dfrac{\partial}{\partial \beta}(SSE) = 0$.

2.3.2 Estimation using Matrix Notation.

Suppose we have k explanatory variables, $x_1, x_2, \dots, x_k$ and n observations, $y_1, y_2, \dots, y_n$; each of which can be expressed in the form of Equation (2.3.1). Then we can express Equation (2.3.1) in matrix form as:

$$Y = X\beta + \epsilon, \qquad (2.3.3)$$

where,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{21} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}_{n\times(k+1)}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The least squares method for estimation of the coefficients involves finding β for which

$$SSE = (Y - X\beta)^T (Y - X\beta) = \epsilon^T \epsilon \qquad (2.3.4)$$

is minimised.

We now estimate $\hat{\beta}$ as follows:

$$\begin{aligned}
SSE &= \sum_{i}^{n} e_i^2 = e^T e \\
&= (Y - \hat{Y})^T (Y - \hat{Y}) \\
&= Y^T Y - \hat{Y}^T X \hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X \hat{\beta} \\
\therefore\ SSE &= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X \hat{\beta}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta}(SSE) &= 0 \\
\implies \hat{\beta} &= -2X^T Y + 2X^T X \hat{\beta} = 0 \\
\therefore\ \ \hat{\beta} &= (X^T X)^{-1} X^T Y
\end{aligned}$$

Therefore for an invertible matrix, $X^T X$ we can find an estimate of the model coefficients such that:

$$\boxed{\hat{\beta} = (X^T X)^{-1} X^T Y} \tag{2.3.5}$$

2.3.3 Statistical Properties of the Least Squares Estimators.

Knowing that $\hat{\beta} = (X^T X)^{-1} X^T Y$, we show that the expected value, $\mathrm{E}(\hat{\beta})$ is an unbiased estimate of β as follows:

$$\begin{aligned}
\mathrm{E}(\hat{\beta}) &= \mathrm{E}\Big[(X^T X)^{-1} X^T Y\Big] \\
&= \mathrm{E}\Big[(X^T X)^{-1} X^T (X\beta + \epsilon)\Big] \\
&= \mathrm{E}\Big[(X^T X)^{-1} (X^T X)\beta\Big] + \mathrm{E}\Big[(X^T X)^{-1} X^T \epsilon\Big] \\
&= (X^T X)^{-1} (X^T X)\mathrm{E}(\beta) + (X^T X)^{-1} X^T \mathrm{E}(\epsilon) \\
\mathrm{E}(\hat{\beta}) &= I\beta = \beta
\end{aligned}$$

We deduce an estimate of the variance of $\hat{\beta}$ as :

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\Big[(X^T X)^{-1} X^T Y)\Big] \\
&= \Big[(X^T X)^{-1} X^T \mathrm{I}\sigma^2 X (X^T X)^{-1}\Big] \\
\therefore\ \mathrm{Var}(\hat{\beta}) &= (X^T X)^{-1} \hat{\sigma}^2
\end{aligned}$$

The elements of the matrix $(X^T X)^{-1} \hat{\sigma}^2$ display the variances of the estimates: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ on the main diagonal and covariances on the off-diagonal (Walpole et al., 2012).

2.3.4 The Residual Standard Error.

For $n - (k + 1)$ degrees of freedom an unbiased estimate, $\hat{\sigma}^2$ of σ^2 is computed using the formula:

$$\hat{\sigma}^2 = \frac{SSE}{n - k - 1}, \qquad \text{where} \quad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

The residual standard error gives us an idea of how large the unexplained variability is. Since standard errors are in the same units as the response, deciding what constitutes large or small standard errors needs to be seen in the context of the units used (Erni et al., 2016).

2.4 Multivariate Linear Regression (MvLR)

There are cases where we might want to find out how a set of covariates affect multiple response variables. Very often, these response variables are correlated among themselves. Performing individual regression (particularly analysis of variance) study on them, would not lead to as efficient estimates as it would if we studied them simultaneously. So instead of looking at several variables separately, in multivariate regression we consider them simultaneously.

Figure 2.1 below shows the conceptual model of the multivariate linear regression.

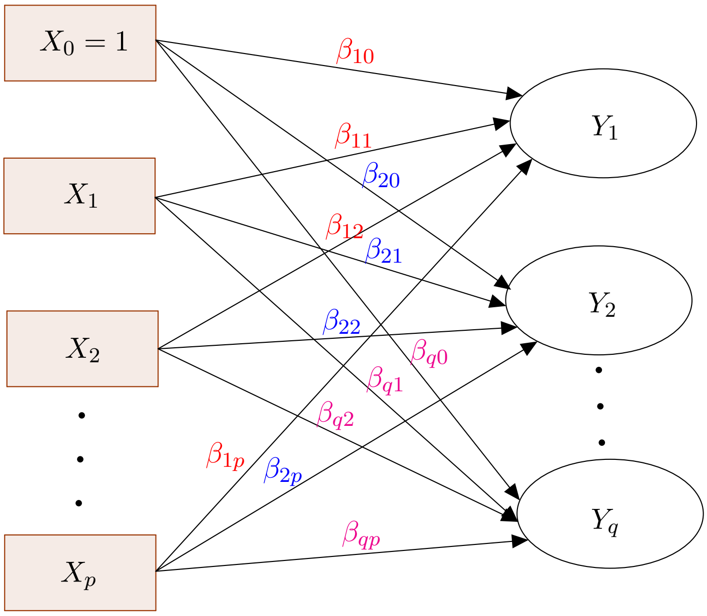

Figure 2.1: MvLR Conceptual Model

Equation (2.4.1) formally defines the multivariate linear regression model; where Y represents the response variables, X the explanatory variables, ϵ the errors, p the number of explanatory variables, and q the number of response variables.

$$\begin{cases} y_1 = \beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2} + \ldots + \beta_{1p}x_{ip} + \epsilon_{i1} \\ y_2 = \beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2} + \ldots + \beta_{2p}x_{ip} + \epsilon_{i2} \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ y_k = \beta_{k0} + \beta_{k1}x_{i1} + \beta_{q2}x_{i2} + \ldots + \beta_{qp}x_{ip} + \epsilon_{i2} \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ y_q = \beta_{q0} + \beta_{q1}x_{i1} + \beta_{q2}x_{i2} + \ldots + \beta_{qp}x_{ip} + \epsilon_{i2} \end{cases} \tag{2.4.1}$$

The MvLR model can be simplified as:

$$\boxed{Y_{n\times q} = X_{n\times(p+1)}\beta_{(p+1)\times q} + \epsilon_{n\times q}} \tag{2.4.2}$$

Assumptions of the Model

(a) Errors $\epsilon_{n\times q}$ are multivariate normal.

(b) Error variances are equal (homogenous) across observations; conditional on predictors.

(c) Errors have common covariance structure across observations.

(d) Independent observations.

2.4.1 Least Squares Estimate of Coefficients.

$$\begin{aligned} Y = X\beta + \epsilon \implies \hat{Y} &= X\hat{\beta} \\ Y_{n\times q} &= X_{n\times(p+1)} \cdot \beta_{(p+1)\times q} + \epsilon_{n\times q} \\ \implies [Y_1 \quad Y_2 \quad \cdots \quad Y_q] &= X[\beta_1 \quad \beta_2 \quad \cdots \quad \beta_q] + [\epsilon_1 \quad \epsilon_2 \quad \cdots \quad \epsilon_q] \end{aligned}$$

The SSE is no longer a scalar as in MLR but a matrix which is symmetric and has diagonal elements assumed to be 0. Let's now denote the SSE as the sum of squares of cross product on errors $(SSCPE_\epsilon)$.

$$SSCP_\epsilon = \epsilon^T_{q\times n} \cdot \epsilon_{n\times q}$$

We want this to be as small as possible, so we take the trace and minimize.

$$\begin{aligned} \text{trace } (SSCP_\epsilon) &= \text{tr } (\epsilon^T \epsilon) \\ \text{tr } (\epsilon^T \epsilon) &= \sum \epsilon_{i1}^2 + \sum \epsilon_{i2}^2 + \ldots + \sum \epsilon_{iq}^2 \\ &= \text{tr}[(Y - X\beta)^T (Y - X\beta)] \end{aligned}$$

Since the diagonal elements add to zero we minimize by setting

$$\frac{\partial \text{tr}(\epsilon^T \epsilon)}{\partial \beta_{kj}} = 0$$

And solving this as with MLR we obtain:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T [Y_1 : Y_2 : \ldots : Y_q] \\ \hat{\beta} &= (X^T X)^{-1} X^T Y \\ \\ \implies \hat{\beta}_1 &= (X^T X)^{-1} X^T Y_1 \\ \hat{\beta}_2 &= (X^T X)^{-1} X^T Y_2 \\ \hat{\beta}_q &= (X^T X)^{-1} X^T Y_q \end{aligned}$$

The difference in the β coefficients for MLR and MvLR comes form the error terms. In MLR, one observation per error. In MvLR, there are q variables.

