FACTORS AFFECTING THE PREFERENCE OF LOCAL AND IMPORTED MILK CONSUMTION IN MATARA DISTRICT OF SRI LANKA: A STATISTICAL APPROACH

S. D.M. Dilshani

(158877D)

Degree of Master of Science in Business Statistics (MBS)

Department of Mathematics

University of Moratuwa

Sri Lanka

October 2019

FACTORS AFFECTING THE PREFERENCE OF LOCAL AND IMPORTED MILK CONSUMTION IN MATARA DISTRICT OF SRI LANKA: A STATISTICAL APPROACH

S. D.M. Dilshani

(158877D)

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Business Statistics

Department of Mathematics

University of Moratuwa

Sri Lanka

October 2019

DECLERATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters Dissertation under our supervision.

Name of the Supervisor: Professor L.A.L.W. Jayasekara
Department of Mathematics
Faculty of Science
University of Ruhuna
Sri Lanka.

Signature of the supervisor: Date:

Name of the Supervisor: Mrs. H. V. S. De Silva
Department of Mathematics
Faculty of Engineering
University of Moratuwa
Sri Lanka.

Signature of the supervisor: Date:

ABSTRACT

Milk is one of the most essential foods to humans and it contains many nutrients such as protein, calcium, phosphorus, vitamin B2 and vitamin B12. Intake of a sufficient amount of milk products is recommended for healthy lifestyle of humans. As an agricultural country, Sri Lanka had become self-sufficient in milk, before adopting the open economic policies in 1977. Because of that, imported milk products were highly consumed since 1977 with very lower prices. The government and private sector data indicated that currently in Sri Lanka, local milk production can supply around 42% of the demand and the country depend on the imported milk powder. Therefore, this study was focused on the socioeconomic and other factors (based on the consumer's attitudes) which are influencing consumer's milk pattern either local milk or imported milk. In this study the data were collected through a consumer survey questionnaire in Matara district. At the beginning of the data analysis study, descriptive statistic and chi-square test of independence have done to identify the significant factors which are related with customer's milk consumption behaviors. Then, the Logistic Regression model was fitted on data using R software. Results from fitted multiple logistic regression model show that Age, Monthly Income, price of the milk, Easy to melt, artificial ingredient and Advertisements are the key determinants of consumers milk type.

Keywords: Milk consumption, Binary Logistic Regression, ROC Curve, Hosmer Lemeshow Goodness-of-fit Test

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor Prof. L.A.L.W. Jayasekara for the continuous support of my study and research for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. Besides my supervisor advisor, I would like to thank Mrs. H.V.S. De Silva, for her encouragement, insightful comments. My sincere thanks also go to Prof. T. S. G. Peiris, for dissemination great knowledge to us on Statistics and Statistically Softwares. His guidance and encouragement for the success of this report is very much appreciated. I would like to thank all the lectures, who giving great knowledge to us in Master of Science in Business Statistics Degree. I wish to thank all my friends for their support and encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Appendix

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

1.1.1 Importance of the Milk

Milk is one of the most important food to humans and it fulfills many nutrients such as protein, calcium, phosphorus, vitamin B2 and vitamin B12. Milk provides calcium essential for strong bones, proteins necessary for brain development and tissue growth, vitamin A for normal vision, and vitamin D for the absorption of calcium. Consumption of an adequate amount of milk and milk products are recommended for the healthy lifecycle of humans (Yayar, 2012).

1.1.2. History of Milk in Sri Lanka

In Sri Lanka, as one of the agricultural countries in the world, the dairy milk industry has survived for thousands of years. Sri Lanka had become self-supporting in milk, before implementing the open economic policies in 1977. Since 1977, the imported milk products were highly attracted to local milk products industry with very lower prices. Because of that, the higher demand for inland dairy products had fallen and the Sri Lankan dairy farmers were discouraged. Then dairy milk industry in Sri Lanka was dropped and it had made many damaging effects on the economy. (Pathumsha, 2016)

Before the implementation of an open economy in 1977, Sri Lanka was approximately 80 percent self-supporting in fulfilling the milk requirements. However, in the recent past, it

was decreasing and Sri Lanka is around 40 percent self-sufficient in milk requirements. This has caused of company, which are importing a large amount of powdered milk to the island.

1.1.3. Milk Consumption Patterns in Sri Lanka

At present, Sri Lanka has different kinds of milk consumption pattern. Some of consumers are interested in imported milk and some of them interested in local milk (milk powder and fresh milk). Also, there is a higher demand for milk powder than fresh milk in Sri Lanka. Therefore, the consumption of fresh milk in Sri Lanka is quite low compared to other countries. In Sri Lanka, since daily milk production is not much enough there is a higher demand for imported powdered milk. There are large scale campaigns appointed to promote imported milk powder in different brand names. Therefore, the majority of the consumers in Sri Lanka are depended on the imported milk products.

In some areas, unpacked fresh milk is preferred by some consumers, especially people who live in rural areas. Unpacked fresh milk is mainly delivered by individual farmers to the customers and it is cheaper than packed milk. The other advantage is that these are delivered at the doorstep with no additional cost. Furthermore, there is no packing cost or processing cost. Hence, unpacked fresh milk is distributed much cheaper than processed milk. Therefore, especially the families with a low-income level, select unpacked fresh milk as their primary milk source. Lack of consumers selecting packed milk than unpacked milk because it's a guarantee of quality, safety, packaging and also store. The need to purchase a safe food product is also a major reason to prefer packed milk (Yayar, 2012).

The milk choice of the consumer depends on different factors such as a person's attitude and socio-economic factors. Furthermore, the education, age, monthly income and other characteristics may be affected to consumers influence for milk consumption pattern. On the other hand, some factors such as increasing consumer awareness and concerns about healthy lifestyle and advertising play very important roles for consumer's milk choices. Today, in

developed countries fresh milk consumption pattern has changed. Because of some factors such as health concerns, increasing educated society and income level factors, low-fat milk consumption has shown an increase but per-capita consumption of whole-fat milk has decreased.

1.1.4. Milk Production in Sri Lanka

In Sri Lanka, the total milk production in 2015 has declined by 4% compared to 2014. The volume of proper milk collection has increased only by 1% in 2015, which result may be affected as a consequence of the negative growth rate of the dairy milk sector. However, the dairy sector has shown significant development in the country for the last few years, but it wasn't sufficient to fulfil customer's requirement. Of the total milk that is available, the volume of milk entering the formal milk market in 2015 was around 218.4 million liters. (Tiskumara, 2015).

The total milk production in Sri Lanka has increased by 3.2% which is to 396.2 million liters in 2017, compared to 2016. Results may be affected due to policy actions such as distributing high-yielding cows and increasing the guaranteed price of milk to farmers, which is highly affected to increase a higher private sector investment into the dairy sector in 2017.

Cow milk, which accounted for 82.7 per cent of the total milk production, increased by 3.1 per cent to 327.6 million liters, while buffalo milk production, which accounted for the rest, increased by 3.7 per cent to 68.6 million liters. The Department of Animal Production and Health (DAPH) estimates that domestic milk production was sufficient to cover 40 per cent of milk consumption of the country during 2017, while the rest was depended on imported milk powder. However, 42 per cent of the domestic milk consumption was met with domestic sources in 2016, highlighting the need for continued efforts in improving domestic production to meet the government objective of increasing food security.

Table 1.1: Annual Milk Production in Sri Lanka from 1998 to 2017

year	Total annual Milk production (Liter)
1998	177,089,045
1999	179,883,600
2000	181,455,748
2001	183,027,600
2002	183,195,000
2003	186,804,000
2004	190,296,000
2005	192,741,600
2006	196,623,360
2007	202,009,200
2008	208,093,090
2009	233,316,240
2010	247,554,000
2011	258,303,600
2012	258,303,600
2013	329,169,600
2014	333,903,600
2015	374,443,200
2016	384,008,400
2017	396,198,000

Source: Agriculture and Environment Statistics Division
Department of Census and Statistics, Sri Lanka

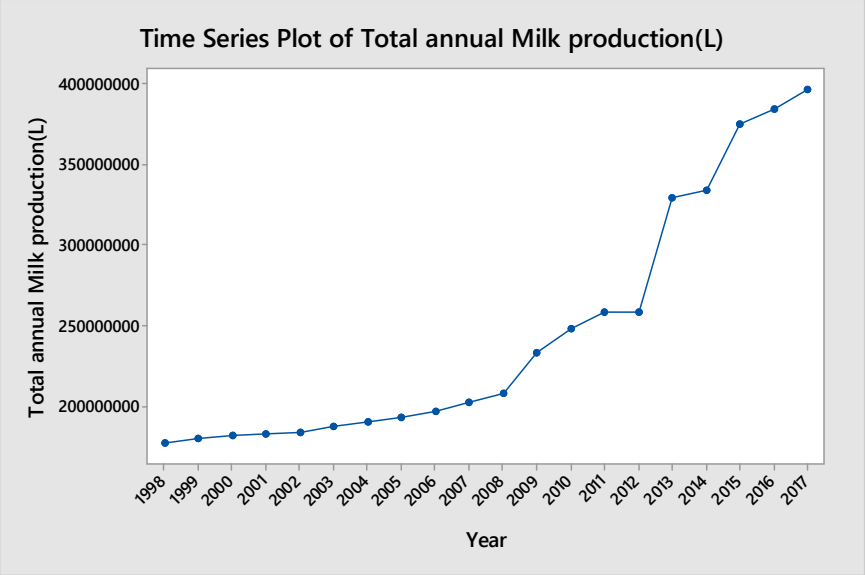


***Figure 1.1*: Time Series Plot for Annual Milk Production in Sri Lanka from 1998 to 2017**

According to figure 1.1, milk production is gradually increasing from 2008 to 2017. There were some reasons for the above result, but the main reason was maybe, end of the war.

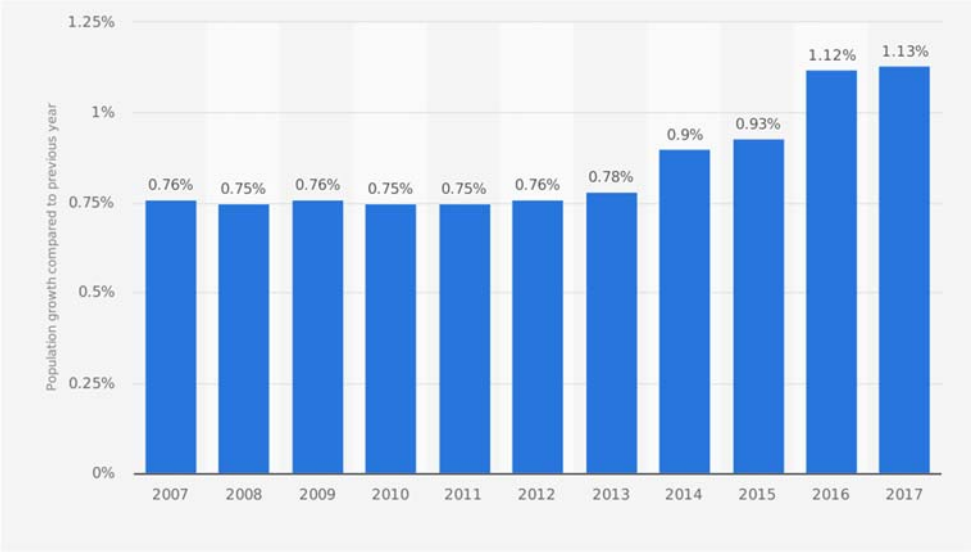


***Figure 1.2*: Population growth from 2007 to 2017(compared to previous year) in Sri Lanka.**
Source: https://www.statista.com/statistics/728536/population-growth-in-sri-lanka/

According to the figure 1.2, the population growth rate is slightly increased up to 2013, but it is gradually increasing from 2013 to 2017. According to the statistics, the local milk production increased throughout the past few years, but it is not sufficient to fulfil for consumer's requirement given the growth rate of the population.

Imports of milk have shown a fast growth increasing per-capita consumption of milk from 45.16 Liter per year in 2014 to 48.56 Liter per year in 2015. Imported milk and milk products have been increased nearly by 22% compared to 2014 while the value of imports reduced by 23%. Unfortunately, 61% of milk and daily requirement depended on imports in 2015. Thus, the market increase of imports may attribute to decreased international market prices of milk and milk commodities as well as to consumer preference towards powdered milk. As a result, even with a negative growth of the sector, the annual per capita availability of milk has increased. (Pathumsha, 2016).

Meanwhile, milk powder imports decreased by 0.9 per cent to 93,127 metric tons in 2017. The Department of Animal Production and Health and the government will start mega farms with imported cattle of higher production capacity with new technology, the sector is expected to grow at a faster rate.

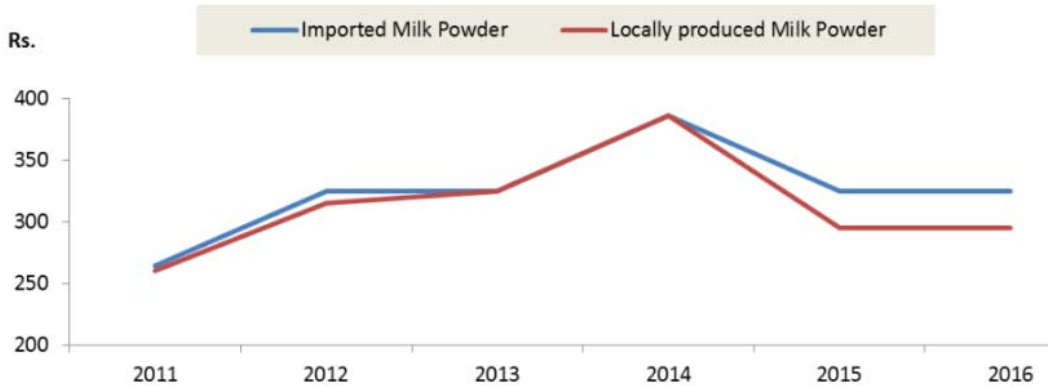


Figure 1.3: Comparison of Prices of Imported milk powder and locally produced milk powder from 2011 to 2016 (Maximum Retail Price of 400g Milk Powder Pack)
Source:https://www.research.advocata.org/price-controls-in-the-dairy-industry/

According to figure 1.3, both imported and local milk prices were increased from 2011 to 2014, meanwhile, it is also represented that the imported milk price was higher than local milk prices. However, in 2014 represents the peak prices for both local and imported milk. After that the prices were decreased from 2014 to 2015, price have remained the same level.

1.1.5 World Milk Production

In the last three decades, world milk production has increased by more than 50 per cent, from 500 million tonnes in 1983 to 769 million tonnes in 2003. India is the world's largest milk producer, with 18 per cent of global production, followed by the United States of America, China, Pakistan and Brazil. Since the 1970s, most of the expansion in milk production has been in South Asia, which is the main driver of milk production growth in the developing world. Milk production in Africa is growing more slowly than in other developing regions, because of poverty and in some countries due to adverse climatic conditions. The countries with the highest milk surpluses are New Zealand, the United States of America, Germany, France, Australia and Ireland. The countries with the highest milk deficits are China, Italy, the Russian Federation, Mexico, Algeria and Indonesia.