2.4.2 Sampling Distribution of $\hat{\beta}$.

We show that $\hat{\beta}$ is an unbiased estimate of β

$$\begin{aligned}
\mathrm{E}(\hat{\beta}) &= \mathrm{E}\Big[(X^TX)^{-1}X^TY\Big] \\
&= (X^TX)^{-1}\mathrm{E}(Y) \\
&= (X^TX)^{-1}X^TX\beta = I\beta \\
\mathrm{E}(\hat{\beta}) &= \beta
\end{aligned}$$

The covariance, Cov is derived as:

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}) &= \mathrm{E}\Big[\{\hat{\beta}-\mathrm{E}(\hat{\beta})\}\{\hat{\beta}-\mathrm{E}(\hat{\beta})\}^T\Big] \\
\hat{\beta}-\mathrm{E}(\hat{\beta}) &= \hat{\beta}-\beta \\
&= (X^TX)^{-1}X^TY-\beta \\
&= (X^TX)^{-1}X^T(X\beta+\epsilon)-\beta \\
&= (X^TX)^{-1}X^TX\beta+(X^TX)^{-1}X^T\epsilon-\beta \\
&= (X^TX)^{-1}X^T\epsilon
\end{aligned}$$

$$\begin{aligned}
(\hat{\beta}-\beta)^T &= \Big[(X^TX)^{-1}X^T\epsilon\Big]^T \\
&= \epsilon^TX(X^TX)^{-1}
\end{aligned}$$

$$\begin{aligned}
\implies \mathrm{Cov}(\hat{\beta}) &= \mathrm{E}\Big[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T\Big] \\
&= \mathrm{E}\Big[(X^TX)^{-1}X^T\epsilon\cdot\epsilon^TX(X^TX)^{-1}\Big] \\
&= (X^TX)^{-1}X^T\mathrm{E}(\epsilon\cdot\epsilon^T)X(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T(I\otimes\Sigma)X(X^TX)^{-1} \\
&= (X^TX)^{-1}X^TX(X^TX)^{-1}\otimes\Sigma \\
&= (X^TX)^{-1}\otimes\Sigma,
\end{aligned}$$

where $\otimes$ is the Kronecker product and,

$$\Sigma_{q\times q} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1q} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2q} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1q} & \sigma_{2q} & \dots & \sigma_q^2 \end{bmatrix}, \qquad \text{where} \begin{bmatrix} \sigma_{11}=\sigma_1^2 \\ \sigma_{22}=\sigma_2^2 \\ \vdots \\ \sigma_{qq}=\sigma_q^2 \end{bmatrix}$$

The Kronecker product $\otimes$ is an operation on two matrices of arbitrary size that results in a block matrix.

By reason of Assumption (c), errors have common covariance structure across observations. We assume that the correlation structure of the $y_i's$ is negligible, thus insignificant. Hence all non-diagonal elements become zero. We can thus write

$$\begin{aligned} Y_1 &= (X^TX)^{-1}\hat{\sigma}_1^2 \\ Y_2 &= (X^TX)^{-1}\hat{\sigma}_2^2 \\ \vdots &= \quad \vdots \\ Y_q &= (X^TX)^{-1}\hat{\sigma}_q^2 \end{aligned}$$

2.5 Logistic Regression

When the response variable is categorical (binary or binomial), we cannot perform a linear regression, as the assumption of linearity is violated. To overcome this problem, we will require a transformation that expresses a non-linear relationship in a linear way. This is where logistic regression comes into play. Instead of predicting a response variable Y from an explanatory variable X, in logistic regression, we predict the probability of Y occurring given some known values of X.

Logistic regression expresses the linear regression equation as logarithmic transformations (called logit). It uses the maximum likelihood to estimate the values of our parameters. The maximum likelihood function selects coefficients that increases the probability that the observed values will occur (James et al., 2013). We can then model the effect of explanatory variables on a binary or binomial response.

The logistic function

$$P(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{2.5.1}$$

can be expressed as

$$\frac{P(X)}{1-P(X)} = e^{\beta_0+\beta_1 X} \tag{2.5.2}$$

The left hand side of Equation (2.5.2), called the *odds*, can take any value between 0 and ∞. Where a value close to 0 indicate very low probabilities of success and high otherwise. Taking log of both sides of Equation (2.5.2) yields:

$$\log\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X \tag{2.5.3}$$

The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable. The left hand side of Equation (2.5.3) is what is termed log odds or logit. This means that, increasing X by one unit changes the log odds by β_1, or equivalently the odds by e^{β_1}. However, in Equation (2.5.1), β_1 does not correspond to the change in $P(X)$ resulting from a unit increase in X. If β_1 is positive, then an increase in X will increase $P(X)$, and if β_1 is negative, then an increase in X decreases $P(X)$ (James et al., 2013).

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize this likelihood function.

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i:y_i=1} P(x_i) \cdot \prod_{i:y_i'=0} (1-P(x_i)) \tag{2.5.4}$$

We can also interpret the estimates in terms of the odds. For example, suppose we want to estimate the odds of a student gaining admission into a college based on grade point average (GPA); then the odds of gaining admission Y, depending on your GPA, X could be evaluated as:

$$\text{odds} = \frac{P(\text{admitted})}{P(\text{not admitted})} \tag{2.5.5a}$$

$$\text{P(event Y)} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}} \tag{2.5.5b}$$

$$\text{P(no event Y)} = 1 - \text{P(event Y)} \tag{2.5.5c}$$

We calculate the change in odds that results from a unit change in GPA, firstly by calculating the odds of gaining admission given that your GPA was not used (Equation (2.5.5b)). Next, we calculate the odds of gaining admission based on your GPA. Finally, we calculate the proportionate change in these two odds as in Equation (2.5.5a). Suppose we had several predictors, then we will use the Equation (2.5.6) below (Andy Field and Miles, 2012):

$$P(Y) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_n X_{ni})}}. \tag{2.5.6}$$

Now, we need to calculate the same thing after GPA has changed by one unit. Note here that our explanatory variable is dichotomous. So we calculate the odds of gaining admission given that GPA was used. So X is 1 (not 0). It is simple to calculate the proportionate change in odds by dividing the odds after a unit change in GPA by the odds before that change (Andy Field and Miles, 2012). Equation (2.5.7) illustrates this.

$$\triangle\text{odds} = \frac{\text{odds after a unit change in GPA}}{\text{original odds}} \tag{2.5.7}$$

If the value of the odds in Equation (2.5.7) is greater than 1, then as a student's GPA increases, so is the odds of gaining admission. Conversely if the odds is less than 1, then in increase in GPA decreases the odds of gaining admission (Andy Field and Miles, 2012).

3. Methodology

The main methods required for analysis of this research are: Cluster Analysis and Compositional Data Analysis. This Chapter gives an overview of what these methods are and how they function.

3.1 Cluster Analysis

In special classes of problems, we only observe input variables, without any corresponding output. In a marketing setting, we might have demographic information for a number of customers. We may want to understand which types of customers are similar to each other by grouping them according to their observed characteristics. So instead of predicting a particular output variable, we are rather interested in determining whether we can cluster together customers with similar profiles (James et al., 2013). This is what is termed cluster analysis. With cluster analysis, for every observation $i = 1, \ldots, n$, we observe a vector of measurements x_i but no associated response y_i. As such we cannot fit traditional linear regression models to such sets of data.

James et al. (2013) note that: "the goal of cluster analysis is to ascertain, on the basis of $x_1, \ldots, x_n$, whether the observations fall into relatively distinct groups." For example, in a market segmentation study we might observe multiple characteristics (variables) on customers, such as: occupation, family composition, income, and shopping habits. We might believe that the customers fall into one of two groups: regular or irregular buyers. We might also want to identify groups that differ with respect to some property of interest, such as brand preference.

One way of doing these is by perform cluster analysis. However, we should be aware that results from clustering do not give the absolute truth of a dataset. They are rather a starting point for the development of a scientific hypothesis and further study, preferably on an independent dataset (James et al., 2013).

The two best-known clustering approaches are the K-means and hierarchical clustering.

3.1.1 K-Means Clustering.

K-means clustering is a simple approach for partitioning a set of n observations, assumed to be independent, into a discrete set of distinct, non-overlapping clusters (Grindrod, 2014; James et al., 2013). To perform K-means clustering, we first pre-specify the desired number of clusters, K. The K-means algorithm iteratively moves the centres of the clusters to minimise the total within cluster variance (Friedman et al., 2001), and then assigns each observation to exactly one of the K clusters.

Let $C_1, \ldots, C_K$ denote the set of observations in each cluster C_i. Then:

(a) each observation belongs to at least one K cluster; that is: $C_1 \cup C_2 \cup \ldots \cup C_K, \ K = 1, \ldots, n,$

(b) no observation belongs to more than one cluster. That is $C_k \cap C_{k'} = \emptyset$.

K-means seeks to partition the observations into K clusters such that the total within-cluster $(W(C_k))$ variation, summed over all K clusters is as small as possible. Hence we solve the problem:

$$\underset{C_1,\ldots,C_K}{\text{minimise}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}. \tag{3.1.1}$$

In order to solve Equation (3.1.1), we need to define the within-cluster variation using the *squared Euclidean distance* as follows:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \tag{3.1.2}$$

where $|C_k|$ denotes the number of observations in the kth cluster and p is number of variables.

By combining Equations (3.1.1) and (3.1.2) we obtain the K-means clustering optimisation problem:

$$\underset{C_1,\dots,C_K}{\text{minimise}} \left\{ \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}. \tag{3.1.3}$$

The algorithm below provides a simple solution to Equation (3.1.3).

Algorithm 1 The K-Means Algorithm

```
Given an initial set of centres, the $K-$means algorithm alternates the two steps:
  (a) for each data point, we use the Euclidean distance to identify the closest cluster centre;
  (b) each cluster centre is replaced by the coordinate-wise means of all data points that are closest to
      it, and this mean vector becomes the new centre for that cluster.
```

One stopping criterion is to iterate these two steps until the centroids remain constant at two successful iterations, so that no more observations would be reclassified. Another stopping rule is to define some threshold for differences in centroids, or run some fixed number of iterations.

It should be emphasised that since the K-means algorithm finds a local optimum rather than a global. The results obtained will depend on the initial (random) cluster assignment of each observation. So we run the algorithm multiple times using different initial random configurations. The best solution is that for which Equation (3.1.3) is the smallest (Friedman et al., 2001).

3.1.2 Hierarchical Clustering.

Hierarchical clustering methods require the user to specify a measure of dissimilarity between (disjoint) groups of observations, based on the pairwise dissimilarities among the observations in the two groups (Friedman et al., 2001). Here, we do not pre-specify the number of clusters.

This method produces tree-like representations. The clusters at each level of the hierarchy are created by merging clusters at the next lower level. While each cluster at the lowest level contains a single observation, at the highest level there is only one cluster containing all of the data (Friedman et al., 2001). There are two strategies from hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). James et al. (2013) describe the algorithm as follows:

Algorithm 2 The Hierarchical Clustering Algorithm

```
1. Begin with $n$ observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = \frac{n(n-1)}{2}$
   pairwise dissimilarities. Treat each observation as its own cluster.
2. For $i = n, n-1, \ldots, 2$:
   (a) Examine all pairwise inter-cluster dissimilarities among the $i$ clusters and identify the pair
       of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The
       dissimilarity between these two clusters indicates the height in the dendrogram at which the
       fusion should be placed.
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.
```

Once this is done, it is represented graphically by a *dendrogram*, which provides a highly interpretable complete description of the hierarchical clustering.

3.1.3 Number of Clusters and Clustering Stability.

Clustering algorithms like K-means require an input of the number of clusters. A naive way of obtaining the right number of clusters is by deducing cluster formation from a dendrogram. However this is not straightforward and requires some expertise. Brock et al. (2011) have built an R package that attempts to suggest the best number of clusters using nine different clustering algorithms and assess their stability.

Given a numeric dataset, we can compare clustering results based on the full data and that of each column, one at time. There are stability measures available in the *clValid* package in R that seek to minimise the average taken over all removed columns. These are the: average proportion of non-overlap (APN), average distance (AD), average distance between means (ADM), and figure of merit (FOM) . Details about these measures can be found in Brock et al. (2011).

3.2 Compositional Data Analysis

Compositional data (co-data) describe parts of some whole and only carry relative information. They can be presented as vectors of proportions, percentages, concentrations, or frequencies. Compositional data have special characteristics and need to be taken into account when analysing proportions, which sometimes exhibit unusual behaviour. Since proportions are expressed as real numbers, an apparent conclusion may be to analyse and interpret them as real multivariate data. The effect is that this practice can lead to paradoxes and/or misinterpretations (Pawlowsky-Glahn et al., 2015).

The components/vectors that make up a composition will be correlated with one another, as a result of the sum constraint. These correlations do not "mean" anything, they are just because of the sum constraint. If we analyse the vectors directly, without any transformation, we can end up finding "spurious correlations;" that is, significant relationships that are just due to the sum constraint. The basic interest with compositional data is to analyse the ratio, not the individual components.

There are two main ways to do compositional data analysis: the logratio approach and the staying-in-the-simplex approach. The logratio approach transforms the constrained $\mathbb{R}^n$ space into the unconstrained $\mathbb{R}^{n-1}$, while the 'staying-in-the-simplex' considers working directly in the simplex. The staying-in-the-simple approach is an alternative to the logratio transformation techniques, but we obtain similar inferences whether we adopt a staying-in-the-simplex approach or a transformation technique (Aitchison, 2003).

3.2.1 Definition. (*D-part composition*). A (row) vector, $x = [x_1, x_2, \ldots, x_D] \in \mathbb{R}^D_+, \quad x_i > 0, \quad \forall\, i = 1, 2, \ldots, D$, is a *D-part* composition when all its components are strictly positive real numbers and carry only relative information.

3.2.2 The Simplex Sample Space.

The sample space for compositions is the *simplex*. This is the set of vectors of positive (or zero) components and constant sum κ, and defined by:

$$S^D = \left\{ x = (x_1, \ldots, x_D) \in \mathbb{R}^D \middle| x_i > 0, \sum_{i=1}^{D} x_i = \kappa \right\} \tag{3.2.1}$$

The simplex has a different structure to $\mathbb{R}^n$, with uniquely defined operations analagous to vectors in $\mathbb{R}^n$. Examples are closure, perturbation and powering operations. See Aitchison (2003) for details.

A usual practice with co-data analysis is to *close* the amounts in the parts of interest to sum up to a constant sum representative (κ). This operation is what is called 'closure.' For any D part vector $z = [z_1, z_2, \ldots, z_D] \in \mathbb{R}^D_+,\ z_i > 0, \quad \forall\, i = 1, 2, \ldots, D]$, the closure of z to $\kappa > 0$ is defined as:

$$\mathcal{C}(z) = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^{D} z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^{D} z_i}, \cdots, \frac{\kappa \cdot z_D}{\sum_{i=1}^{D} z_i} \right] \tag{3.2.2}$$

3.2.3 Principles of Compositional Data Analysis.

A compositional problem recognises the sizes of its components to be irrelevant. Hence its analysis should be independent of κ or even independently of whether the closure was applied or else the data vectors sum up to different values (Van den Boogaart and Tolosana-Delgado, 2013). Any statistical method applied to compositions must satisfy three conditions: scale invariance, permutation invariance, and subcompositional coherence (Aitchison, 1986).

(a) *Scale Invariance:* The basic principle of co-data analysis is that: any meaningful scale-invariant function $f(\cdot)$ for any composition $x \in \mathcal{S}^D$ must yield the same result for all compositionally equivalent vectors. That is, $f(\alpha x) = f(x)$ for any positive proportionality constant, $\alpha \in \mathbb{R}^+$. This is only possible if $f(\cdot)$ can be expressed as a function of logratios of the parts in x.

(b) *Permutation Invariance:* Standard compositional analysis does not consider the information considering ordering. Permutation invariance requires that we obtain equivalent results even if the ordering of the parts in a composition are changed.

(c) *Subcompositional Coherence:* Subcompositional coherence ensures that studies performed on subcompositions do not contradict with those performed on the full composition. Three implications of this condition are: (i) the distance measured between two full compositions must be greater than when measured in any subcomposition (the Euclidean distance defaults here); (ii) the total dispersion of a D-part compositional dataset must be higher than the dispersion in any subcomposition; and (iii) addition of a new non-informative component to a model fitted to a D-part composition should not change the result (Van den Boogaart and Tolosana-Delgado, 2013).

3.2.4 Logratio Analysis of Compositional Data.

Logratio analysis for compositional data problems arose due to the realisation of the importance of the principle of scale invariance and required working with ratios of components. This transformation technique with logratios of the components of a co-data was motivated by the fact that logarithms of ratios are mathematically more tractable. Two main transformations are the additive logratio transformation (alr) and the centred logratio transformation (clr) (Aitchison, 2003). In an attempt to overcome the limitations of these transformations, Egozcue et al. (2003) came up with the isometric logratio transformation; but interpretation of results with this technique is very difficult.

As outlined by Aitchison (2003), the procedure for performing logratio analysis is as follows:

(a) Formulate the compositional problem in terms of the components of the composition.

(b) The formulation is then translated into terms of the logratio vector of the composition.

(c) Transform the compositional data into logratio vectors.