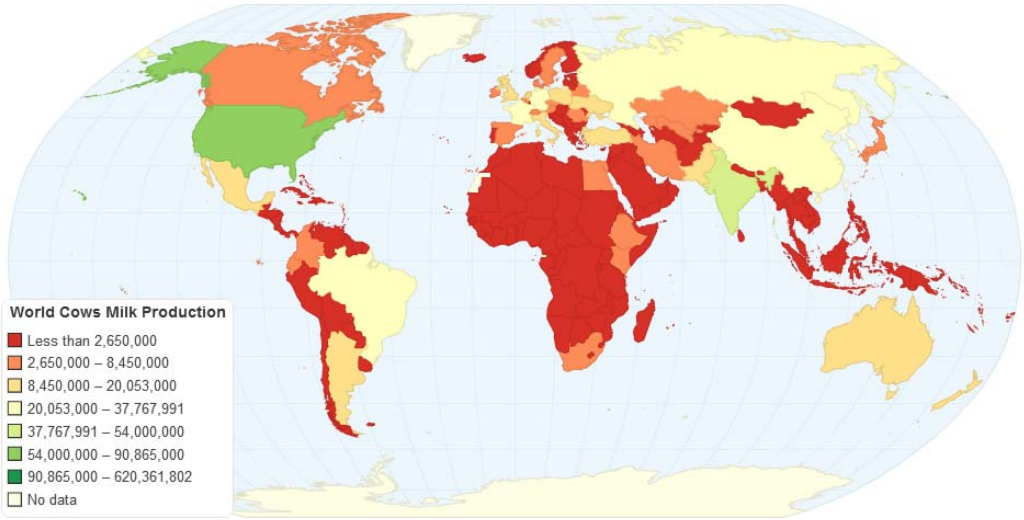


Figure 1.4: World Milk Production in tonnes
Source: http://chartsbin.com/view

1.2 Objectives

This study focused on socio-economic and demographic factors influencing consumer's different milk consumption preference in Matara district. On the view of the above, the objectives of the study are to;

- Identify the socioeconomic factors, influencing on consumer's milk choices.
- Identify how these socioeconomic factors affect with milk consumption pattern in Matara district.

1.3 Outline of the Dissertation

The aim of this thesis is to identify the factors affecting the preference of local milk and imported milk consumption in Matara district. We will specially focus on binary logistic regression approach. The first chapter mainly described the milk consumption patterns in Sri Lanka and the milk production in Sri Lanka as well as in the world. Sri Lanka has served milk consumption patterns such as fresh milk, local milk powder and imported milk powder. Because of the negative growth rate of local milk, the majority of the consumers in Sri Lanka depended on the imported milk products. In chapter two use present related literatures, which have reviewed for complete this study. In third chapter, data collection and statistical methods were discussed and this study mainly carried out the binary logistic regression since the response variable is a binary variable.

In chapter four carried out all the statistical results in this study. In this analysis, there were 21 covariates have considered and at the first, the simple descriptive analysis has done for defined covariates. Then the Chi-square test of independence was used to investigate the effects of socioeconomic characteristics on consumers' local and imported milk consumption behaviour. The multinomial logistic regression model was used to determine

the extent, how selected socioeconomic characteristics of consumers influence these milk types. At this step, the forward selection criteria and the backward selection criteria applied to select most significance covariates. The Likelihood Ratio test and Akaika Information Criteria were applied to selected most significance model. Then assessing the fitted model were done using Pearson's Chi-squared test, Hosmer Lemeshow test and Roc curve.

In the last chapter included the general conclusion for this study. The data were collected using questionnaire and the analysis has done using RStudio software. The relevant source code and the sample questionnaire are included at the end of this thesis.

CHAPTER 2

LITERATURE REVIEW

2.1 Milk Consumption Pattern

There are different studies on milk consumption pattern in different countries. The choice of milk consumption and preference can be categorized as packed or unpacked milk, local or imported milk, fresh milk or powdered milk. There is a study which has focused on consumption of packed milk and unpacked milk (Yayar, 2012). In this survey studied the factors affecting fluid milk purchasing sources concerning packed and unpacked fluid milk produced in Turkey households. As a drink, fresh milk has competition from soft drink and powdered drink. The soft drink production has increased and additionally, their low material cost helps to extensively promoted. Also, there are different campaigns to promote milk powdered with different brand names. Therefore, there is a higher demand for powdered milk (De Alwis A.E.N, 2009). However, the nutritive values are destroyed by heat. Therefore, De Alwis and others (2009) suggest fresh milk is the most nutritive milk than others.

2.2 Effect of Socioeconomics Characteristics on Milk Consumption Pattern

There are many studies on the effect of socioeconomic characteristics on milk consumption pattern and preferences. Many studies have investigated consumers' attitudes toward aggregate fluid milk purchases and consumption (De Alwis A.E.N, 2009; Yayar, 2012; Health D, 2012). One finding of these studies is that socioeconomic and demographic factors can be important in determining consumer's preference and milk consumption.

According to the study, De Alwis (2009), has focused on analysing the consumer attitudes, demographic and economic factors that affect fresh milk consumption among the mid-country consumers of Sri Lanka. Some studies are highlighted that consumers' attitude and their beliefs are affected to predict their consumption pattern. There should be an attention to the development of fresh milk consumption to promote a more healthful lifestyle. This study was developed to hypothesize that fresh milk consumption is associated with socioeconomic and demographic factors of consumers. In this study, the factor analysis was carried out to introduce to the weight up the consumer attitude and factor scores. As a final model, it proposed that the independent variables categorized as cost and usage, nutrition, sensory factor and availability.

In this study, the logistic regression was applied to find out the relationship between fresh milk consumption and socioeconomics, demographics and attitudinal factors of the individual consumers. The results indicated that gender and household size did not significantly impact on fresh milk consumption. The consumers age was affected with consumption of fresh milk. Results showed that consumers with higher income level are more preferable with fresh milk than lower income level. Further, household composition is related to fresh milk consumption, which is increasing with the probability of fresh milk consumption. But consumers who had health problems had less interested in fresh milk. Furthermore, De Alwis (2009) found out that the increase in cost and usage affected by reducing the probability of fresh milk consumption. De Alwis mentioned that some previous studies showed, the 95% consumers believed, their risk of certain diseases may be reduced by nutrition foods. However, the consumers consider such kinds of health benefits, and they assess other products based on some characteristics such as appearance, price, taste and naturalness.

Bus and Worsley (2003) found out that perceptions such as cost, family habits, nutritional awareness, beliefs and perceived sensory properties affect milk consumption and attitudes of different types of milk. Through a review of the study, it mentioned that the majority of consumers had positive responses in the taste of milk. Besides, women have more positive

beliefs about price, taste, health, and nutrition than men. Furthermore, the type of milk consumption is influence with socio-demographic variables such as gender, age, education level, socio-economic status and ethnicity. This study focused on the perception of the milk among food shoppers. Kruskal_Wallis test and Chi-square analysis were performed to examine the consumers' perception of the three types of milk (whole milk, reduced-fat milk and soy milk) with demographic factors. There was a significant interaction between educational level and type of milk consumption. Low-educated consumers' had lower interest on reduced-fat milk and whole milk than tertiary-educated consumers. In this study majority of the consumers agreed that dairy milk has good sensory properties. Among them, whole milk has highest agreement on taste, although reduced-fat milk closely followed.

The review of the study (Yayar, 2012), the chi-square test of independence was used to investigate the effects of socioeconomic and demographic characteristics on consumer's packed and unpacked fluid milk behaviour. The results show that the cross-tabulations of unpacked fluid milk, packed fluid milk and unpacked-packed fluid milk choices considering households socioeconomic and demographic characteristics. All of the socioeconomic and demographic variables were statistically significant at the 5% level of probability. Furthermore, Multinomial logistic regression model was used to analyze household's packed and unpacked milk consumption decisions as a function of socioeconomic and demographic factors. Also, Yayar (2012) found out that the consumer with higher education and small families were more prefer for packed milk. The large families were more like for unpacked milk. The results show that household with a middle income had a negative impact on unpacked fluid milk consumption, which means they less likely to purchase unpacked fluid milk than lower income households. The non-working housewives are interested in non-packed milk, others were more inclined to choose packed milk (Yayar, 2012).

Jane and Yu (2006) found that the fluid milk consumption patterns and attribute perception of responses can be explained by under three segments. The highest percentage of housewives, senior high school graduates and shoppers more preferred purchasing fluid milk

at the supermarket. Also, higher household incomes and large household sizes have appeared in the same pattern. It mentioned that higher educated household shoppers more tended to reduced flavour milk consumption. The price of milk is another influencing factor in milk consumption. The consumers who purchase less fresh milk are more influenced by price. Jane and Yu (2006) pointed out that the shoppers were more interested to purchase a large quantity of high-quality brands of fresh milk.

CHAPTER 3

METERIAL AND METHODS

3.1 Data Collection

The participants were selected from Matara district surveying from March to May 2018. The development of a questionnaire was based on a qualitative study of consumer's milk choices. The final questionnaire mainly consisted of three parts: (1) the personal factors, which are related to the milk choice; (2) socioeconomic factors that are likely to influence consumer's milk choices; (3) daily milk consumption pattern. There are 21 factors included in the questionnaire. The questionnaire is shown in the appendix. Before collecting data, a pilot survey was carried out using a group of randomly selected consumers and these pre-tested surveys were not included in the final data set.

A random sample of 421 households was surveyed. Through the questionnaire, consumers answered questions about their choices of purchasing milk alternatives and provided socioeconomic information.

3.1.1 Involved Variables in the Model Building Process

Response Variable

- Daily milk consumption

This is a categorical variable and according to the responses of the consumers there are two categories; those are (1) consumption local milk and (2) consumption imported milk.

Predictor Variables

In the first section of the questionnaire there are two numerical variables, which are;

- Age
- Number of members in the family

The other categorical variables are as follows;

- Gender
- Marital Status
- Educational attainment
- Educational attainment of head of the household
- Monthly Income

The second part of the questionnaire includes some factors, which are related to the consumer's milk choice. In this part, we asked about the consumer's opinion about their milk choice. Each factor consists of three categories, which are "agreed", "neither agree nor disagree" and "disagree". Based on the given question, "Are you considering the following factors for your milk choice?", then the Consumer should select one category according to the following factors.

- Good quality
- Reasonable price than others
- Taste
- Nutrition
- Thickness
- Easy to melt
- Smell
- Easy to buy in the market
- Well-known brand name
- Easy to use/store
- Consider about artificial ingredients
- Influence by others (friends, relations)
- Affected by advertisements

The characteristics of the data set are as follows. The response binary variable Y represented with 1, stands for local milk and 0, stands for imported milk is as follows.

$$Y = \begin{cases} 1, & Local\ milk \\ 0, & Imported\ Milk \end{cases}$$

There are two types of predictor variables, which are continuous and categorical predictor variables. Age and Number of members in the family are two continuous predictor variables and there are 17 categorical predictor variables identified in this survey. Table 3.1 shows the description of the involved variables and categories of the categorical variables in this analysis.

Table 3 .1: Description of the Response variable and Predictor Variables.

Variable Name	Description	Categories
Age	Age	
Gender	Male or Female	
No_Members	Number of family members in the family	
Education	Consumer's Educational level	Up to O/L (below O/L) Up to A/L Graduate/Postgraduate (Professional)
H_Education	Education level of the head of the family	Up to O/L Up to A/L Graduate/Postgraduate (Professional)
Monthly_Income	Monthly Income	Less than Rs 35,000 Rs 35,000-50,000 Rs 51,000-65,000 Rs 66,000-80,000 (Greater than Rs 80,000)
Quality	Quality of milk	Agree Neither agree nor disagree (Disagree)
Price	Price of milk	Agree Neither agree nor disagree (Disagree)
Taste	Taste of milk	Agree Neither agree nor disagree (Disagree)
Nutrition	Nutrition level of milk	Agree Neither agree nor disagree (Disagree)
Thickness	Opinion on thickness	Agree

		Neither agree nor disagree (Disagree)
Easy_melt	Easy to melt	Agree Neither agree nor disagree (Disagree)
Smell	Smell	Agree Neither agree nor disagree (Disagree)
Easy_buy	Easy to buy	Agree Neither agree nor disagree (Disagree)
Brand_name	Brand Name	Agree Neither agree nor disagree (Disagree)
Easy_use	Easy to Use	Agree Neither agree nor disagree (Disagree)
Arf_ingredient	Artificial Ingredients	Agree Neither agree nor disagree (Disagree)
Advertisement	Advertisement	Agree Neither agree nor disagree (Disagree)
milk_type	Type of milk Consumption	Agree Neither agree nor disagree (Disagree)

The reference category is shown in parenthesis in categorical variables (3rd column in Table 3.1).

3.2 Methodology

3.2.1 Contingency Table

For a single categorical variable, we can summarize the data by counting the number of observations in each category. The sample proportion in the categories estimate the category probabilities.

Suppose there are two categorical variables, denoted by X and Y. Let I denote the number of categories of X and J the number of categories of Y. A rectangular table having I rows for the categories of X and J columns for the categories of Y has cells that display the IJ possible combinations of outcomes. A table of this form that displays counts of outcomes in the cells is called a contingency table. A table that cross classifies two variables is called a two – way contingency table. ; One that cross classifies three variables is called three-way contingency table, and so forth. A two-way table with I rows and J columns is called an $I \times J$ (read I –by-J) table.

3.2.2 Chi-Square Test of Independence

The entries in the cells in a contingency table may be frequencies or proportions. It can be applied for qualitative data classified into two or more categories, or nominal scaled variables. Chi-Square test is not a parametric test, it is a nonparametric test to check whether if the two or more classifications of samples are dependent or independent. Therefore, the hypothesis in this test are;

H_0: The variable 1 and variable 2 are associated

H_1: The variable 1 and variable 2 are not associated

Table 3.2: Contingency Table with Observed frequencies

		Variable 2		
		Category C	Category D	Total
Variable 1	Category A	O_{11}	O_{12}	n_3
	Category B	O_{21}	O_{22}	n_4
	Total	n_1	n_2	N

In Chi- Square test, the corresponding expected frequencies calculate as follows;

Table 3.3: Expected frequencies

		Variable 2	
		Category C	Category D
Variable 1	Category A	$E_{11} = \frac{n_1 \times n_3}{N}$	$E_{12} = \frac{n_2 \times n_3}{N}$
	Category B	$E_{21} = \frac{n_1 \times n_4}{N}$	$E_{22} = \frac{n_2 \times n_4}{N}$

In Chi-Square test of independence, the test is based on chi-square (χ^2) distribution. To compare the observed frequencies and expected frequencies, we can calculate test statistics using the following equation;

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (3.1)$$

In this equation, O_i stands for observed frequencies (Table 3.1), E_i stands for expected frequencies (Table 3.2) and i goes from 1, 2, …, n, where n is the total number of cells in the contingency table. To assess the significance of the test statistics, we refer to the standard chi-square table, which contains the critical χ^2 values for levels of probabilities on different degrees of freedom. For the contingency table with "r" rows and "c" columns, we can calculated the degrees of freedom for the above contingency table using the following formula.

$$df = (r-1)(c-1)$$

After that, we can made conclusion comparing chi-square test statistic and probability level (significance level). If the value of chi-square lies on the probability level, chi-square test rejects the null hypothesis. Therefore, we concluded that the two variables are do not independent each other.

There are some limitations, when we are applying chi-square test of independence. In the standard chi-square table presented the chi-square values, which computed using the equation (3.1) assuming for the large expected values. Therefore, the use of chi-square test

is restricted to large samples. However, there are some ways, when the small samples are considered. The one way is to apply a correction of continuity, also known as *Yates correction.* The most common way is applying *Fisher's Exact Test.* The fisher's exact test is recommended for used when the total sample size is less than 20 or when the one of the expected frequencies less than 5 with sample is less than 40.