(d) Next, analyse the logratio data by an appropriate standard multivariate statistical method.

(e) Finally, translate back into terms of the compositions the inference obtained at step (d).

3.2.5 Additive Logratio Transformation.

The additive logratio transformation takes the composition into the whole of $\mathbb{R}^{D-1}$. We can thus use standard unconstrained multivariate analysis on the transformed data. Because of the bijective nature of the transformation, we can transfer any inferences back to the simplex and to the components of the composition (Aitchison, 1986).

Let $x = [x_1, \ldots, x_D]$. Then the additive logratio transformation, $alr : \mathcal{S}^D \to \mathbb{R}^{D-1}$ is defined by:

$$y = alr(x) = \left(\ln \frac{x_1}{x_D},\ \ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right). \tag{3.2.3}$$

Its inverse transformation alr^{-1}, $\mathbb{R}^{D-1} \to \mathcal{S}^D$ is defined by:

$$x = alr^{-1}(y) = \mathcal{C}\,[\,\exp(y_1), \exp(y_2), \ldots, \exp(y_{D-1})\mathbf{1}]; \tag{3.2.4}$$

where $\mathcal{C}$ denotes the closure operation.

If we choose another component x_i as divisor other than the last component, we obtain different *alr* transformations. However, in the end, we will make similar inferences for different reference components (Aitchison, 2003).

For example, the *alr* transformation of a 4-part composition is estimated as:

$$alr(x) = [y_1; y_2; y_3] = \left[\ln \frac{x_1}{x_4}; \ln \frac{x_2}{x_4}; \ln \frac{x_3}{x_4} \right], \tag{3.2.5}$$

and its inverse transformation is obtained by defining

$$x = \frac{[\exp(y_1); \exp(y_2); \exp(y_3); 1]}{\exp(y_1) + \exp(y_2) + \exp(y_3) + 1}. \tag{3.2.6}$$

However, a drawback with this technique is that the transformation is asymmetric in the parts and not isometric. This is because the reference part x_D, is the divisor of the logratios of the components. Sometimes, it is more convenient to treat the parts symmetrically and this can be achieved by the centred logratio transformation.

3.2.6 Centred Logratio Transformation.

The centred logratio transformation, $clr = \mathcal{S}^D \to \mathbb{U}^D$ is

$$z = clr(x) = \left(\ln \frac{x_1}{g(x)}, \ldots, \ln \frac{x_D}{g(x)} \right), \qquad g(x) = \prod_{i=1}^{D} x_i^{1/D};$$

where

$$\mathbb{U}^D = \{[u_1, \ldots, u_D] \mid u_1 + \ldots + u_D = 0\}$$

is a hyperplane of $\mathbb{R}^D$. Its inverse transformation $clr^{-1} : \mathbb{U}^D \to \mathcal{S}^D$ is defined by:

$$x = \mathcal{C}[\,\exp(z_1), \ldots, \exp(z_D)].$$

For 3-part composition, the *clr* transformation obtained by:

$$clr(x) = z_i = \ln\left(\frac{x_i}{\sqrt[3]{x_1 x_2 x_3}}\right), \tag{3.2.7}$$

and its inverse transformation by

$$x_i = \frac{\exp(z_i)}{\exp(z_1) + \exp(z_2) + \exp(z_3)}. \tag{3.2.8}$$

Similarly to the additive logratio transformation, this transformation to a real space allows us to use standard unconstrained multivariate methods. A major drawback is that we have a constrained transformed vector (Aitchison, 2003).

4. Analyses, Findings and Discussions

In this chapter, we present the analyses and discussions, as well as findings from the results obtained. All analyses were done using R. We performed a cluster analysis, and applied logistic regression on cluster results. Next, we run compositional analysis and applying the techniques of multivariate linear regression.

4.1 Data Description

The data used for this study is a two-year panel collected on 171 customers' purchases of four chocolate brands: Cadbury, Kit Kat, Lindt and M&M. It consists of the number chocolates bought by a customer and the price paid in each year. A customer's profile is characterised by the state he or she lives in, the number of kids in a house, the number of people in a house, life description, shopping pattern description and household spending description. Table 4.1 is a summary of the total number of purchases and spending (in Australian dollar) for both years.

	Counts		Spending (AUD)	
	Year 1	Year 2	Year 1	Year 2
Cadbury	2354	2093	7929.58	6745.97
Kit Kat	600	424	1363.82	891.40
Lindt	607	402	2303.94	1533.72
M&M	587	490	1845.75	1551.41

Table 4.1: Total Number of Purchases and Total Spending on Brands

All panellists made their purchases in one of the following Australian states: Austrialian Capital Territory (ACT), New South Wales (NSW), Northern Territory (NT), Queensland (QLD), South Australia (SA), Tasmania (TAS), Victoria (VIC), Western Australia (WA). The most purchases were made in VIC (26.90%), followed respectively by NSW (26.32%), QLD (18.13%), SA (14.04%), WA (8.77%), TAS (3.51%), NT (1.75%) and finally ACT (0.58%).

The number of kids ranged from 0-7, and the number of people from 1-8. Under life description, young families were the most consumers representing 42.11%. Older families followed with 25.73%, then adult households and older singles/couples respectively with, 15.79% and 14.04%. The least (2.34%) consumers were young singles/couples. The description of customers shopping pattern was such that 36.26% bought more than once a week, 29.24% weekly, 26.32% bought randomly, 7.6% fortnightly and only 0.58% bought monthly. For customer's household spending description, 42.11% spent between \$151-\$250, 25.73% between \$101-\$150. 21.05% spent over \$250+ with 11.11% spending between \$41-\$100.

4.1.1 Distributions and Correlations.

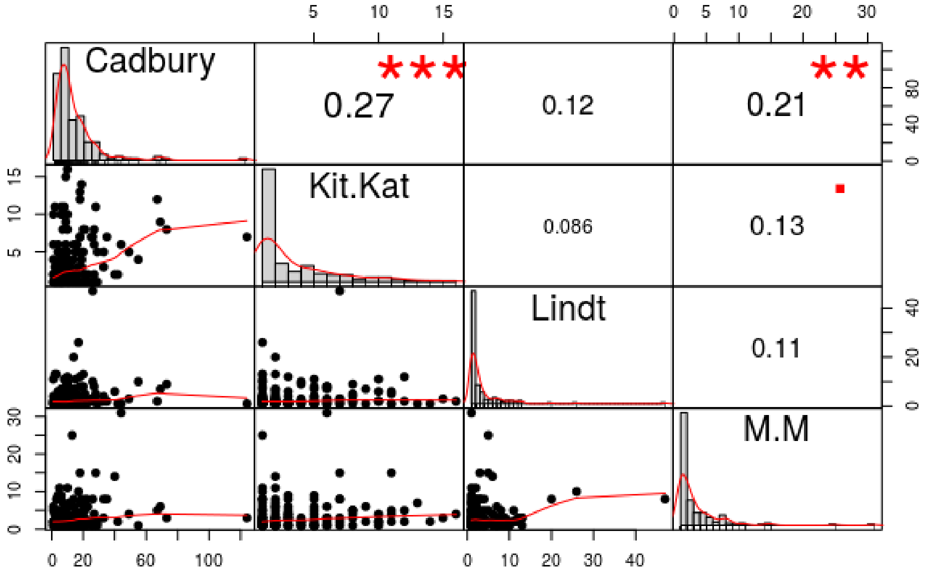

Figure 4.1: Distributions and Correlations of the Number of Chocolates Bought in Year 1

Figure 4.1 above shows the distribution of each chocolate brand on the diagonal. Below the main diagonal are the bivariate scatter plots of the brands with a fitted line. The correlation value and its p-value significance is displayed above the diagonal. A p-value significance is denoted by: (0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’1). The figure was obtained using the *PerformanceAnalytics* package in *R*. From the figure we observe that there is significant correlation between Cadbury and Kit Kat (0.27), Cadbury and M&M (0.21) and Kit Kat and M&M (0.13). We also observe that there is positive correlation between each pair of brands. This connotes that a consumer is likely to purchase at least one of each brand.

4.2 Results and Analysis from Clustering

This research considered the K-means clustering algorithm. The method is known to be one of the best for market segmentation, since it seeks to minimise the within-cluster variation rather than relying on distance measures. However, in order to get the best initial cluster input, we use the stability measures proposed by Brock et al. (2011).

Tables 4.2 to 4.5 give the optimal scores of the best performing algorithms for our dataset. The methods considered were: Hierarchical, K-means, DIANA, FANNY, MODEL, SOTA, PAM and CLARA. After several runs, results showed that the average distance (AD) and figure of merit (FOM) measures suggested high cluster numbers. However, the average proportions of non-overlap (APN) and the average distance between means (ADM) consistently suggested almost the same number of clusters.