3.2.3 Relative Risk

A Risk Ratio or Relative Risk is the ratio of the probability of an outcome in an exposed group to the probability of an outcome in an unexposed group. Risk ratio is used in the statistical analysis of the data of experimental, cohort and cross-sectional studies, to estimate the strength of the association between treatments or risk factors, and outcome. For example, it is used to compare the risk of an adverse outcome when receiving a medical treatment versus no treatment (or placebo), or when exposed to an environmental risk factor versus not exposed (Agresti, 2007).

Assuming the causal effect between the exposure and the outcome, values of RR can be interpreted as follows:

- RR = 1 means that exposure does not affect the outcome;
- RR < 1 means that the risk of the outcome is decreased by the exposure;
- RR > 1 means that the risk of the outcome is increased by the exposure.

For 2 × 2 tables, the relative risk is the ratio;

$$Relative\ Risk = \frac{\pi_1}{\pi_2}$$

where π_1 denoted probability of an outcome in an exposed group and, π_2 denoted the probability of an outcome in an unexposed group.

3.2.4 The Definition of the Odds

An odds ratio (OR) is a measure of association between an exposure (event) and an outcome. The OR represents the odds that an outcome will occur given a particular expose, compared to the odds of the outcome occurring in the absence of that expose.

The odds ratio of an event is defined as follows;

$$odds\ of\ an\ event = \frac{P(event\ occur)}{P(event\ does\ not\ occur)} = \frac{\pi_1}{1-\pi_1}$$

For instance, if π_1 = 0.75, then the odds of success equal 0.75/0.25 = 3. The odds are nonnegative, with value greater than 1.0 when a success is more likely than a failure. For example, when odds = 4.0, a success is four times as likely as a failure. The probability of success is 0.8, the probability of failure is 0.2, and the odds equal 0.8/0.2 = 4.0.We then expect to observe four successes for every one failure. When odds = 1/4, a failure is four times as likely as a success. We then expect to observe one success for every four failures.

3.2.5. The Odds Ratio

In 2 × 2 tables, within 1st row the *odds* of success are $odds_1 = {}^{\pi_1}/_{1-\pi_1}$, and within 2nd row the odds of failure equal $odds_2 = {}^{\pi_2}/_{1-\pi_2}$. The ratio of the odds from the two rows,

$$odds\ ratio = \frac{odds_1}{odds_2} = \frac{{}^{\pi_1}/_{1-\pi_1}}{{}^{\pi_2}/_{1-\pi_2}} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Whereas the relative risk is a ratio of two probabilities, the odds ratio (θ) is a ratio of two odds.

3.2.5.1 Properties of Odds Ratio

The odds ratio can be equal to any real number, which should be non-negative. If $\pi_1 = \pi_2$, that means probability of success of event 1 and probability of success of event 2 are equal, therefore $odds_1 = odds_2$ and $odds_1/odds_2 = 1$. Odds ratios on each side of 1 reflect certain types of associations. When odds ratio greater than 1 ($\theta > 1$), the odds of success of event 1 is higher than odds of success of event 2. For example, when $\theta = 3$ (odds ratio is equal to 3), the odds of success of event 1 is three times the odds of success of event 2. Thus, subjects in first event is more likely to have successes than is subjects in second event; that is, $p_1 > p_2$. When $\theta < 1$, a success of first event is less likely than in second event; that is, $p_1 < p_2$.

When the table orientation reverse, the odds ratio does not change. In this situation, the rows and columns are interchange with each other. So that the rows become the columns and the columns become the rows. The same odds ratio will occur when columns treated as response variable and the rows treated as explanatory variable, or columns treated as explanatory variable and as well as rows as response variable. Thus, it is unnecessary to identify one classification as a response variable in order to estimate θ. By contrast, the relative risk requires this, and its value also depends on whether it is applied to the first or to the second outcome category.

3.2.6 Binary Logistic Regression

Regression analysis is popular and widely used analysis concerned with describing the relationship between a response variable and one or more explanatory variables. In linear regression, the response variable (dependent variable) is continuous. It can have any one of an infinite number of possible values. For instance, weight, height, number of hours, etc. It is often the case that the response variable is categorical in nature, taking on two or more possible values. For instance, yes/no, true/false, red/green/blue, 1st/2nd/3rd/4th, etc. The logistic regression model has become, the standard method of analysis in this situation.

Therefore, the main distinguishes a logistic regression model from the linear regression model is that the response variable in logistic regression is binary or dichotomous.

The goal of an analysis is to find the best model fitting for given data, to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables. These independent variables are also called *covariates*.

In any regression analysis the key point is "what is the mean value of the response variable, given that the value of the predictor variable". This quantity can be expressed as $"E(Y|x)"$ where Y denotes the response variable and x denotes the independent variable. In regression analysis, we can consider this quantity as a linear equation in x and it can be expressed such that,

$$E(Y|x) = \beta_0 + \beta_1 x$$

Where β_0 and β_1 are the unknown parameters of the model.

To simplify notation, in logistic regression, we use the notation $\pi(x) = E(Y|x)$ to represent the above quantity, which is conditional mean of Y given x. The standard logistic regression model form is as follows;

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{3.2}$$

When a binary outcome variable is modeled using logistic regression, it is assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables. In the logistic regression, this transformation is defined, in terms of $\pi(x)$ as

$$g(x) = logit[\pi(x)] = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \tag{3.3}$$

The logit, $g(x)$ is linear in its parameters, may be continuous and may range from $-\infty$ to $+\infty$, depending on the range of x.

According to formula (3.3) the parameter β_1 indicated that the rate of increase or decrease of the S-shape curves (figure 3.1) for $\pi(x)$.

In logistic regression model, we can express the value of the response variable given x , as $y = \pi(x) + \varepsilon$. Here the quantity ε can have one of two possible values, which depends on the value of outcome variable y. If $y = 1$ then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$, and if $y = 0$ then $\varepsilon = -\pi(x)$ with probability $1-\pi(x)$. Therefore, ε has a distribution with mean zero and variance equal $\pi(x)[1-\pi(x)]$. That is the conditional distribution of the outcome variable follows a binomial distribution with probability given by the mean, $\pi(x)$.

3.2.6.1 Use of the logistic curve

Binary dependent variables have only two outcomes. To define the relationship boundary by 0 and 1, logistic regression uses the logistic curve to represent the relationship between the independent and dependence variables.

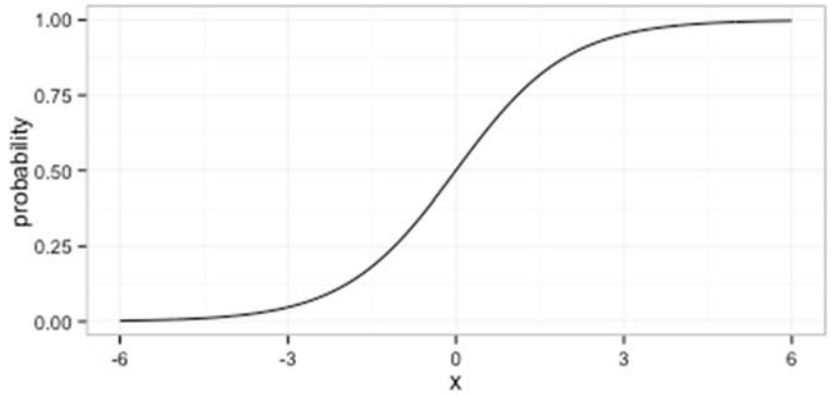


Figure 3.1: Logistic Regression Curve

The logit model uses the specific forms of the logistic curve, which is appearing S-shaped of the model for $\pi(x)$ to study within the range of 0 to 1. It is curved rather than a straight line, the rate of change in $\pi(x)$ per unit when increasing in x, depends on the value of x.

If the values of independent variable with very low levels, then the probability of predicted values approaches to zero, but it never reaches to 0. As well when the independent variable increases, the predicted values also increase up the curve, but then the slope starts decreasing so that at any level of the independent variable, the probability will approach 1 but never go beyond from it.

3.2.6.2 The Binary Logistic Regression Model

Consider a sample of n independent observations of the pairs (x_i, y_i), $i = 1,2,3..,n$, where y_i denotes the value of a dependent variable with two outcomes (dichotomous) and x_i denotes the value of the independent variable for the i^{th} observation. Furthermore, we assume that the two outcomes of the response variable have been coded as 0 or 1.

The general method of estimation that leads to the least square function under the linear regression model is called "Maximum Likelihood". In this method, we have to define a function, called "likelihood function", this function expresses the probability of the observed data as a function of the unknown parameters.

Now consider, Y that is coded as 0 or 1. According to the equation (3.2), the expression for $\pi(x)$, represents the conditional probability of Y taking value 1, given x, which is denoted by $P(Y = 1|x)$. Similarly, the quantity $1-\pi(x)$ indicated the probability of Y is equal to 0, denoted by $P(Y = 0|x)$.

For the pair(x_i, y_i), the contribution to the likelihood function can be defined as following expression.

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \qquad (3.4)$$

Since the observations are independent, then the likelihood function can be defined as;

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \qquad (3.5)$$

The expression for the log likelihood is defined as;

$$L(\beta) = \ln l(\beta) = \sum_{i=1}^{n}\{y_i \ln[\pi(x)] + \ln[1-\pi(x)] - y_i \ln[1-\pi(x)]\} \quad (3.6)$$

$$= \sum_{i=1}^{n} \ln[1-\pi(x_i)] + \sum_{i=1}^{n} y_i ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)$$

$$= \sum_{i=1}^{n} \ln\left[1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right] + \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i)$$

$$= \sum_{i=1}^{n} ln\ \frac{1}{1+e^{\beta_0+\beta_1 x_i}} + \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i)$$

$$= \sum_{i=1}^{n} \ln(1) - \ln(\,1+e^{\beta_0+\beta_1 x_i}) + \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i)$$

$$= -\sum_{i=1}^{n} ln\ \left(1+e^{\beta_0+\beta_1 x_i}\right) + \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i)$$

To find the maximum likelihood estimators we would partially differentiate the log likelihood with respect to the parameters β_0 and β_1. Take the derivatives with respect to β_0;

$$\frac{\partial L(\beta)}{\partial \beta_0} = -\sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} + \sum_{i=1}^{n} y_i$$

$$= \sum_{i=1}^{n} [y_i - \pi(x_i)]$$

Then set the resulting expression equal to zero, we get;

$$\sum_{i=1}^{n}\big(y_i - \pi(x_i)\big) = 0 \quad (3.7)$$

Now take the derivative with respect to β_1;

$$\frac{\partial L(\beta)}{\partial \beta_1} = -\sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} x_i + \sum_{i=1}^{n} y_i x_i$$

$$= x \sum_{i=1}^{n} [y_i - \pi(x_i)]$$

Then set the resulting expression equal to zero, we get;

$$x \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \qquad (3.8)$$

The above equations (3.7) and (3.8) are commonly called as the likelihood equations. The maximum likelihood estimators are given by the solutions for the above likelihood equations (equation (3.7) and(3.8)). In logistic regression, the expressions show in equation (3.7) and (3.8) are non-linear in β_0 and β_1 , thus require numerical methods and these methods are iterative in nature. Therefore, we can use available logistic regression software to find the model coefficients. (David W. Hosmer, 2000)

3.2.6.3 Significance of the Coefficients

After estimating the model coefficients, the significance of the variables in the fitted model is considered. These methods are testing the significance of the statistically hypothesis, to check whether the independent variables in the fitted model significantly related to the response variable. First, we discuss general methods for a simple case: binary logistic regression model. That is model with a single independent variable.

There are two hypothesis testing approaches applied to testing for the significance of the coefficients in the model as below;

- Likelihood Ratio Test
- Wald Test

Likelihood Ratio Test

One approach to testing the significance of the estimated coefficient of a variable in logistic regression model is compare observed values of the response variable to predicted values obtained from models with variable and without the variable. In logistic regression, the log likelihood function can be defined to do this comparison of observed to predicted values. The log likelihood function defined in equation 3.6.

The following expression (3.9) indicated the comparison of observed to predict values using likelihood function;

$$D = -2\ln\left[\frac{(likelihood\ of\ the\ fitted\ model)}{(likelihood\ of\ the\ saturated\ model)}\right] \tag{3.9}$$

An observed value of the response variable as also being a predicted value resulting from a saturated model. The quantity inside the square bracket in the above expression (3.9) is called the **likelihood ratio**. The quantity in the above whole expression can be used for hypothesis testing for significance of the estimated coefficients. Such a test is called the *likelihood ratio test.* (David W. Hosmer, 2000)

Using equation (3.6) and (3.9);

$$D = -2\frac{\Sigma_{i=1}^{n}\{y_i\, ln[\hat{\pi}(x_i)]+(1-y_i)ln[1-\hat{\pi}(x_i)]\}}{\Sigma_{i=1}^{n}\{y_i \ln[y_i]+(1-y_i)\ln[1-y_i]\}}$$

$$D = -2\sum_{i=1}^{n}\left\{y_i \ln\left[\frac{\hat{\pi}(x_i)}{y_i}\right] + (1-y_i)\ln\left[\frac{1-\hat{\pi}(x_i)}{1-y_i}\right]\right\} \tag{3.10}$$

where $\hat{\pi}(x_i)$is the maximum likelihood estimate of $\pi(x_i)$. (estimate of the conditional probability that y is equal to 1, given that x_i).

The statistic, D in equation (3.10) is known as the deviance, which acts for logistic regression same role that the sum of squares (SSE) plays in linear regression of testing for the significance of a fitted model. For purpose of testing the significance of an independent variable (x), compare the value D with and without the independent variable in the equation.

$$G = D(model\ without\ the\ variable) - D(model\ with\ variable)$$

$$G = -2\ ln\ \left[\frac{likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable}\right]$$

The value of G statistic can be simplified as follows;

$$G = 2\left\{\sum_{i=1}^{n}\{y_i\ ln[\hat{\pi}(x_i)] + (1-y_i)\ln[1-\hat{\pi}(x_i)]\} - [n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)]\right\} \tag{3.11}$$

The null hypothesis and alternative hypothesis as follow;

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Under the null hypothesis, β_1 equal to zero, the Test statistic G, follows a chi-square distribution with 1 degree of freedom.

Wald Test

The Wald test is obtained by comparing the maximum likelihood estimate of the slope, $\hat{\beta}_1$, to an estimate of its standard error.

he null hypothesis and alternative hypothesis are follows;

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

The test statistic for the Wald test is;

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

Under the null hypothesis, that means β_1 equal to zero, the Test statistic W, follows a standard normal distribution.

3.2.7 Multiple Logistic Regression

Multiple logistic regression is used to predict the probability of categorical response variable based on multiple independent variables. The independent variables can be either dichotomous (i.e., binary) or continuous (i.e., interval or ratio in scale). Similar to binary logistic regression, the multiple logistic regression also uses maximum likelihood estimation method to evaluate the probability of categorical membership.

When using multiple logistic regression, the following assumptions are required;

- Data should not have multicollinearity.
- Data should not have outliers.
- Have a linear relationship between any continuous independent variables and the logit transformation of the dependent variable.
- Should have independence of observations and the dependent variable should have mutually exclusive and exhaustive categories.

3.2.7.1 The Multiple Logistic Regression Model

Let consider a set of p independent variables, which is denoted by the vector $X' = (x_1, x_2, .., x_p)$ and also assume that we have a sample of n independent observations. Same as in the univariate case, when we are fitting the model, the multivariate case also requires that we obtain estimates of vector $\beta' = (\beta_1, \beta_2, .., \beta_p)$. The method of estimation used in multivariate case will be the same as in the univariate cases, which is maximum likelihood. (David W. Hosmer, 2000)

The conditional probability that the outcome is present (when $Y = 1$) be denoted by;

$$P(Y = 1|x) = \pi(x)$$

where $\pi(x)$ represent the probability of an event that depends on n covariate or independent variables. Then using a logit transformation for modeling the probability, we have;

$$\pi(x) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots\ldots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots\ldots+\beta_px_p}}$$

To obtain corresponding logit function from this, we calculate

$$logit[\pi(X)] = g(x) = ln\left[\frac{\pi(X)}{1-\pi(X)}\right]$$

$$= \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots \ldots . + \beta_pX_p \qquad (3.12)$$

The logit of the probability of an event given x is a simple liner function.

The equation (3.12) shows the logistic regression model, once the dichotomous outcome is transformed by the logit transform. This transform changes the range of $\pi(X)$ from 0 to 1 to $-\infty$ to $+\infty$, as usual for linear regression.

After differentiating the log likelihood function with respect to $(p+1)$ coefficients (including constant), there are $(p+1)$ likelihood equations can be obtained. The maximum likelihood estimators are obtained by maximizing these functions. Thus, the results may be expressed as follows;

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$$

and

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0 \qquad j = 1,2,..,p$$

As a univariate case, to find the likelihood estimators in multivariate case, it requires special software.

3.2.7.2 Fitting the Multiple Logistic Regression Model with Design Variables

Suppose that some independent variables are discrete, nominal scale variable such as gender, treatment group, and educational level etc. If those variables were interval scale variables, which are not appropriate to include them in the model. The numbers can be used to identify the different levels of these nominal scale variables merely identifiers, have no numerical significance. In this kind of situation, use the collection of design variables (or called dummy variables) is the best method. (David W. Hosmer, 2000)

In general, if we assume that the nominal scale variable has "k" possible values, then there will be "k -1" design variables generated. Thus, the logit for a model with p variables and j^{th} variable being discrete would be;

$$g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \sum_{m=1}^{k-1} \beta_{jm} D_{jm} + \ldots . + \beta_p X_p$$

where $k - 1$ design variables are denoted as D_{jm} and the coefficients of these design variables are denoted as β_{jm}, $m = 1,2,..,k-1$.

3.2.7.3 Testing for the Significance of the Model

After we have fitted a multiple logistic regression model, we will check the model assessment. As in the univariate case, the first step in this process is to assess the significant variables in the model.

Likelihood Ratio Test

The likelihood Ratio test for overall significance of the p number of coefficients for the independent variables in the model is carried out in the same manner as in the univariate case. The test is based on the G statistic defined as follows;

$$G = -2\ ln\ \left[\frac{likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable}\right] \tag{3.13}$$

$$G = -2(\ln(likelihood\ without\ the\ variable) - \ln(likelihood\ with\ the\ variable))$$

The only difference is that the fitted values under the model are based on the vector containing $(p + 1)$ parameters. The appropriate null hypothesis and alternative hypothesis are as follows;

$$H_0: The\ slope\ coefficients\ in\ the\ model\ are\ equal\ to\ zero$$
$$H_1: at\ least\ one\ slope\ coefficient\ in\ the\ model\ is\ different\ from\ zero$$

If the p-value for the test is less than the significance level, we can reject the null hypothesis, and conclude that at least one coefficient is different from zero.

Wald Test

The Wald test is obtained by comparing the maximum likelihood estimate of the slope, $\hat{\beta}_j$, to an estimate of its standard error.

The null hypothesis and alternative hypothesis as follow;

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

The test statistic for the Wald test is;

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

If the corresponding p-values for each coefficient are less than the significance level we can reject the null hypothesis, and conclude that the regarding coefficient is different from zero.

Then our main objective in multiple logistic regression model is to obtain the best fitting model while minimizing the number of parameters. To achieve this goal, the next step is to fit a reduced model with only containing variables, which are significant in full model and then compare it with the full model.

3.2.8 Assessing the Fitted Model

After fitting a suitable model to the data, one of the next important steps is to examine how well the fitted model fits the observed data. As in linear regression, assessing the logistic regression model is required to evaluate the quality or suitability of the model. When the model building step has been finished, the number of logical steps can be applied to assess the fitted model. They are evaluation of the overall measures of fits, examination of the individual components of summary statistics and examination of the measures of difference between observed and fits. A Goodness-of-fit test statistic is one of the popular methods to determine the suitability of the fitted logistic model. The one of the main advantages of the Goodness-of-fit statistic is that it provides an easily interpretable single numerical value that can be used to assess the fitted model.

3.2.8.1 Hosmer Lemeshow Test

In order to evaluate overall Goodness-of-fit, Hosmer and Lemeshow introduced grouping estimated method, which according to the values of the estimated probabilities from the logistic regression model. In this test, subjects divided into groups based on predictive probabilities and then computes a chi-square test statistic from observed and expected frequencies. In this approach, the predicted probabilities are arranged as an ascending order and the separated into several groups (generally recommended with ten groups) of approximately equal size. For example, suppose that the n columns of the estimated probabilities. The first column corresponding to the smallest estimated probability value and the nth column to the largest estimated probability value. The grouping strategy defined based on percentiles of the estimated probabilities. When we use g=10 groups, in the first group it contains $n_1 = {}^{n}/_{10}$ subjects, which are having the smallest estimated probabilities. The last group contains $n_{10} = {}^{n}/_{10}$ subjects, which are having the largest estimated probabilities. For y =1 row, the estimates of the expected values are obtained by adding the estimated probabilities in a group from all over subjects. For y =0 row, the estimates of the expected values are obtained by adding the one minus estimated probability in a group from all over subjects. The Hosmer- Lemeshow Goodness-of-fit test statistic, denoted by $\hat{C}$, is obtained by calculating the Pearson chi-square statistic. The estimated expected and observed frequencies are obtained from the $g \times 2$ contingency table. The Hosmer-Lemeshow Goodness-of-fit test statistics is follows;

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

Where g denotes the number of groups, n_k is number of observations in the k^{th} group, O_k is the sum of the Y values for the k^{th} group and $\bar{\pi}_k$ is the average of the ordered $\bar{\pi}_k$ for the k^{th} group.

The null hypothesis of the Hosmer- Lemeshow Goodness-of-fit test is "the fitted logistic model is correct", then the test statistics distributed approximately chi-squared distribution with $g - 2$ degrees of freedom. The large p-value represents that there is no significant difference between observes and estimated expected observation. This indicated that the fitted model is quite reasonable. (Sarkar S.K H. M., 2010)

3.2.8.2 ROC Curve

Classification Table is one of the ways to summarized results of a fitted regression model. This table represents the results of cross-classifying the dichotomous outcome variable and values of the classification table derived from the estimated logistic probabilities. We should define a cut-point c, to obtain the derived dichotomous variable. Then we compare each estimated probabilities and defined cut-point (c). If the estimated probability greater than cut-point, then we derived variable to 1 and otherwise it is equal to 0. Commonly we used 0.5 as a cut-point. The following Table 3.4 represents the common classification table for binary logistic regression model (Sarkar S.K H. M., 2010)

Table 3.4: Classification Table Based on the Logistic Regression Model

Classified	Observed	
	Y = 1	Y = 0
Y = 1	a	c
Y = 0	b	d
Total	a + b	c +d

According to the Table 3.4, the correct classifications are "a" and "d" and also "b" and "c", are the misclassifications. The theoretical background of the terms sensitivity and specificity come from the classification table. Sensitivity is the proportion of true positive or proportion of cases correctly classified by the certain subject (Y = 1). The specificity is the proportion f true negative or the proportion of cases correctly classified by the other condition (Y = 0).

$$Sensitivity = \frac{a}{a+b}$$

$$Specificity = \frac{d}{c+d}$$

A ROC curve is a graphical representation, which plots the probability of true positive (sensitivity) against the probability of false positive (1-specificity) for all positive cutoff points

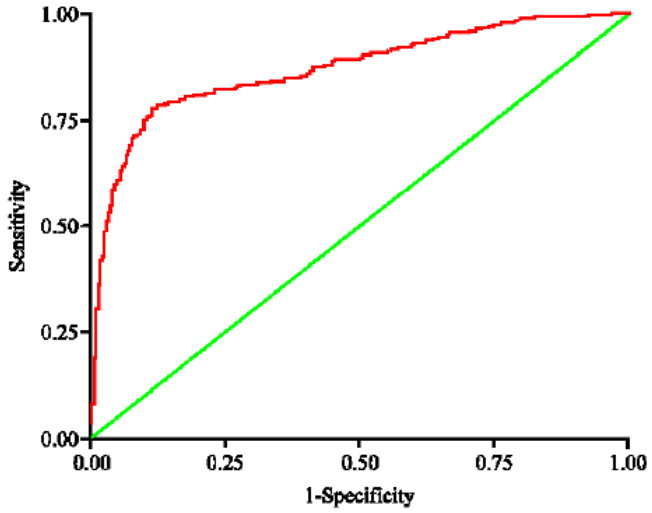


***Figure 3.2*: Receiver Operating Characteristics (ROC) Curve**

The area under a ROC curve (AUC) is a popular measure of the accuracy of a diagnostic test. In general, higher AUC values indicate better test performance. The AUC has an interpretation as follows (David W. Hosmer, 2000).

If AUC = 0.5; this suggests no discrimination

If $0.7 \leq$ AUC < 0.8; this is considered acceptable discrimination.

If $0.8 \leq$ AUC < 0.9; this is considered excellent discrimination.

If AUC ≥ 0.9; this is considered outstanding discrimination.

3.2.9 Interpretation of the Fitted Logistic Regression Model

After fitting a logistic regression model now moves from the computations and valuation of significance of the estimated coefficients to the interpretation of their values. The interpretation is very important, it provides practical inference of the fitted model. However, interpretation of the coefficient of independent variables are very useful for making decisions. It provides the slope or rate of change of a function of dependent variable when change per unit in the independent variable.

3.2.9.1 Interpretation of Odds Ratio in the presence of Categorical Dichotomous Independent Variable

Suppose that we have only two categories in the independent variable. This kind of independent variables are called "Dichotomous independent variables". The function of response (dependent) variable is a linear function of the predictor (independent) variables in a model, which is called link function. In the logistic regression model this link function is the logit transformation

$$g(x) = \beta_0 + \beta_1 x$$

The estimated coefficient for the independent variable represents the slope or rate of change. In logistic regression model, the slope coefficient (β_1), provides the expected change in the logit corresponding to a change of one unit in the independent variable.

$$\beta_1 = g(x+1) - g(x)$$

Let us consider independent variable x, which is coded as either zero or one. Then the logit for a subject with $x = 1$ denoted as $g(1)$ and the logit for a subject with $x = 0$ denoted as $g(0)$. Thus the difference in the logit for a subject with $x = 1$ and $x = 0$ is as follows;

$$g(1) - g(0) = (\beta_0 + \beta_1) - (\beta_0) = \beta_1$$

In this case logit difference is equal to β_1, or rate of change in the independent variable (David W. Hosmer, 2000). For further interpretations, we need to discuss have a

proper idea about the odds ratio. The next section we will discuss the interpretation of the odds ratio.

The possible values of the logistic probabilities for the dichotomous independent variable, are displayed in the 2×2 contingency table in Table 3.5.

Table 3.5: Logistic Probabilities for the Dichotomous Independent Variable

Dependent Variable (Y)	Independent Variable (X)	
	$X = 1$	$X = 0$
$y = 1$	$p(y = 1\|x = 1)$ $= \pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$p(y = 1\|x = 0)$ $= \pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y = 0$	$p(y = 0\|x = 1)$ $= 1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$(y = 0\|x = 0)$ $= 1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$

The odds of the outcome being present among individuals with $x = 1$ is defined as follows;

$$\frac{\pi(1)}{1-\pi(1)}$$

The odds of the outcome being present among individuals with $x = 0$ is defined as follows;

$$\frac{\pi(0)}{1-\pi(0)}$$

The odds ratio, denoted as OR, is defined as the ratio of the odds for $x = 1$ to the odds for $x = 0$. The OR defined as follows;

$$R = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \tag{3.14}$$

$$= \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \Big/ \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} \Big/ \frac{1}{1+e^{\beta_0}}}$$

$$= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}$$

$$= e^{\beta_1} \tag{3.15}$$

$$OR = \quad e^{\beta_1}$$

For logistic regression model, when the dichotomous independent variable coded as 0 and 1, then there is a close relationship between the odds ratio and the regression coefficient as follows;

$$\ln(OR) = \ln(e^{\beta_1})$$

$$\ln(OR) \ = \beta_1 \tag{3.16}$$

The above equation 3.16 provides very important relationship between the model coefficient and log odds ratio. It shows that the model coefficient is equal to the log odds ratio. This is very powerful research tool for interpretation in practical scenario. β_1 represents the change in the logit corresponding to a change of one unit in the independent variable.

$$OR = \quad e^{\beta_1}$$

3.2.9.2 Interpretation of Odds Ratio when Categorical Polychotomous Independent Variable

Suppose that we have more than two categories in the independent variables instead of two categories. These types of independent variables we called "polychotomous Independent variables". For example, race, school types, educational level, etc. There are more than two categories in each of the above examples. Therefore, we need to define set of design variables to represent each levels (categories) of the variables. In this section we are going to explain the method for creating design variables to represent the categories of the variable for polychotomous independent variables.

We assume that there is a polychotomous independent variable with four levels. Therefore, we want to create the design variables necessary to include the variables in the logistic regression. Since the independent variable has four categories, three design variables must be created on the goal of the analysis and model development. The corresponding coding system as the design variables for the polychotomous independent variable showed as following table 3.6.

Table 3.6: Coding of the Design variables for polychotomous independent variable using Reference Cell Coding with Level 1 as the reference group

Variable	Variable_2	Variable_3	Variable_4
Level 1 (1)	0	0	0
Level 2 (2)	1	0	0
Level 3 (3)	0	1	0
Level 4 (4)	0	0	1

Table 3.7 represents the odds ratios and log odds ratios for each level in polychotomous independent variable. At the bottom of the Table 3.7, shows the odds ratio for each variable, noted that the Level 1 as the reference group. The reference group is indicated by a value of

1 for the odds ratio. At the last row of the same table, shows that log odds ratios for each Level, using Level 1 as the reference group.

Table 3.7: Specification of the Design variables for polychotomous independent variable using Reference Cell Coding with Level 1 as the reference group

Program	Level 1	Level 2	Level 3	Level 4
Academic (1) General (0)	a e	b f	c g	d h
Estimated Odds Ratio	1	$\frac{e \times b}{a \times f}$	$\frac{e \times c}{a \times g}$	$\frac{e \times d}{a \times h}$
Estimated ln(OR)	0	$\ln\left(\frac{e \times b}{a \times f}\right)$	$\ln\left(\frac{e \times c}{a \times g}\right)$	$\ln\left(\frac{e \times d}{a \times h}\right)$

3.2.9.3 Interpretation Odds Ratio when Continuous Independent Variables

Logistic regression model may contain both continuous and categorical independent variables. If continuous independent variables are included in a logistic regression model, to interpret the model coefficients, then we will assume that the logit is linear in the continuous variable. Based on the assumption that the logit is linear in the continuous variable x, the equation for the logit is can express as $g(x) = \beta_0 + \beta_1 x$. The interpretation of this slope coefficient, β_1 is, it gives the rate of change in the log odds for an increase of one unit in x, that is $\beta_1 = g(x+1) - g(x)$ for any value of x (David W. Hosmer, 2000).