	Score	Method	Clusters
APN	0.01	kmeans	2
AD	10.19	clara	6
ADM	1.39	kmeans	2
FOM	6.46	pam	2

Table 4.2: Optimal Values: Year 1 Purchases

	Score	Method	Clusters
APN	0.01	hierarchical	2
AD	9.99	clara	5
ADM	1.40	diana	3
FOM	6.45	model	4

Table 4.3: Optimal Values: Year 2 Purchases

	Score	Method	Clusters
APN	0.01	diana	2
AD	33.37	pam	6
ADM	4.21	hierarchical	2
FOM	21.58	diana	6

Table 4.4: Optimal Values: Year 1 Prices

	Score	Method	Clusters
APN	0.01	hierarchical	2
AD	32.53	pam	6
ADM	1.54	hierarchical	2
FOM	22.51	diana	6

Table 4.5: Optimal Values: Year 2 Prices

4.2.1 Clustering with K-Means.

We chose K=2 for the K-means algorithm based on the APN and ADM measures. The results showed that the majority of consumers preferred Cadbury chocolates. The two distinct clusters suggested 'Heavy' Cadbury consumers and 'Moderate' Cadbury consumers. The results with K=3 also suggested that we would have had to label our clusters as 'Heavy' Cadbury consumers, 'Usual' Cadbury consumers and 'Low' Cadbury consumers. We opted for K=2 in order that 'Moderate' Cadbury Consumers would represent 'Usual' and 'Low' Cadbury consumers.

The K-means no doubt performs good clustering. In R, it is ideal to use the *set.seed()* function to obtain the same results. However one challenging aspect is that, if you run the algorithm for some number of times, it may swap the cluster means, but maintains the same values.

Clusters sizes: 9, 162

Clusters means:

	Cadbury	Kit.Kat	Lindt	M.M
1	62.56	6.11	4.44	7.67
2	11.06	3.36	3.50	3.20

Table 4.6: Number of Purchases in Year 1

Clusters sizes: 7, 164

Clusters means:

	Cadbury	Kit.Kat	Lindt	M.M
1	82.29	3.71	4.29	7.43
2	9.25	2.43	2.27	2.67

Table 4.7: Number of Purchases in Year 2

Clusters sizes: 15, 156

Clusters means:

	Cadbury	Kit.Kat	Lindt	M.M
1	173.27	10.88	17.70	22.34
2	34.17	7.70	13.07	9.68

Table 4.8: Amount Spent in Year 1

Clusters sizes: 7, 164

Clusters means:

	Cadbury	Kit.Kat	Lindt	M.M
1	266.86	8.35	19.69	27.10
2	29.74	5.08	8.51	8.30

Table 4.9: Amount Spent in Year 2

For the number of purchases in year 1, K-means suggested two cluster sizes of 162 and 9. From cluster 1 ('Heavy' Cadbury) of Table 4.6, we observe that on the average, customers will buy approximately 63 Cadbury chocolates, 6 Kit Kat, 4 Lindt and 8 M&M chocolates. Similarly, from cluster 2 ('Moderate' Cadbury), customers buy approximately 11 Cadbury chocolates, 3 Kit Kat, 4 Lindt and 3 M&M chocolates. We can make similar inferences from the results in Tables 4.7 to 4.9.

We sought to investigate whether or not a customer assigned to cluster i in year 1 would be assigned to the same cluster in year 2. From Tables 4.6 to 4.9, we observe the same overall clusters. The patterns in the data in year 1 were similar to year 2, implying that customers portray similar in year 2 as they do in year 1. The benefit that we obtain from the results is that it helps us to identify customers who spend larger proportions of their spending on a particular brand most.

4.2.2 Model Fitting and Analysis.

In this subsection, we used logistic regression analysis to predict a customer's cluster assignment using some demographics. The results from the fitted model showed non-significance of the demographic predictors on clustering group.

The fitted model is:

$$\log\left[\frac{P(\text{HeavyCadCons}|\textit{dems})}{P(\text{ModCadCons}|\textit{dems})}\right] = \beta_0 + \beta_1(\text{Kids}) + \beta_2(\text{People}) + \beta_3(\text{HouseHoldSpendDesc})$$

where *dems* denotes demographics, with the results displayed in Table 4.10.

Table 4.10 shows a summary of the results of the logistic regression model with cluster group as response and demographics as covariates for the number of chocolates bought in year 1. Also included are estimates of the odd ratios (OR) and their confidence interval (at 5% level) in the last two columns.

	Estimate	Std. Error	z value	Pr(>|z|)	OR	2.5 %	97.5 %
(Intercept)	2.6772	1.3698	1.95	0.0506	14.54	1.33	394.76
NoKidsInHouse	-0.3874	0.4328	-0.90	0.3707	0.68	0.26	1.51
NoPeopleInHouse	0.4832	0.4374	1.10	0.2693	1.62	0.74	4.22
HSpendDesc$151-$250	-1.6295	1.1552	-1.41	0.1584	0.20	0.01	1.36
HSpendDesc$250+	-1.3544	1.3289	-1.02	0.3081	0.26	0.01	3.22
HSpendDesc$41-$100	-0.5762	1.4686	-0.39	0.6948	0.56	0.02	15.27

Table 4.10: Effects of Demographics on Cluster Classification.

We can obtain some valuable information here. The results tells us that, for every unit change in the number of kids in house, the log odds of being a 'Heavy' Cadbury consumer (against 'Moderate' Cadbury) decreases by 0.3874. Also, for a unit increase in the number of people in house, the log odds of being a 'Heavy' Cadbury consumer increases by 0.4832. We interpret the estimates of spending description quite differently. The reference (baseline) covariate is the spending range: $101-$150. So spending between $151-$250 against spending between $101-$150 changes the log odds of being a 'Heavy' Cadbury consumer by -1.6295. We could make similar inferences for the other spending ranges.

Alternatively we could make inferences from estimates of the odds ratio. For example, the estimate for number of kids (0.68) means that, for a unit increase in the number of kids, the odds of being a 'Heavy' Cadbury consumer compared to a 'Moderate' Cadbury consumer increases by a factor of 0.68. We usually do not expect a confidence limit to exceed 1. But values greater than 1 mean that as a covariate increases, the odds of being a 'Heavy' Cadbury consumer also increases (or decreases if less than 1).

We can infer from the lower limit ($0.26 < 1$) of the confidence interval for number of kids that, there is a chance that the direction of the relationship in the population is opposite to what we have observed. This means that we cannot trust that the number of kids increases the odds of being a 'Heavy' Cadbury consumer. If both limits of the confidence interval are above 1, then we can be confident that the direction of relationship that we observed is true in the population. That is, it is likely that having more kids compared to not, increases the odds of being a 'Heavy' Cadbury consumer.

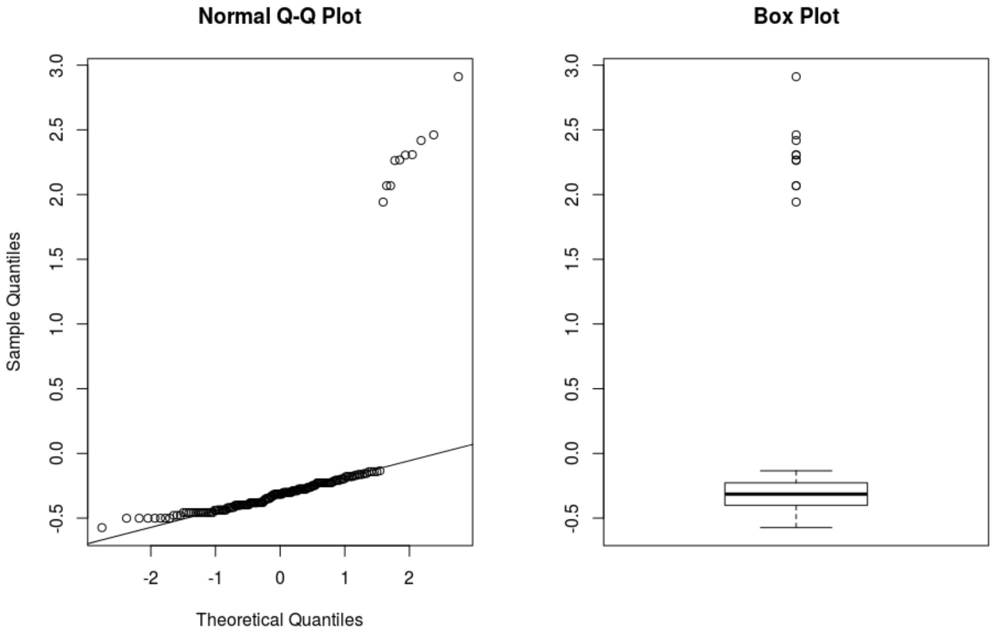

Figure 4.2: QQ/Box Plots of Residuals for Clustering Model

Figure 4.2 suggests a non-normal distribution of residuals, confirmed by a Shapiro-Wilk normality test ($W = 0.44119$, p-value $< 2.2e-16$). The figure suggests that the residuals are right-skewed; hence large observations are not adequately fitted by the model. An alternative would have been to consider a distribution (like the gamma) that can model right-skew data. It should therefore, be noted that any conclusions must be drawn with caution due to the violation of the assumption of normality.