CHAPTER 4

RESULTS AND DISCUSSION

At the first step, the descriptive analysis was done for the collected data. Then the chi-square test of independence and multiple logistic regression models were applied to analyze household's milk consumption in Matara district. Chi-square test of independence was used to investigate the effects of socioeconomic characteristics on consumers' local and imported milk consumption behavior. The multiple logistic regression model was used to determine the extent, how selected socioeconomic characteristics of consumers influence these milk types. The RStudio Statistical package was used to analyze the data.

4.1 Descriptive Data Analysis

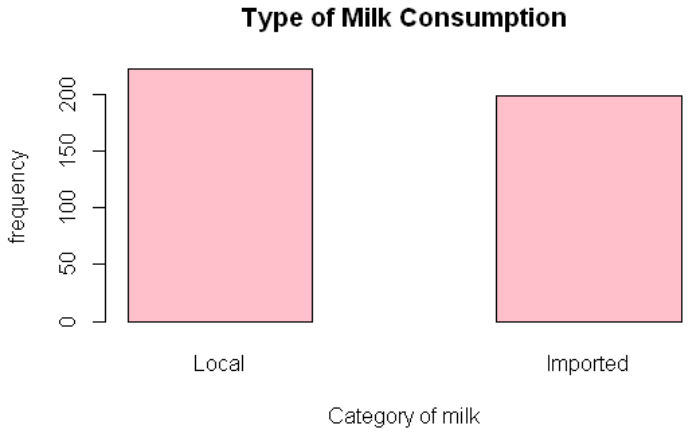


Figure 4.1: **Bar Plot for Type of Milk Consumption**

According to Figure 4.1, the number of consumers interested in local milk is higher than imported milk. To identify the behavior of predictor variables with response variable (milk type), each predictor variable was examined with the type of milk consumption.

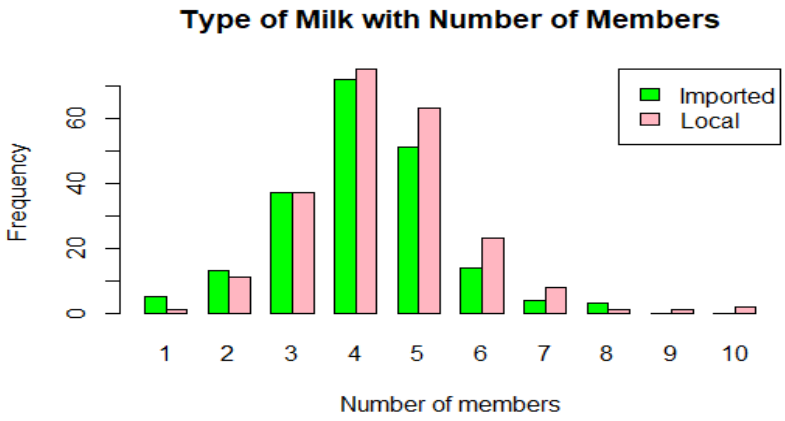


***Figure 4.2*: Bar plot for Number of Family Members with Type of Milk**

We expected that household size influence for household's milk choices. According to Figure 4.2, it illustrated that large families are more preferred to local milk than imported milk. The families with less than three members, they are more likely to imported milk than local milk.

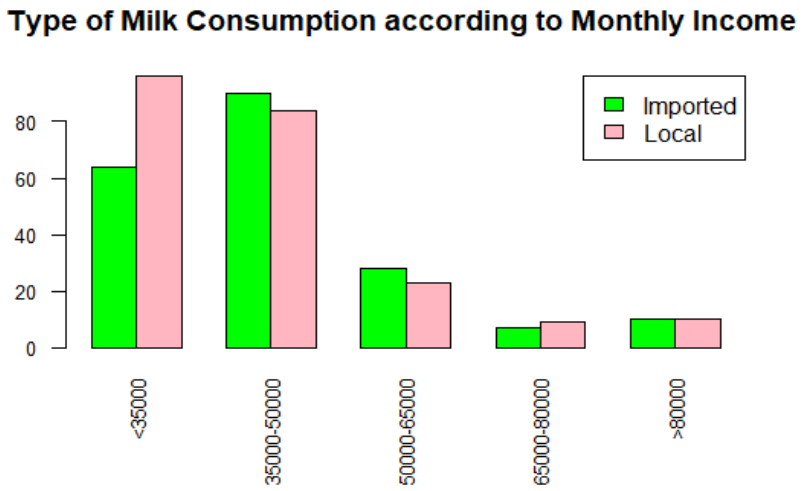


***Figure 4.3*: Bar plot for Monthly Income with Type of Milk**

Figure 4.3 shows the distribution of milk consumption with monthly income. According to Figure 4.3, the highest difference between local and imported milk consumption among lower-income rate and also number local milk consumers are greater than that of the

imported milk consumers. The lower income consumers are more likely to prefer local milk. Figure 4.4 shows that the effect of educational level for their milk choices.

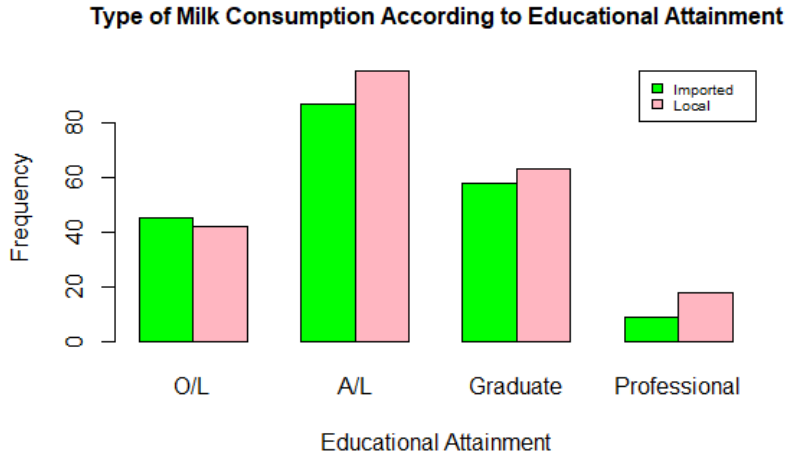


***Figure 4.4*: Bar plot for Type of Milk Consumption according to Education Level**

Figure 4.4 indicated that the consumers who have passed the GCE (A/L) examination and above tend to used local milk. Therefore, we can conclude that the better-educated consumer has a higher preference for local milk than imported milk.

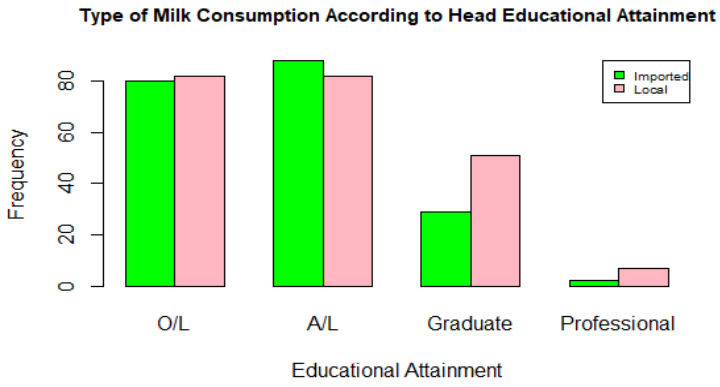
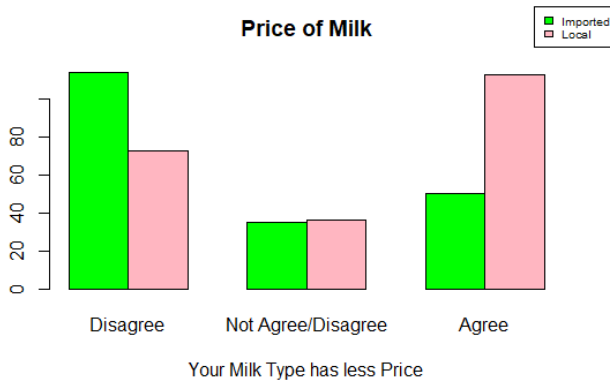
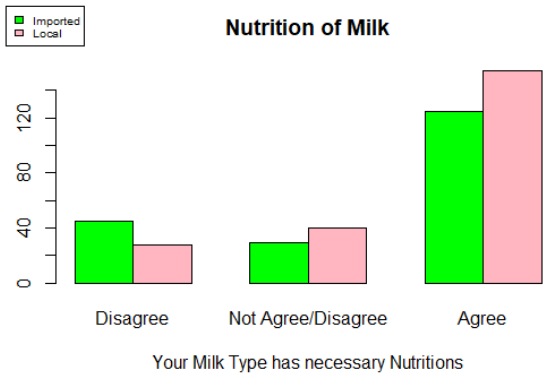


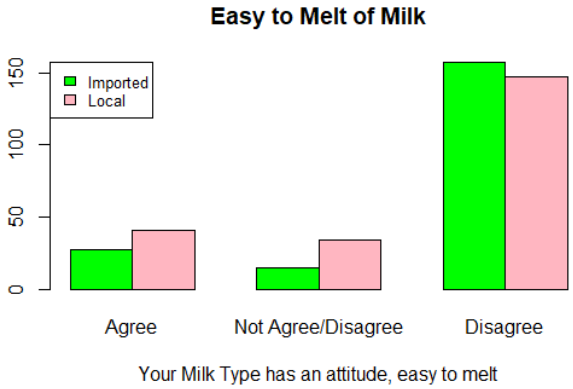
***Figure 4.5*: Bar plot for Education Level of Household Head with Type of Milk Consumption**

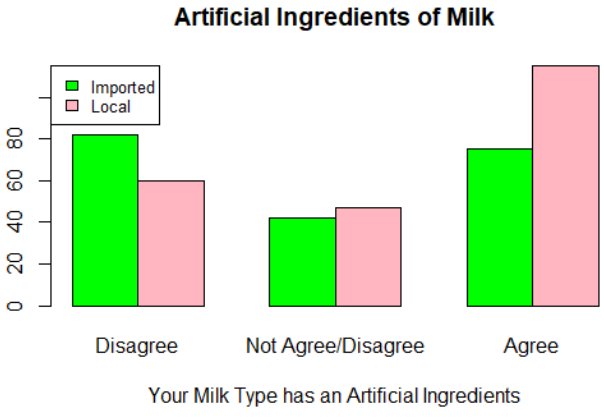
As we expect, education of the household head is influenced by households' milk choices. Figure 4.5 indicated that the higher education levels of household heads (Graduate and Professional) have a higher preference for local milk than imported milk.

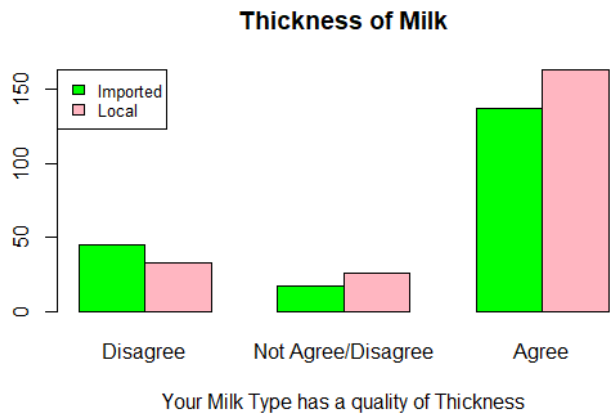
In this study, there were 12 factors we have considered to observe, consumer's opinion about local and imported milk. Under each factor, there are three categories have defined to identify prefer for their type of milk consumption. They are agree, neither agree nor disagree, disagree.

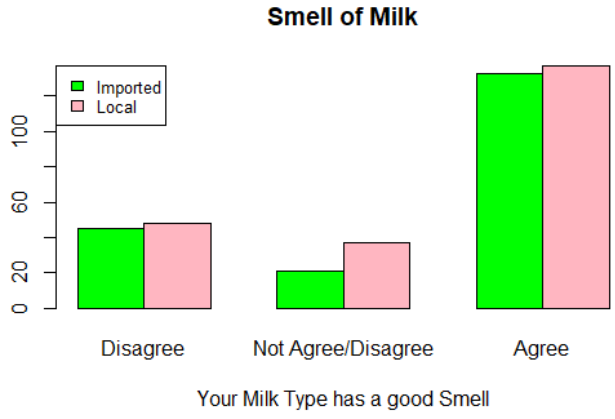


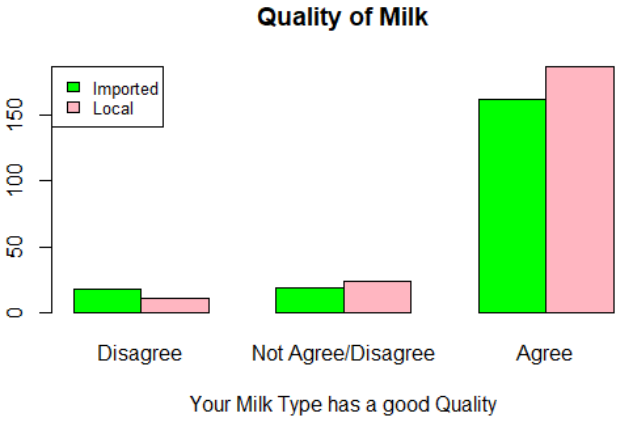


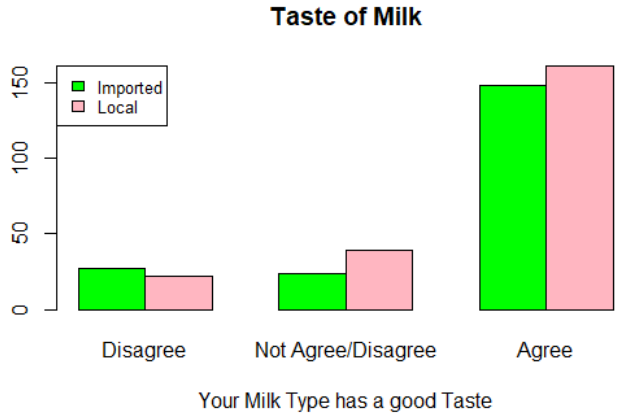


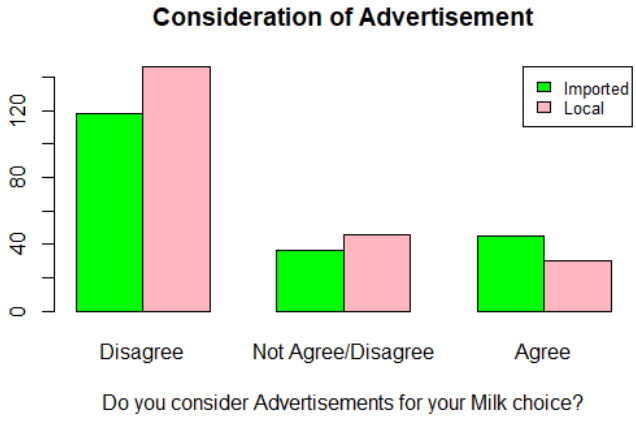


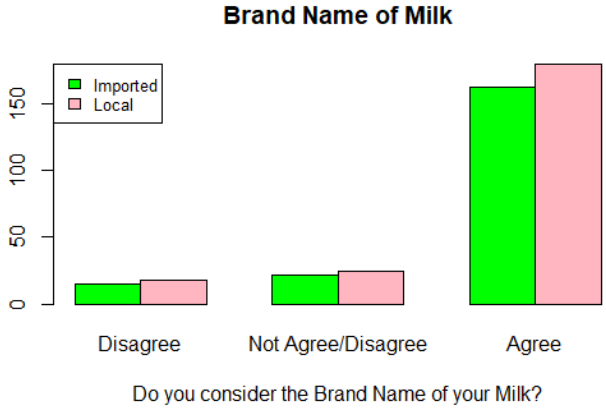

Thickness of Milk
Imported
Local
0
50
100
150
Disagree
Not Agree/Disagree
Agree
Your Milk Type has a quality of Thickness


Smell of Milk
Imported
Local
0
20
60
100
Disagree
Not Agree/Disagree
Agree
Your Milk Type has a good Smell


Quality of Milk
Imported
Local
0
50
100
150
Disagree
Not Agree/Disagree
Agree
Your Milk Type has a good Quality


Taste of Milk
Imported
Local
0
50
100
150
Disagree
Not Agree/Disagree
Agree
Your Milk Type has a good Taste


Consideration of Advertisement
Imported
Local
0
40
80
120
Disagree
Not Agree/Disagree
Agree
Do you consider Advertisements for your Milk choice?