4.3 Results and Discussions from Compositional Analysis

The theory of linear models rests particularly on the assumptions of normality and homoscedasticity. The first step is to check if our data satisfactorily fits a normally distribution. For compositions, we can visually check the assumption of normality on the set of all pairwise logratios with QQ plots, box plots or histograms. This gives us a good global impression of how our data is distributed. In this section, we will investigate the dependence or non-dependence of the buying patterns of consumers in year 2 on year 1 using the counts dataset.

With regards to linear models, compositions can be treated as either response or explanatory variables; and even as both response and explanatory variables. In our case, we treated the chocolate brand compositions in year 2 as the response variables. This allows us to fit a multivariate linear regression to the unconstrained dataset obtained after the *alr* transformation.

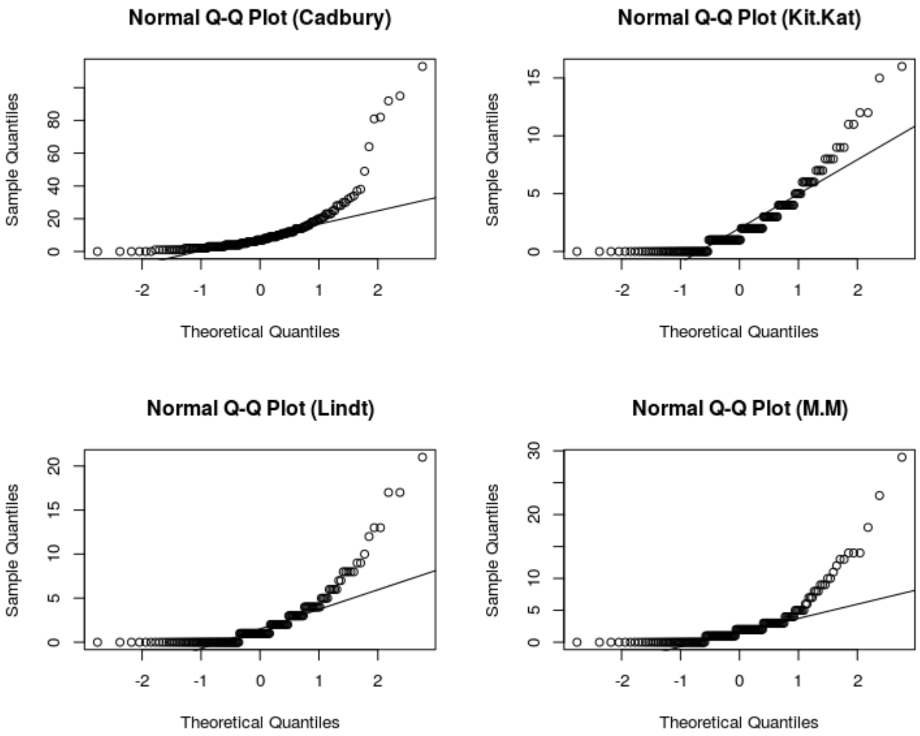

Figure 4.3: QQ Plot of the Compositions as Response of the Number of Purchases (Counts) in Year 2

The counts and spending datasets for year 2 were not normally distributed. Figure 4.3 shows the normal quantile-quantile plot of the number of purchases in year 2. The *alr* transformation of this same counts and spending dataset as shown in Figure 4.4 again depicted a non-normal distribution.

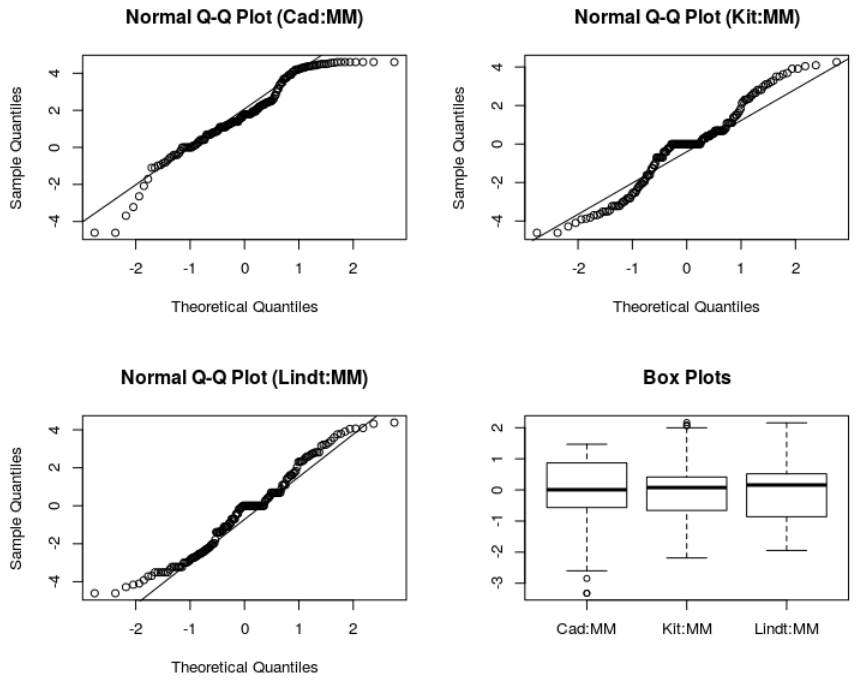

Figure 4.4: QQ/Box Plots of the *alr*-Transformed Compositions for Year 2 Count Dataset.

The original counts and spending datasets in year 1, clearly showed non-normal distributions (similar to Figure 4.3). However, an *alr* transformation to these datasets resulted in an obvious normal distribution. The Henze-Zirkler's, Mardia's and Royston's tests of multivariate normality and the Shapiro-Wilk univariate normality test were consistent with results from the QQ and box plots shown in Figure 4.5.

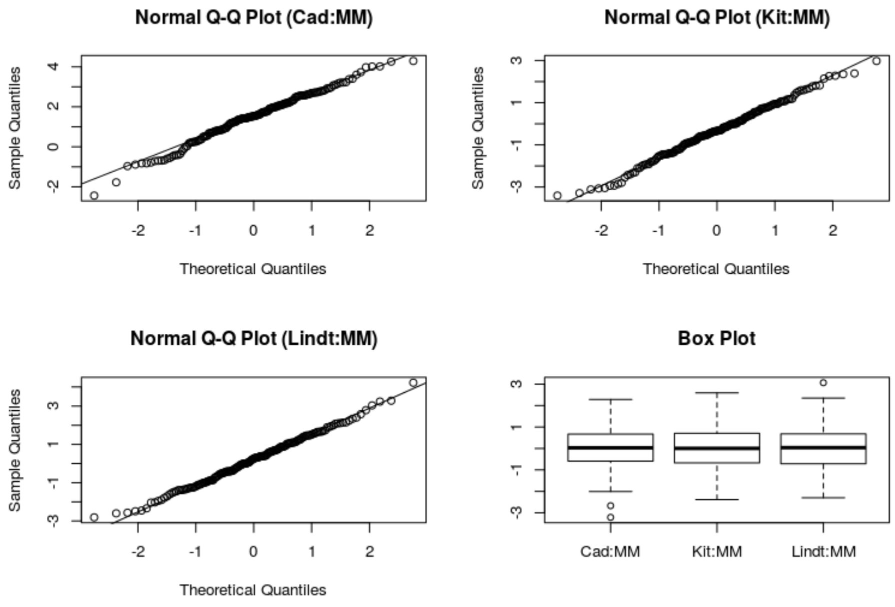

Figure 4.5: QQ/Box Plots of the *alr*-Transformed Compositions for Year 1 Count Dataset.

4.3.1 Model Results of the *alr*-Transformed Data.

This subsections describes the results from the fitted model to the *alr*-transformed compositions dataset. The model considered here is:

```
> model2 <- lm(codata2 ~ codata1 + State + NoKidsInHouse + NoPeopleInHouse +
+            LifeDescription + ShoppingPatternDesc)
```

Interpreting summaries in terms of compositions is slightly different. Here, we analyse each pairwise ratio on a log scale rather than analysing each brand separately (Van den Boogaart and Tolosana-Delgado, 2013). Table 4.11 displays the estimates of the coefficients of the model and their respective indications of significance. Non-significant covariates were ignored.

	Cad:MM	Kit:MM	Lindt:MM
(Intercept)	5.02	3.27	5.19
Cad:MM-Year1	0.81***	0.30·	0.27
Kit:MM-Year1	0.02	0.46**	-0.02
Lindt:MM-Year1	-0.19	0.09	0.63***
StateNSW	-3.76*	-3.93*	-5.52**
StateNT	-3.14	-4.32·	-5.02*
StateQLD	-3.97*	-3.99*	-5.48**
StateSA	-4.11*	-4.69*	-6.24**
StateTAS	-2.13	-4.05·	-4.11·
StateVIC	-3.82*	-3.89·	-5.14*
StateWA	-3.34·	-3.59·	-4.60*
NoKidsInHouse	0.37·	0.15	0.23
NoPeopleInHouse	-0.28·	-0.12	-0.08
Young Families	-0.80	-0.52	-1.73**
Young Singles/Couples	1.35	1.47	2.09*
More than once a week buyers	0.75*	1.18**	0.75·

Table 4.11: Summary Results of Estimates of the Coefficients of the Model

There is lot of information that we can obtain from these results. For instance, we can infer that: a unit change in the number of people living in New South Wales relative to those living in the Australlian Capital Territory in year 1, will significantly change the log of the ratio between Cadbury and M&M consumers by -3.76 in year 2, if all other covariates remain constant. Another inference we can make is that, a unit increase in the ratio of Cadbury to M&M consumers in year 1 on the average, will significantly increase the log of the ratio between Cadbury and M&M consumers in year 2 by 0.81 if all other covariates remain constant. We can make similar inferences from the Kit:MM and Lindt:MM response variables.

Taken together, the coefficients allow us to describe the following:

(a) Cad:MM will tend to be higher where people had a high Cad:MM ratio in year 1, and where people shop more than once a week. It will tend to be moderately lower (i.e. a greater proportion going to M&M) in the states NSW, QLD, SA, and VIC, relative to ACT.

(b) Kit:MM will tend to be higher where people had a high Kit:MM ratio in year 1, and, stongly, where people shop more than once a week. It will tend to be moderately lower (i.e. a greater proportion going to M&M) in the states NSW, QLD, and SA, relative to ACT.

(c) Lindt:MM will tend to be higher where people had a high Lindt:MM ratio in year 1, and with young singles/couples. It will tend to be strongly lower (i.e. a much greater proportion going to M&M) in young families, and in the states NSW, QLD, and SA, relative to ACT, and moderately lower in states NT and VIC.

The strong dependence of ratios in year 2 on the corresponding ratios in year 1 is a clear indication of brand loyalty and the importance of getting customers to try your brand. That is, customers maintain a similar behaviour from year to year. This was also found in the cluster analysis. The remaining significant variables offer marketers some strategic insights e.g. to advertise to young singles in the case of Lindt.

4.3.2 Model Diagnostics.

In this subsection we assess whether our fitted model satisfies the assumptions of multivariate linear regression.

	Kit:MM	Lindt:MM
Cad:MM	0.50	0.57
Kit:MM	1	0.51

Table 4.12: Pearson's Correlation Coefficient of Residuals

A Pearson's test of correlation revealed that, there is true correlation among the residuals. Table 4.12 below displays the correlation results. Figure 4.6 is a scatter plot of the residuals. The plot depicts a positive correlation between the residuals.

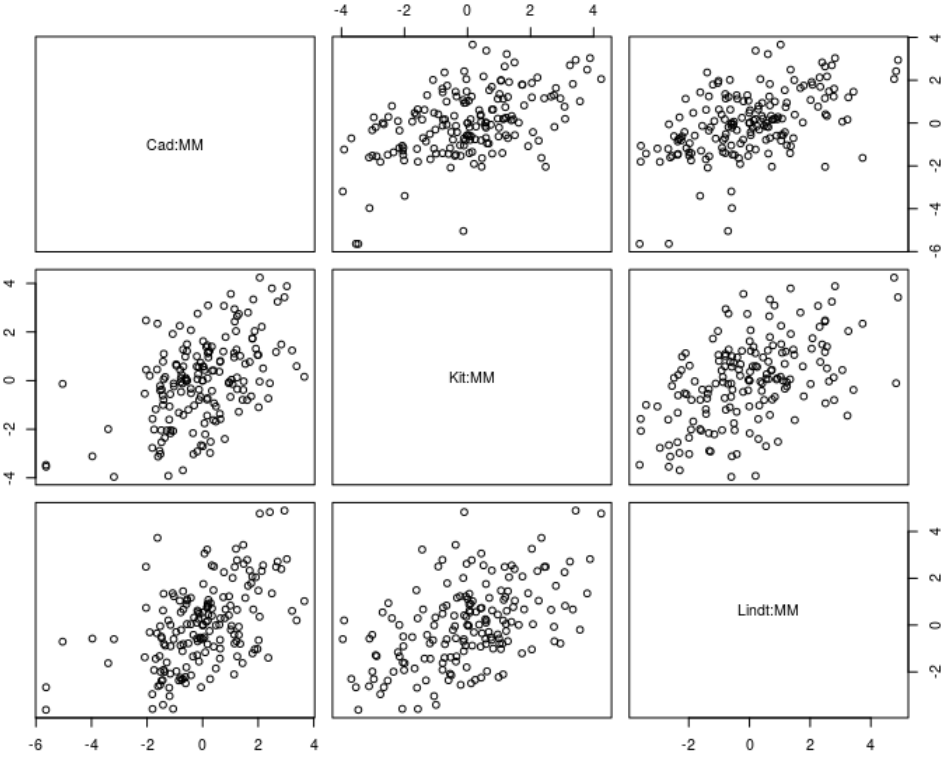

Figure 4.6: Scatter Plot of Model Residuals

It is worth understanding the consequences of fitting a multivariate linear model to a dataset that deviates from the assumptions of multivariate normality. A way to detect deviations from the assumption that errors have the same variance (i.e. homoscedastic) is to plot the residuals against the predicted values. Plots of residuals against predicted values in Figure 4.7 show some potential violations of homoscedacity, particularly in the Cad:MM and Lindt:MM residuals.

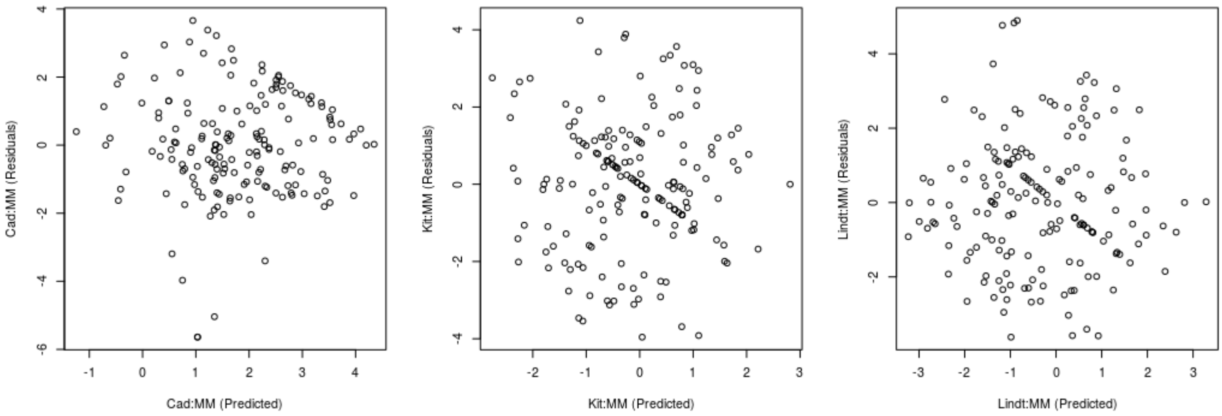

Figure 4.7: Plots of Residuals against Predicted Values

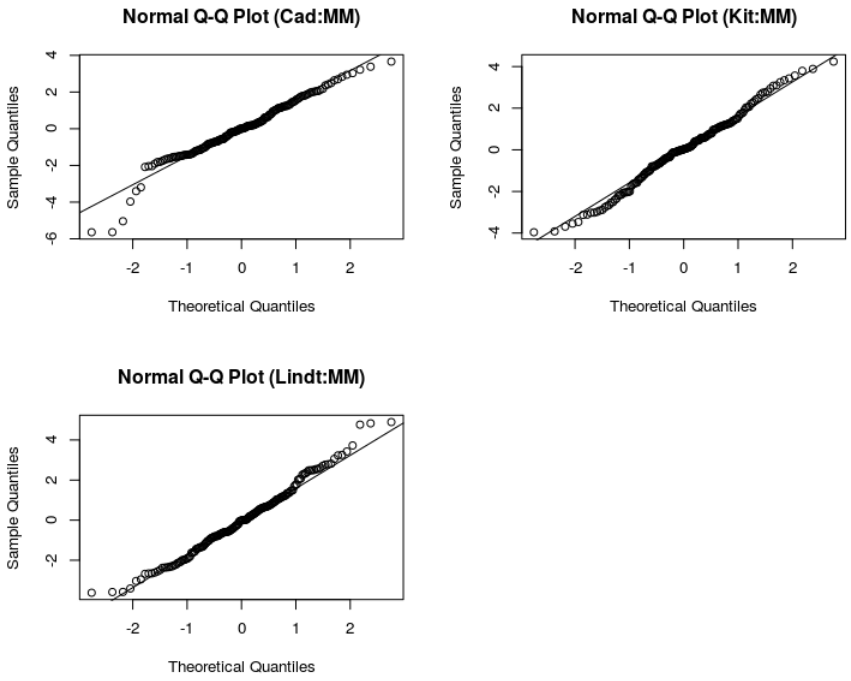

Figure 4.8: Residual QQ Plots

It is important to test the assumption that for any multivariate linear regression model, residuals are normal distributed. The normal QQ residual plot in Figure 4.8 tends to suggest a normal distribution of the residuals (except that of Cadbury:M&M).