Brand Name of Milk
Imported
Local
0
50
100
150
Disagree
Not Agree/Disagree
Agree
Do you consider the Brand Name of your Milk?

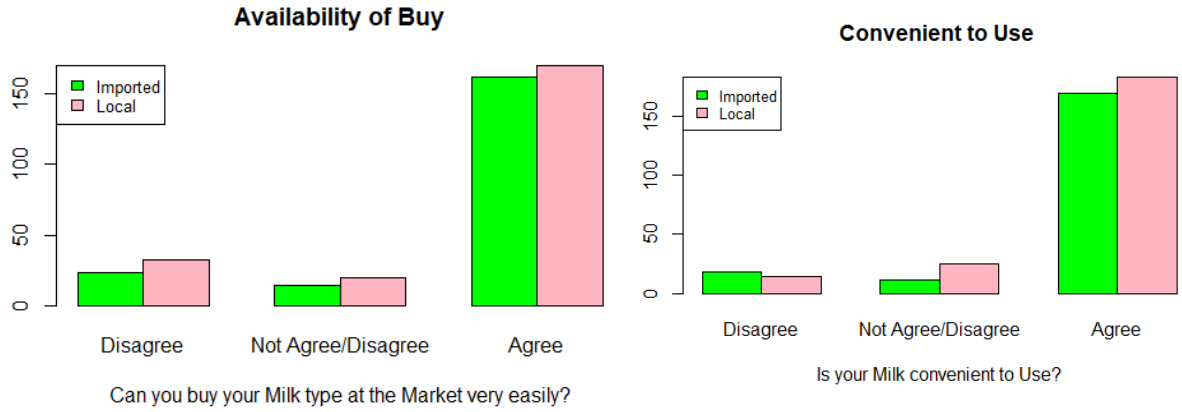


***Figure 4.6*: Bar plots for Type of Milk Consumption according to Consumer's Opinion about their selected milk type**

Based on the above plots (figure 4.6), there is a reasonable gap between local milk consumers and the imported milk consumers based on the price of the milk. Therefore, the consumers' are considering the price of the milk for their milk choices. Most consumers are considering about the nutrition, thickness and smell on their milk choice. The bar plot (figure 4.6) shows that there is no difference between the smell of the local and imported milk, and additionally the imported milk is easier to melt than the local milk.

Most of the people are considering the factors, which are artificial ingredients, quality, taste, brand name, availability of buy their milk and convenient to use on their milk choice. The advertisement is an important point in this analysis. The majority of consumers do not consider an advertisement to their choices, among them, the local milk consumers are higher than imported milk consumers. In contrast, the consumers who considering an advertisement for their milk choices, the imported milk consumers are higher than local milk consumers.

4.2 Univariate Analysis

In this section, cross-tabulations were applied for the type of milk consumption and considering factors as mention above. The main objective in this section is to test whether significance association between type of milk consumption (local or imported) and socioeconomic factors.

Table 4.1: Results of Chi-Squared Test of Independence for Milk Type and Selected factors.

Data	χ^2 Test statistic	p-value
Milk Type and Education	2.8362	0.4176
Milk Type and Education of the household Head	7.8311	0.04963
Milk Type and Monthly Income	6.1088	0.1912
Milk type and Quality of milk	2.7815	0.2449
Milk type and Price of milk	31.666	1.33e-07
Milk type and Taste of milk	4.4921	0.1802
Milk type and Nutrition of milk	7.4006	0.0258
Milk type and Thickness	4.6806	0.0963
Milk type and Easy to melt	9.0018	0.0111
Milk type and Smell of milk	3.401	0.1826
Milk type and Easy to buy	1.5314	0.465
Milk type and Brand Name	0.0649	0.968
Milk type and Easy to Use	5.2155	0.0737
Milk type and artificial ingredient	10.614	0.0049
Milk type and Advertisements	5.8551	0.0535

Table 4.1 indicated that the results of Pearson's Chi-Square Test, to identify the association between each factor and type of milk consumption. Here we checked the statistical significance at 0.05 level. Table 4.1 represents the Chi-Squared test statistic and relevant probability value (*p*-value) for each case.

According to Table 4.1, education of the head of household is one of the important factors which significantly associated with choice of the milk type (*p*- value= 0.049). As mentioned above, figure 4.5 also shows the distribution of household heads education level with their milk consumption. As we expect, education of the household head is influenced by households' milk choices. Figure 4.5 also indicated that the higher education levels of household heads (Graduate and Professional) have a higher preference for local milk than less educated heads.

In the second part of the questionnaire, the consumers provided their opinion about some factors based on their milk choices. According to Table 4.1, there is a significant association between milk price and type of milk consumption. According to figure 4.6, the consumer who tend to buy imported milk, they disagree about the statement "the price of the type of your milk is lower than other milk types". And also, the consumer who tend to buy local milk, they agree with the statement which we mentioned above. Therefore, we can conclude that the prices of imported milk are reasonably higher than the local milk.

The variables, Nutrition of milk (p-value = 0.025), Easy to melt (p-value = 0.011) and artificial ingredients (p-value = 0.0049), which are also significantly associated with type of milk consumption.

According to the statement "It contains necessary nutrition for the human body", the Figure 4.6 illustrated that, both local milk and imported milk buyer's opinion about nutrition based on their milk choices. Considering Figure 4.6, the higher rate of consumers agrees to the above statement and, besides, among them, local milk consumers are higher than imported milk consumers. Therefore, we can conclude that the majority of consumers believe that, local milk is more essentially helps to a healthy life than imported milk.

Based on the statement “Easy to melt”, Figure 4.6 shows the bar chart for consumer’s opinion for the above statement. When we are comparing the local milk consumers with imported milk consumers who disagreed with the above statement, the local milk consumers are higher than the imported milk consumers. In contrast, the number of imported milk consumers are highly agreed with the above statement that the local milk consumers. Therefore, we can conclude that imported milk is more melt than the local milk.

Figure 4.6 illustrated that the answers for the question “Are you consider an artificial ingredient, before your milk choice?”. Figure 4.6 indicated that the higher rate of consumers is considered about the artificial ingredients before their milk choice, among them local milk consumers are considering about artificial ingredients with their milk choices than imported milk consumers. In contrast with, the few rates of local milk consumers are do not considering about artificial ingredients of milk. Therefore, we can conclude that the majority of consumers believed that the local milk does not contain artificial ingredients rather than imported milk.

4.3 Fitting a Logistic Regression Model

The results of the binary logistic regression model for household milk consumption with socioeconomics and other identified factors are presented in this section. The model coefficients have been estimated by the maximum likelihood method. Series of design variables have been defined for all the categorical variables when fitting a logistic regression model. The last category in each categorical variable defined as a reference category.

The result of binary logistic regression model with all predictor variables given in Table 4.2. In this table first and second columns are represented, variable and corresponding odds ratios respectively. The other columns represent model coefficients, 95% confidence Interval, standard error of each coefficient, Wald test statistic and probability value respectively.

Table 4.2: Binary Logistic Regression Model with all predictor variables

Variables	Odds Ratio	Estimate	CI for estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.59	0.46494	(-1.74 , 2.67)	1.12537	0.413	0.6795
Age	1.028	0.028	(0.004 , 0.051)	0.011842	2.365	0.018
EducationDV3	0.617	-0.481	(-1.644, 0.682)	0.593617	-0.81	0.417
EducationDV2	0.626	-0.467	(-1.641, 0.706)	0.599	-0.78	0.417
EducationDV1	0.26	-1.344	(-2.675, -0.014)	0.678	-1.981	0.047
H_EducationDV3	059	-0.514	(-2.476, 1.446)	1	-0.514	0.607
H_EducationDV2	0.279	-1.273	(-3.21, 0.670)	0.992	-1.284	0.199
H_EducationDV1	0.336	-1.089	(-3.057, 0.878)	1.004	-1.085	0.277
Monthly_IncomeDV4	2.170	0.775	(-0.763, 2.313)	0.784	0.987	0.323
Monthly_IncomeDV3	1.098	0.094	(-1.124, 1.312)	0.621	0.151	0.879
Monthly_IncomeDV2	1.653	0.502	(-0.617, 1.622)	0.571	0.879	0.379
Monthly_IncomeDV1	3.11	1.136	(-0.017, 2.290)	0.588	1.93	0.053
QualityDV2	0.809	-0.211	(-1.062, 0.638)	0.433	-0.488	0.625
QualityDV1	0.475	-0.743	(-1.728, 0.242)	0.502	-1.478	0.139
PriceDV2	0.351	-1.044	(-1.735, -0.352)	0.352	-2.96	0.003
PriceDV1	0.278	-1.277	(-1.801, -0.753)	0.267	-4.777	0.001
TasteDV2	1.249	0.222	(-0.627, 1.158)	0.428	0.519	0.603
TasteDV1	1.303	0.265	(-1.158, 0.627)	0.455	0.582	0.56
NutritionDV2	1.011	0.011	(-0.730, 0.752)	0.378	0.029	0.976
NutritionDV1	0.886	-0.12	(-0.897, 0.656)	0.396	-0.305	0.76
ThicknessDV2	0.425	-0.854	(-1.771 ,0.063)	0.467	-1.825	0.067
ThicknessDV1	0.448	-0.801	(-1.573, -0.028)	0.394	-2.032	0.042
Easy_meltDV2	3.108	1.134	(1.190, 0.190)	0.481	2.355	0.018
Easy_meltDV1	2.061	0.723	(-1.422, 2.078)	0.356	2.029	0.042
SmellDV2	1.520	0.418	(-0.443, 1.281)	0.44	0.952	0.341
SmellDV1	1.009	0.009	(-0.678, 0.697)	0.351	0.028	0.977
Easy_buyDV2	0.995	-0.004	(-1.013, 1.005)	0.515	-0.008	0.993
Easy_buyDV1	1.580	0.457	(-0.374, 1.289)	0.424	1.078	0.281
Brand_nameDV2	0.653	-0.425	(-1.316, 0.465)	0.454	-0.935	0.349
Brand_nameDV1	0.996	-0.003	(-1077, 1.069)	0.547	-0.007	0.994
Easy_useDV2	2.092	0.738	(-0.216, 1.693)	0.487	1.151	0.129
Easy_useDV1	0.810	-0.21	(-1.261, 0.841)	0.536	-0.392	0.695
Arf_ingredientDV2	0.594	-0.52	(-1.174, 0.134)	0.334	-1.557	0.119
Arf_ingredientDV1	0.583	-0.538	(-1.071, -0.006)	0.271	1.984	0.047
other_influenceDV2	1.542	0.433	(-0.403, 1.270)	0.426	1.016	0.309
other_influenceDV1	0.903	-0.101	(-0.654, 0.450)	0.281	-0.36	0.718
AdvertisementDV2	1.074	0.071	(-0.864, 1.007)	0.477	0.15	0.88
AdvertisementDV1	1.816	0.596	(-0.045, 1.238)	0.327	1.823	0.068

According to the Table 4.2, the following variables are statistically significance at 0.05 level. (In the above outputs, the design variables are defined as DV1, DV2. etc.). Age of the respondent, education level, monthly income, price of the milk, thickness, attitude of easy melt and considering artificial ingredients. The AIC value for the Fullmodel is indicated that 561.84.

At the second step, the most important variables are selected using the backward elimination method. The results of backward elimination method given as follows;

Results of Backward Elimination Method

```
Start:  AIC=561.84
milktype ~ Age + EducationDV + H_EducationDV + Monthly_IncomeDV +
    QualityDV + PriceDV + TasteDV + NutritionDV + ThicknessDV +
    Easy_meltDV + SmellDV + Easy_buyDV + Brand_nameDV + Easy_useDV +
    Arf_ingredientDV + other_influenceDV + AdvertisementDV

                     Df Deviance    AIC
- NutritionDV         2   485.94 557.94
- TasteDV             2   486.30 558.30
- Brand_nameDV        2   486.75 558.75
- SmellDV             2   486.84 558.84
- Easy_buyDV          2   487.06 559.06
- other_influenceDV   2   487.53 559.53
- QualityDV           2   488.15 560.15
- Easy_useDV          2   488.64 560.64
- H_EducationDV       3   491.09 561.09
<none>                    485.84 561.84
- AdvertisementDV     2   489.94 561.94
- Arf_ingredientDV    2   490.70 562.70
- EducationDV         3   492.83 562.83
- ThicknessDV         2   492.09 564.09
- Monthly_IncomeDV    4   496.19 564.19
- Age                 1   491.57 565.57
- Easy_meltDV         2   494.34 566.34
- PriceDV             2   511.23 583.23

Step:  AIC=557.94
milktype ~ Age + EducationDV + H_EducationDV + Monthly_IncomeDV +
    QualityDV + PriceDV + TasteDV + ThicknessDV + Easy_meltDV +
    SmellDV + Easy_buyDV + Brand_nameDV + Easy_useDV + Arf_ingredientDV +
    other_influenceDV + AdvertisementDV

                     Df Deviance    AIC
- TasteDV             2   486.35 554.35
- Brand_nameDV        2   486.86 554.86
- SmellDV             2   486.98 554.98
- Easy_buyDV          2   487.14 555.14
- other_influenceDV   2   487.69 555.69
- QualityDV           2   488.40 556.40
- Easy_useDV          2   488.80 556.80
```

```
- H_EducationDV      3   491.20 557.20
<none>                   485.94 557.94
- AdvertisementDV    2   490.21 558.21
- Arf_ingredientDV   2   490.89 558.89
- EducationDV        3   493.01 559.01
- Monthly_IncomeDV   4   496.27 560.27
- ThicknessDV        2   493.14 561.14
- Age                1   491.60 561.60
- Easy_meltDV        2   494.63 562.63
- PriceDV            2   513.02 581.02

Step:  AIC=554.35
milktype ~ Age + EducationDV + H_EducationDV + Monthly_IncomeDV +
    QualityDV + PriceDV + ThicknessDV + Easy_meltDV + SmellDV +
    Easy_buyDV + Brand_nameDV + Easy_useDV + Arf_ingredientDV +
    other_influenceDV + AdvertisementDV

                     Df Deviance    AIC
- Brand_nameDV        2   487.27 551.27
- Easy_buyDV          2   487.48 551.48
- SmellDV             2   488.07 552.07
- other_influenceDV   2   488.17 552.17
- QualityDV           2   488.61 552.61
- Easy_useDV          2   489.37 553.37
- H_EducationDV       3   491.55 553.55
<none>                   486.35 554.35
- AdvertisementDV     2   490.91 554.91
- Arf_ingredientDV    2   491.33 555.33
- EducationDV         3   493.46 555.46
- Monthly_IncomeDV    4   496.57 556.57
- ThicknessDV         2   493.15 557.15
- Age                 1   492.52 558.52
- Easy_meltDV         2   495.00 559.00
- PriceDV             2   513.17 577.17

Step:  AIC=551.27
milktype ~ Age + EducationDV + H_EducationDV + Monthly_IncomeDV +
    QualityDV + PriceDV + ThicknessDV + Easy_meltDV + SmellDV +
    Easy_buyDV + Easy_useDV + Arf_ingredientDV + other_influenceDV +
    AdvertisementDV

                     Df Deviance    AIC
- SmellDV             2   488.61 548.61
- Easy_buyDV          2   488.96 548.96
- other_influenceDV   2   488.99 548.99
- QualityDV           2   489.65 549.65
- Easy_useDV          2   490.10 550.10
- H_EducationDV       3   493.18 551.18
<none>                   487.27 551.27
- AdvertisementDV     2   492.05 552.05
- Arf_ingredientDV    2   492.05 552.05
- EducationDV         3   494.86 552.86
- Monthly_IncomeDV    4   497.34 553.34
- ThicknessDV         2   494.36 554.36
- Age                 1   493.14 555.14
- Easy_meltDV         2   496.01 556.01
- PriceDV             2   514.61 574.61
```

```
Step:  AIC=548.61
milktype ~ Age + EducationDV + H_EducationDV + Monthly_IncomeDV +
    QualityDV + PriceDV + ThicknessDV + Easy_meltDV + Easy_buyDV +
    Easy_useDV + Arf_ingredientDV + other_influenceDV + AdvertisementDV

                     Df Deviance    AIC
- Easy_buyDV          2   490.34 546.34
- other_influenceDV   2   490.51 546.51
- QualityDV           2   490.95 546.95
- Easy_useDV          2   491.41 547.41
- H_EducationDV       3   494.56 548.56
<none>                    488.61 548.61
- AdvertisementDV     2   493.12 549.12
- Arf_ingredientDV    2   493.17 549.17
- EducationDV         3   496.73 550.73
- Monthly_IncomeDV    4   499.09 551.09
- ThicknessDV         2   496.09 552.09
- Age                 1   494.60 552.60
- Easy_meltDV         2   498.79 554.79
- PriceDV             2   515.22 571.22

Step:  AIC=546.34
milktype ~ Age + EducationDV + H_EducationDV + Monthly_IncomeDV +
    QualityDV + PriceDV + ThicknessDV + Easy_meltDV + Easy_useDV +
    Arf_ingredientDV + other_influenceDV + AdvertisementDV

                     Df Deviance    AIC
- other_influenceDV   2   492.24 544.24
- QualityDV           2   492.25 544.25
- Easy_useDV          2   492.85 544.85
- H_EducationDV       3   496.23 546.23
<none>                    490.34 546.34
- Arf_ingredientDV    2   495.32 547.32
- AdvertisementDV     2   495.58 547.58
- EducationDV         3   498.57 548.57
- Monthly_IncomeDV    4   500.80 548.80
- ThicknessDV         2   497.45 549.45
- Age                 1   496.60 550.60
- Easy_meltDV         2   501.35 553.35
- PriceDV             2   516.68 568.68

Step:  AIC=544.24
milktype ~ Age + EducationDV + H_EducationDV + Monthly_IncomeDV +
    QualityDV + PriceDV + ThicknessDV + Easy_meltDV + Easy_useDV +
    Arf_ingredientDV + AdvertisementDV

                    Df Deviance    AIC
- QualityDV          2   494.05 542.05
- Easy_useDV         2   495.08 543.08
- H_EducationDV      3   497.80 543.80
<none>                   492.24 544.24
- AdvertisementDV    2   496.47 544.47
- Arf_ingredientDV   2   498.15 546.15
- EducationDV        3   500.36 546.36
- ThicknessDV        2   498.96 546.96
- Monthly_IncomeDV   4   503.16 547.16
- Age                1   498.61 548.61
- Easy_meltDV        2   502.82 550.82
- PriceDV            2   518.92 566.92
```

```
Step:  AIC=542.05
milktype ~ Age + EducationDV + H_EducationDV + Monthly_IncomeDV +
    PriceDV + ThicknessDV + Easy_meltDV + Easy_useDV + Arf_ingredientDV +
    AdvertisementDV

                   Df Deviance    AIC
- Easy_useDV        2   496.70 540.70
- H_EducationDV     3   498.73 540.73
<none>                  494.05 542.05
- AdvertisementDV   2   498.46 542.46
- EducationDV       3   501.79 543.79
- Monthly_IncomeDV  4   504.84 544.84
- Arf_ingredientDV  2   501.09 545.09
- ThicknessDV       2   501.34 545.34
- Age               1   499.62 545.62
- Easy_meltDV       2   504.93 548.93
- PriceDV           2   521.43 565.43

Step:  AIC=540.7
milktype ~ Age + EducationDV + H_EducationDV + Monthly_IncomeDV +
    PriceDV + ThicknessDV + Easy_meltDV + Arf_ingredientDV +
    AdvertisementDV

                   Df Deviance    AIC
- H_EducationDV     3   501.08 539.08
<none>                  496.70 540.70
- AdvertisementDV   2   501.06 541.06
- EducationDV       3   504.14 542.14
- Monthly_IncomeDV  4   506.64 542.64
- Arf_ingredientDV  2   503.73 543.73
- ThicknessDV       2   503.83 543.83
- Age               1   502.34 544.34
- Easy_meltDV       2   510.90 550.90
- PriceDV           2   524.20 564.20

Step:  AIC=539.08
milktype ~ Age + EducationDV + Monthly_IncomeDV + PriceDV + ThicknessDV +
    Easy_meltDV + Arf_ingredientDV + AdvertisementDV

                   Df Deviance    AIC
<none>                  501.08 539.08
- AdvertisementDV   2   505.98 539.98
- Monthly_IncomeDV  4   510.70 540.70
- ThicknessDV       2   507.29 541.29
- Arf_ingredientDV  2   508.61 542.61
- Age               1   507.48 543.48
- EducationDV       3   513.71 545.71
- Easy_meltDV       2   514.80 548.80
- PriceDV           2   531.78 565.78
```

Backward selection algorithm starts with a complex model with all the variables, and its AIC value is 561.84. Then sequentially remove one by one variables at each step. At the first step it removes, "`NutritionDV`" and at the same step the AIC value is 557.94. In this procedure, detection of variables from a full model is done based on the importance of the variable.

Finally, it has removed 9 variables within nine steps. There are eight variables remain in the final step, which are AdvertisementDV, `Monthly_IncomeDV`, `ThicknessDV`, `Arf_ingredientDV`, `Age`, `EducationDV`, `Easy_meltDV` and `PriceDV`. The AIC value for the final step is 539.08. The summary of the binary logistic regression model, with variables which are selected from backward elimination method as following Table 4.3.

Table 4.3: Summary Table for the Model with Backward Elimination Method

Variables	**Coefficients**	**Std. Error**	**Z value**	**Pr (> |z|)**
(Intercept)	0.01401	0.80508	0.017	0.98612
Age	0.02742	0.01097	2.501	0.01240
EducationDV3	-0.73910	0.49460	-1.494	0.13509
EducationDV2	-1.00456	0.55804	-3.149	0.00164
EducationDV1	-1.75749	0.49118	-2.045	0.04084
Monthly_IncomeDV4	0.73082	0.74375	0.983	0.32579
Monthly_IncomeDV3	0.01730	0.58659	0.029	0.97647
Monthly_IncomeDV2	0.31875	0.53109	0.600	0.54839
Monthly_IncomeDV1	0.74767	0.52836	1.415	0.05070
PriceDV2	-1.03644	0.33200	-3.122	0.00180
PriceDV1	-1.30683	0.24848	-5.259	1.45e-07
ThicknessDV2	-0.51594	0.42363	-1.218	0.22327
ThicknessDV1	-0.73570	0.31732	-2.318	0.02042
Easy_meltDV2	1.32184	0.42856	3.084	0.00204
Easy_meltDV1	0.75554	0.32062	2.357	0.01845
Arf_ingredientDV2	-0.42110	0.31085	-1.355	0.17553
Arf_ingredientDV1	-0.68031	0.25262	-2.693	0.00708
AdvertisementDV2	0.47677	0.38117	1.251	0.21100
AdvertisementDV1	0.65688	0.30010	2.189	0.02861

Table 4.3 shows the number of variables selected by backward elimination method. The mo del with all covariates, defined as “Fullmodel” and the model with backward elimination m

ethod, defined as “Reducedmodel”. Then, Likelihood ratio test and Akaike Information Cri teria (AIC) have considered to select best model among full model and backward model.

4.4 Model Selection Criteria

Likelihood Ratio Test

Output 4.2 shows the output for likelihood ratio test for “Fullmodel” and “Reducedmodel ”. The “Fullmodel” contains all the covariates in this analysis. The difference between the two models is the exclusion of H_Education, Quality, Taste, Nutrition,smell, Easy_buy, Brand name, Easy_use and Other_influence from the full model.

Results of Likelihood Ratio Test (Fullmodel and Reducedmodel)

```
Likelihood ratio test

Fullmodel: milk_type ~ Age + EducationDV + H_EducationDV + Monthly_Incom
eDV + QualityDV + PriceDV + TasteDV + NutritionDV + ThicknessDV +
    Easy_meltDV + SmellDV + Easy_buyDV + Brand_nameDV + Easy_useDV +
    Arf_ingredientDV + other_influenceDV + AdvertisementDV

Reducedmodel: milk_type ~ Age + EducationDV + Monthly_IncomeDV + PriceDV
       +ThicknessDV + Easy_meltDV + Arf_ingredientDV + AdvertisementDV

  #Df   LogLik   Df    Chisq    Pr(>Chisq)
1  38  -242.92
2  19  -250.54  -19    15.243    0.7071
```

Based on the results from the Likekihood Ratio test, the p-value is 0.701, which greater than 0.05 significance level. Therefor we can conclude that the reduced model is as good as the full model. Thus there is no advantage to including H_Education, Quality, Taste, Nutrition,smell, Easy_buy, Brand name, Easy_use and Other_influence in the model.

Akaike Information Criterion (AIC)

The following Table 4.4 shows the AIC values for different proposed models.

Table 4.4: Comparison of the AIC Values

Model	AIC Value
Full Model (Fullmodel)	561.84
Model with variables are selected by back ward elimination method (Reducedmodel)	539.08

Based on the AIC criteria, we should select the best model since it has the smallest value of AIC. According to the results of Table 4.4, among the two models, the second model having the minimum value of AIC. The model with variables selected by backward elimination method is the best model which included Age of the respondent, Educational level, Monthly Income, Price, Thickness, Easy to melt, Artificial Ingredient and advertisement. Thus, these variables are important and should be in the model.

4.5 Assessing the Fitted Model

4.5.1. Hosmer Lemeshow Goodness-of-fit Test

By assessing a few summary measurements, we can check the predictive power of the selected model. The summary measures of goodness-of-fit, give an overall indication of the fitted model. The commonly used summary measures Goodness-of-fit is represented in Table 4.5. This test indicated that significance of the overall logistic regression model and adequately used for predictions.

Table 4.5: Summary measure of Hosmer-Lemeshow test.

Summary Statistics	**Value**	**df**	**P-value**
Hosmer-Lemeshow	7.5788	8	0.4757

In the Hosmer-Lemeshow goodness-of-fit test, the subjects were divided into nearly ten groups based on the percentile of the estimated probabilities. According to Table 4.5, the p-values of Hosmer-Lemeshow test is greater than 0.05 (significance level). This indicated that we do not reject the null hypothesis, which the model is fit well. Thus, we can say that the fitted logistic model performance at an acceptable level.

4.5.2. ROC Curve

In this analysis, ROC (Receiver Operating Characteristics) curve also considered to measure the model's predictive power, which is one of the useful summary measures in logistic regression. The predictive capability of the fitted model can be quantified by the area under the ROC curve. This curve has plotted the probability of correctly classifying a positive response (Sensitivity) against the probability of incorrectly classifying a negative response (1-Sensitivity) for the entire set of possible cut-off-point.

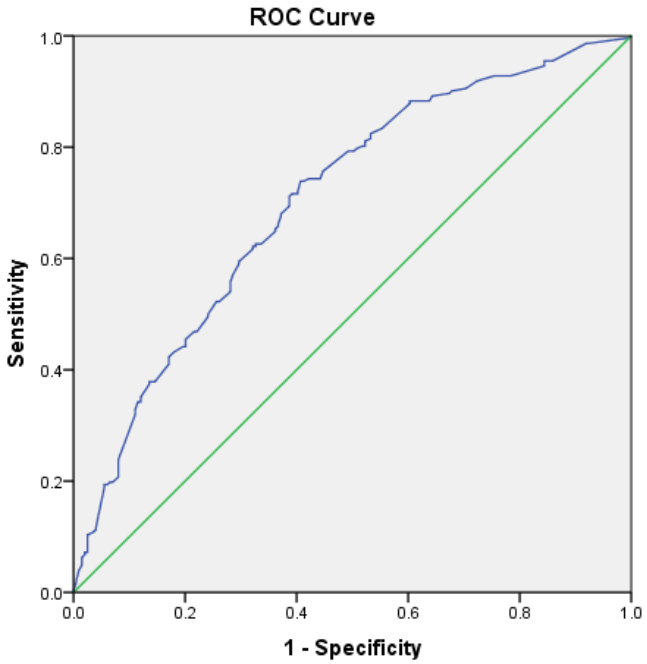


Figure 4.7: ROC Curve

Figure 4.7, presents the ROC curve for the fitted model. The area under the ROC curve is

0.7487, which indicated that the predictive ability is satisfactory.

Therefore, the model with variables selected by backward elimination method is the best m odel which included Age of the respondent, Educational level, Monthly Income, Price, Thi ckness, Easy to melt, Artificial Ingredient and Advertisement. Table 4.6 shows the odds rat ios of each coefficient and 95% confidence Interval for each odds ratio.

Table 4.6: Odds ratios and 95% Confidence Interval to the odds ratios for the final fitted model.

Variables	**OR**	**95% CI of OR**
(Intercept)	1.01	(0.20,4.91)
Age	1.03	(1.00,1.05)
Education(3)	0.47	(0.18,1.26)
Education(2)	0.37	(0.14,0.96)
Education(1)	0.17	(0.06,0.51)
Arf_ingredient(2)	0.66	(0.36,1.21)
Arf_ingredient(1)	0.51	(0.31,0.83)
Advertisement(2)	1.61	(0.76,3.40)
Advertisement(1)	1.92	(1.07,3.47)
MonthlyIncome(4)	2.08	(0.48,8.92)
MonthlyIncome(3)	1.02	(0.31,3.10)
MonthlyIncome(2)	1.38	(0.49,3.89)
MonthlyIncome(1)	2.11	(0.86,7.41)
Price(2)	0.35	(0.19,0.68)
Price(1)	0.27	(0.17,0.44)
Thickness(2)	0.59	(0.26,1.37)
Thickness(1)	0.47	(0.26,0.89)
Easy_melt(2)	3.75	(1.62,8.69)
Easy_melt(1)	2.13	(1.14,3.99)

4.6 Discussion

Based on the comparison of the AIC value of the Fullmodel and Reducedmodel (using backward elimination method), it indicated that the Reducedmodel has a minimum AIC value than Fullmodel. Therefore, we can conclude that the Reducedmodel (model selected by backward elimination method) is more appropriate for this study.