To verify this, we perform the hypothesis test that:

- H_0 : residuals are normally distributed,
- H_1 : residuals are not normally distributed.

Results from the Shapiro-Wilk test of normality reveal that all except residuals of Cadbury to M&M are normally distributed (Table 4.13). Our fitted model could thus be useful.

Residuals	Shapiro-Wilk Statistic	p-value
Cad:MM	0.96557	0.0003073
Kit:MM	0.99081	0.3413
Lindt:MM	0.98713	0.1191

Table 4.13: Shapiro-Wilk's Normality Test on Residuals

4.3.3 Model Prediction.

Table 4.14 shows prediction results from the *alr*-transformation for the first ten consumers. We can infer from the table that, the log of the ratio of the number of Cadbury chocolates bought to the number of M&M chocolates bought for say panellist 1, is 1.40. Similarly, the log of the ratio between the number of Kit Kat chocolates bought and that of M&M is -0.01. Such interpretations could be tricky but quite easier when we transform back to compositions (in proportions). Table 4.15 shows the inverse transformation of the *alr* predicted estimates.

	Cad:MM	Kit:MM	Lindt:MM
1	1.40	-0.01	0.80
2	1.91	-1.40	-1.54
3	2.67	0.63	1.81
4	2.82	0.54	-0.32
5	0.94	-0.16	-1.03
6	2.01	-0.56	-1.08
7	1.03	-1.06	-1.95
8	2.50	0.79	-1.08
9	1.67	-1.38	-1.42
10	1.12	-0.02	-0.69

Table 4.14: Predicted Estimates

	Cadbury	Kit Kat	Lindt	M&M
1	0.49	0.12	0.27	0.12
2	0.82	0.03	0.03	0.12
3	0.62	0.08	0.26	0.04
4	0.83	0.08	0.04	0.05
5	0.54	0.18	0.07	0.21
6	0.80	0.06	0.04	0.11
7	0.65	0.08	0.03	0.23
8	0.78	0.14	0.02	0.06
9	0.78	0.04	0.04	0.15
10	0.55	0.18	0.09	0.18

Table 4.15: Inversely Transformed Predicted Estimates

We obtain these back-transformed estimates by using the definitions of Equations (3.2.5) and (3.2.6). For instance, the estimates for panellist 10 in Table 4.15 is evaluated as:

$$\begin{aligned} x_{10} &= \frac{[\exp(1.12); \exp(-0.02); \exp(-0.69); 1]}{\exp(1.12) + \exp(-0.02) + \exp(-0.69) + 1} \\ &= [0.55; 0.18; 0.09; 0.18]. \end{aligned}$$

We can thus say that, on the average, panellist 1 will likely buy Cadbury, Kit Kat, Lindt and M&M chocolates in proportions of 0.49, 0.12, 0.27 and 0.12 respectively in year 2 relative to what he/she bought in year 1. What a manager for say Lindt can do, is to identify consumers that buy large proportions of their product; for example, panellist 7 who buys a good proportion (0.23) of Lindt chocolates.

5. Conclusion

In this essay we sought to investigate the relationship between customer's profile and the spread of their spending on four chocolate brands – Cadbury, Kit Kat, Lindt and M&M. We did this in two ways. Firstly, by clustering together customers with similar profiles and assessing the effect of demographics on clustering groups. Secondly by modelling how consumers spread their spending proportionally across brands.

Previous studies by Bass et al. (1984) had shown that consumers naturally switch brands frequently; and that a consumer's choice of brands followed the zero order process, hence did not depend previous purchase. Uncles et al. (1995) also found out typical regularities in consumer behaviour; one of which is the law of double jeopardy. For this reason 'big' brands have double benefits: they have more users who often use these brands. The comprehensive Dirichlet model is the most widely used approach to modelling consumer purchasing behaviour. But Uncles et al. (1995) and Bongers and Hofmeyr (2010) found that modelling averages such as the Dirichlet sometimes makes systematic under-predictions since they are based on observations alone. Bongers and Hofmeyr (2010) further suggested attitudinal surveys are the best ways to successfully understand consumer behaviour.

The results from our cluster analysis showed two distinct clusters groups which we termed as 'Heavy' Cadbury consumers and 'Moderate' Cadbury consumers. It was observed that a very small number of consumers are 'Heavy' Cadbury chocolate consumers, whereas a very large group are 'Moderate' Cadbury chocolate consumers. Cluster groups between the two years looked similar, such that a customer described as a 'Heavy' Cadbury consumer in the first year was likely to be described the same in the second year. Findings showed that customer's demographics were not good predictors of their cluster groups.

Also, the results from compositional analysis showed a strong pattern of brand loyalty. The ratio of consumers in year 1 strongly associated (positively) with the same ratio in year 2. It was observed that some demographics were significant predictors of brand choice; the most being consumers who purchased more than once a week. For certain brands, young singles/couples and young families proved to be good predictors of brand choice.

Acknowledgements

First of all, let me be religious. Isaiah 65:24 (KJV) speaks of it: "*And it shall come to pass, that before they call, I will answer; and while they are yet speaking, I will hear*." Thank you dear Lord for taking me through. Secondly, I'll be natural. My sincerest gratitude goes to my supervisor: Assoc Prof Ian Durbach for his guidance throughout this essay. "You are a precious gem oh dear Ian." I also appreciate the support from my family and loved ones, for understanding the times and conditions. My kindest appreciation goes to all and sundry in the AIMS community. Your humane love, care and support will forever remain in memory. Finally, I send a deep appreciation to all authors and contributors of the scientific papers and resources that aided this research.

References

J. Aitchison. *The statistical analysis of compositional data*. Chapman and Hall London, 1986.

J. Aitchison. A concise guide to compositional data analysis. https://www.scribd.com/document/3799370/A-Concise-Guide-to-Compositional-Data-Analysis, 2003.

Z. Andy Field and J. Miles. *Discovering statistics using R*. Sage, Thousand Oaks, 2012.

F. M. Bass, M. M. Givon, M. U. Kalwani, D. Reibstein, and G. P. Wright. An investigation into the order of the brand choice process. *Marketing Science*, 3(4):267–287, 1984.

M. Bongers and J. Hofmeyr. Why modeling averages is not good enough. *Journal of Advertising Research*, 50(3):323–333, 2010.

G. Brock, V. Pihur, S. Datta, S. Datta, et al. clValid, an R package for cluster validation. *Journal of Statistical Software*, 2011.

J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

B. Erni, R. Altwegg, and T. Photopoulou. Regression course notes. *Unpublished Manuscript, University of Cape Town*, 2016.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics, Berlin, 2001.

G. J. Goodhardt, A. S. Ehrenberg, and C. Chatfield. The dirichlet: A comprehensive model of buying behaviour. *Journal of the Royal Statistical Society. Series A (General)*, pages 621–655, 1984.

P. Grindrod. *Mathematical Underpinnings of Analytics: Theory and Applications*. OUP Oxford, 2014.

C. Hurlin. Panel data econometrics. *General introduction Panel Data Econometrics Master of Science in Economics - University of Geneva*, 2010.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.

D. G. Morrison and D. C. Schmittlein. Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort? *Journal of Business & Economic Statistics*, 6(2):145–159, 1988.

V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. *Modeling and analysis of compositional data*. John Wiley & Sons, 2015.

O. Torres-Reyna. Panel data analysis fixed and random effects using Stata (v. 4.2). *Data & Statistical Services, Princeton University*, 2007.

M. Uncles, A. Ehrenberg, and K. Hammond. Patterns of buyer behavior: Regularities, models, and extensions. *Marketing Science*, 14(3_supplement):G71–G78, 1995.

K. G. Van den Boogaart and R. Tolosana-Delgado. *Analyzing compositional data with R*, volume 122. Springer, 2013.

R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye. *Probability & Statistics for Engineers & Scientists*. Prentice Hall, 2012.