According to Backward elimination method the following variables can be identified as the most important variables with milk consumption in Matara district. The variables are; age, Education Level of the respondent, Monthly Income, Price of the milk, Thickness, the attitude of easy to melt, Artificial Ingredients and Advertisements.

After model selection, the assessment of the fitted model has become an important step in the model building. This study was demonstrated a comparison of the Hosmer-Lemeshow test for the fitted model which indicated that the fitted model fits well. The predictive capability of the fitted model can be quantified by the area under the ROC curve. The area under the ROC curve is 0.7487, which indicated that the fitted model is reasonable to predict.

According to the fitted model, the odds ratios are less than one, on the subjects with Education, Artificial Ingredients in the milk, Price and Thickness of the milk.

- The consumption of local milk is less likely to occur among those who disagree considering the artificial ingredients than among who agree considering the artificial ingredients. The consumers highly considered about artificial ingredients when they buy local milk. Furthermore, they believed that, the local milk does not contain an artificial ingredient than the imported milk.

- Similarly, the consumers considered about the price of the milk when they buy local milk. Furthermore, they believed that, the local milk is cheaper than the imported milk.

- The consumers considered about the thickness of the milk when they buy local milk. Furthermore, they believed that, the local milk has thickness compared with the imported milk.

According to the fitted model, the odds ratios are greater than one, on the subjects with Age, Advertisements, Monthly Income and Easy to melt.

- The consumption of local milk is more likely to occur among those who disagree with the considering advertisement for their milk choice than among who agree with the considering advertisement. The most of the local milk consumers not considering about an advertisement for their milk choices.

- Similarly, the most of the local milk consumers not satisfied with the attitude of easy to melt. The consumers with lower income level more likely to buy local milk than imported milk for their day today milk requirements.

CONCLUSIONS

This study is based on the consumers' preferences for milk choices either local or imported milk in Matara district. Based on the results, the majority of the consumers are interested in local milk then imported milk.

Further, in this study some factors affecting the preference of local and imported milk consumption in Marata district were examined. The findings of this study suggest that the socioeconomic and demographic factors of the householders play an important role in milk consumption.

- According to the results it can be concluded that the families with members more than four or above are interested in local milk than imported milk.
- Consumer education is one of the primary reasons for purchasing local milk. When the education level of consumers increases substantially, their preferences shift from imported milk to local milk.

Then the study has focused on the associated demographic, socioeconomics and attitudinal factors of consumers' with the milk consumption in Matara district. Based on the results of chi-square test, household head education, price of the milk, included nutrition, the property of easy to melt, include artificial ingredients and advertisements for marketing are the statistically associated variables with milk type.

- The price of the milk is a reasonable factor influencing consumer's milk choices. The result shows that the prices of imported milk are higher than those of local milk. Therefore, in this study, a higher number of consumers are interested in local milk than imported milk.
- Majority of consumers believed that the local milk contains necessary nutrition for the human body than the imported milk.
- Results revealed that the majority of consumers believed that local milk does not contain artificial ingredients compared with imported milk.

- The property of melting is also one of the considering factor by consumers on their milk pattern. The results indicated that the imported milk is easily melt than the local milk.
- Advertisements are also affective factor for consumers' milk choices. Empirical findings of this study have provided important information to marketers to further marketing strategies and provide products with high quality to meet the needs of consumers.

There is a high degree of willingness to buy local milk when these constraints are addressed. Understand consumer preference for local milk and trends in consumption and their impact on determining dairy production and marketing opportunities. Hence, there is a good opportunity for the development of the production and marketing system of milk in the country, while minimizing dependency on imported products.

References

1. Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (Second Edition ed.). A John Wiley & Sons.
2. Allison, P. D. (2014). Measures of Fit for Logistic Regression. *SAS Global Forum*.
3. Arboretti Giancristofaro, S. L. (2003). Model Performance Analysis and Model Validation. *Statistica*, LXIII.
4. Badi, N. H. (2017). Asymptomatic Distribution of Goodness-of-Fit. *Open Journal of Statistics*, 434-445.
5. Bus A E M, W. A. (2003). Consumers' sensory and nutritional perceptions of three. *Public Health Nutrition*, 6(2), 201–208.
6. David W. Hosmer, S. L. (2000). *Applied Logistic Regression,Second Edition* (Second edition ed.). Canada: A Wiley Interscience Publication.
7. De Alwis A.E.N, E. J. (2009). Analysis of Factors Affecting Fresh Milk Consumption Among the Mid-Country Consumers. *Agricultural Research and Extension*.
8. Hallett, D. C. (1999). *Goodness of Fit Test in Logistic Regression.*
9. Health, D.O. (2012). *A Survey of Infant and Young Child Feeding in Hong Kong.* Hong Kong.
10. Hosmer D.W, H. T. (1997). A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model. *Statistics in Medicine*, 965Ð980.
11. Hsu, J. L. (2018). Consumption and attribute perception of fluid milk in Taiwan. *Emerald, Vol. 36*(Issue: 3), 177-182.
12. Jane Lu Hsu, Y.-T. L. (2006). Consumption and attribute perception of fluid milk in Taiwan. *EEmerald Group Publishing Limited*, Vol. 36 No. 3, pp. 177-182.
13. Katsaragakis, S., Koukouvinos, C., Stylianou, S., Theodoraki, E.-M., & and Theodoraki, E.-M. (2005). Comparison Of Statistical Tests In Logistic:The Case Of Hypernatreamia. *Journal of Modern Applied Statistical Methods*, 514-521.
14. Mengchao Wang, J. W. (2013). A Comparison of Approaches to Stepwise Regression for Global Sensitivity Analysiis used with Evolutionary Optimization. Chambery, France: International Building Performance Simulation association.
15. Pathumsha, V. (2016). Daily Industry Trends in Sri Lanka.

16. Plecher. (2019). *Sri Lanka: Population growth from 2007 to 2017 (compared to previous year)*. Retrieved from Statista: https://www.statista.com/statistics/728536/population-growth-in-sri-lanka/

17. *Price Controls in daily Industry*. (n.d.). Retrieved from https://www.research.advocata.org/price-controls-in-the-dairy-industry/

18. Sarkar S.K, H. M. (2010). Importance of Assessing the Model Adequacy of Binary Logistic Regression. *Journal of Applied Science, 10(6)*(1812-5654), 479-486.

19. Sarkar S.K, H. M. (2010). Model Selection in Logistic Regression and Performance of its Predictive Ability. *Australian Journal of Basic and Applied Science*(ISSN 1991-8178), 5813-5822.

20. Smyth, G. K. (n.d.). *Pearson's Goodness of Fit Statistic as a Score Test*. Retrieved from https://projecteuclid.org/download/pdf_1/euclid.lnms/1215091138

21. Statistics, D. O. (2018). *Agriculture and Environment Statistics Devision*. Retrieved from Department of Census and Statistics : http://www.statistics.gov.lk/agriculture/Livestock/MilkProduction.html

22. Tiskumara. (2015). *Livestock Statistical Bulletin.* Peradeniya, Sri Lanka: Department of Animal Production and Health.

23. V, P. H. (2016). Daily Industry Trends in Sri Lanka.

24. Wang, M. (2013). A comparison of Approaches to Stepwise Regression for Global Sensitivity Analysis used with Evolutionary Optimization. France.

25. *World Cows' Milk Production (in tonnes)*. (2014, April 04). Retrieved from ChartsBin.com: http://chartsbin.com/view/23557

26. Yayar, R. (2012). Consumer Characteristics Influencing Milk Consumption Preference. The Turkey case. *Theoretical and Applied Economics, XIX,No.7(572)*, 25-42.

27. Zibran, M. F. (n.d.). *Chi-Squared Test of Independence.* University of Calgary, Alberta, Canada.

APPENDIX A: R Codes

Read the Data Set

```
setwd("D:/msc/Final project/final_dataset")

library(haven)

new_data<- read_sav("D:/msc/Final project/final_dataset/final.sav")

View(new_data)

attach(new_data)
```

Descriptive Statistic

```
a<-table(milktype)

barplot(a,col=c("light pink"),beside=TRUE, main="Type of Milk Consumption",
ylab="Frequency", xlab="Milk Type", names.arg = c("Imported","Local"),
width=c(3,3) )

b<-table(milktype,No_Members)

barplot(b,beside=TRUE,col=c("green","light pink"), main="Type of Milk with Number of
Members", ylab="Frequency", xlab="Number of members",legend=c("Imported","Local"),
args.legend = list(x="topright",text.width=4.2))

c<-table(milktype,Monthly_Income)

barplot(c,col=c("green","light pink"),beside=TRUE,legend=c("Imported", "Local"),
main="Type of Milk Consumption according to Monthly Income ", names.arg =
c("<35000","35000-50000","50000-65000","65000-80000",">80000"), las=2,args.legend
= list(x="topright",text.width=2.0),cex.axis=0.8, cex.names=0.8)

d<-table(milktype,Education)

barplot(d,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Type of Milk Consumption According to Educational Attainment",
cex.main=1,xlab="Educational Attainment",ylab="Frequency", names.arg =
c("O/L","A/L","Graduate","Professional"), args.legend =
list(x="topright",text.width=1.4,text.font=1,cex=0.6))
```

```
e<-table(milktype,H_Education)

barplot(e,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Type of Milk Consumption According to Head Educational Attainment",
cex.main=0.95,xlab="Educational Attainment", ylab="Frequency",cex.lab=0.8,
names.arg = c("O/L","A/L","Graduate","Professional"),args.legend =
list(x="topright",text.width=1.0,text.font=1,cex=0.6))


f<-table(milktype,Price)

barplot(f,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Price of Milk",cex.lab=0.6,xlab="Your Milk Type has less Price",
cex.lab=1,names.arg = c("Disagree","Not Agree/Disagree","Agree"), args.legend =
list(x="topright",text.width=0.8,text.font=0.5, cex=0.6,inset=c(-0.1, -0.3)))


g<-table(milktype,Nutrition)

barplot(g,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Nutrition of Milk",xlab="Your Milk Type has necessary Nutritions",
cex.lab=1,names.arg = c("Disagree","Not Agree/Disagree","Agree"), args.legend =
list(x="topleft",text.width=0.8,text.font=1,cex=0.6, inset=c(-0.1, -0.3)))


h<-table(milktype,Easy_melt)

barplot(h,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Easy to Melt of Milk",xlab="Your Milk Type has an attitude, easy to melt",
names.arg = c("Agree","Not Agree/Disagree","Disagree"), args.legend =
list(x="topleft",text.width=0.9,text.font=1,cex=0.8, inset=c(0, 0)))


i<-table(milktype,Arf_ingredient)

barplot(i,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Artificial Ingredients of Milk",xlab="Your Milk Type has an Artificial
ingredients", names.arg = c("Disagree","Not Agree/Disagree","Agree"), args.legend =
list(x="topleft",text.width=0.9,text.font=1,cex=0.8))

j<-table(milktype,Thickness)

barplot(j,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Thickness of Milk",xlab="Your Milk Type has a quality of Thickness",
names.arg = c("Disagree","Not Agree/Disagree","Agree"),args.legend =
list(x="topleft",text.width=0.9,text.font=1,cex=0.8))
```

```
k<-table(milktype,Smell)

barplot(k,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Smell of Milk",xlab="Your Milk Type has a good Smell", names.arg =
c("Disagree","Not Agree/Disagree","Agree"), args.legend =
list(x="topleft",text.width=0.9,text.font=1,cex=0.8))


l<-table(milktype,Quality)

barplot(l,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Quality of Milk",xlab="Your Milk Type has a good Quality", names.arg =
c("Disagree","Not Agree/Disagree","Agree"), args.legend =
list(x="topleft",text.width=0.9,text.font=1,cex=0.8))


m<-table(milktype,Taste)

barplot(m,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Taste of Milk",xlab="Your Milk Type has a good Taste", names.arg =
c("Disagree","Not Agree/Disagree","Agree"), args.legend =
list(x="topleft",text.width=0.9,text.font=1,cex=0.8))


n<-table(milktype,Advertisement)

barplot(n,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Consideration of Advertisement",xlab="Do you consider Advertisements for your
Milk choice?", names.arg = c("Disagree","Not Agree/Disagree","Agree"),  args.legend =
list(x="topright",text.width=0.9,text.font=1,cex=0.8))


p<-table(milktype,Brand_name)

barplot(p,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Brand Name of Milk",xlab="Do you consider the Brand Name of your Milk?",
names.arg = c("Disagree","Not Agree/Disagree","Agree"),  args.legend =
list(x="topleft",text.width=0.9,text.font=1,cex=0.8))


q<-table(milktype,Easy_buy)

barplot(q,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Availability of Buy",xlab="Can you buy your Milk type at the Market very
easily?", names.arg = c("Disagree","Not Agree/Disagree","Agree"), args.legend =
list(x="topleft",text.width=0.9,text.font=1,cex=0.8))
```

```
r<-table(milktype,Easy_use)

barplot(r,col=c("green","light pink"),beside=TRUE,legend=c("Imported","Local"),
main="Convenient to Use ",xlab="Is your Milk convenient to Use?",names.arg =
c("Disagree","Not Agree/Disagree","Agree"), args.legend =
list(x="topleft",text.width=0.9,text.font=1,cex=0.8))
```

Create Design Variables

```
options(contrasts = c("contr.SAS","contr.SAS"))

EducationDV<-C(factor(Education),contr=treatment)

EducationDV<-relevel(factor(EducationDV),ref="3")

H_EducationDV<-C(factor(H_Education),contr=treatment)

H_EducationDV<-relevel(H_EducationDV,ref="3")

Monthly_IncomeDV<-C(factor(Monthly_Income),contr=treatment)

Monthly_IncomeDV<-relevel(Monthly_IncomeDV,ref="4")

QualityDV<-C(factor(Quality),contr=treatment)

QualityDV<-relevel(QualityDV,ref="2")

PriceDV<-C(factor(Price),contr=treatment)

PriceDV<-relevel(PriceDV,ref="2")

TasteDV<-C(factor(Taste),contr=treatment)

TasteDV<-relevel(TasteDV,ref="2")

NutritionDV<-C(factor(Nutrition),contr=treatment)

NutritionDV<-relevel(NutritionDV,ref="2")

ThicknessDV<-C(factor(Thickness),contr=treatment)

ThicknessDV<-relevel(ThicknessDV,ref="2")

Easy_meltDV<-C(factor(Easy_melt),contr=treatment)

Easy_meltDV<-relevel(Easy_meltDV,ref="2")

SmellDV<-C(factor(Smell),contr=treatment)

SmellDV<-relevel(SmellDV,ref="2")

Easy_buyDV<-C(factor(Easy_buy),contr=treatment)
```

```
Easy_buyDV<-relevel(Easy_buyDV,ref="2")

Arf_ingredientDV<-C(factor(Arf_ingredient),contr=treatment)

Arf_ingredientDV<-relevel(Arf_ingredientDV,ref="2")

other_influenceDV<-C(factor(other_influence),contr=treatment)

other_influenceDV<-relevel(other_influenceDV,ref="2")

AdvertisementDV<-C(factor(Advertisement),contr=treatment)

AdvertisementDV<-relevel(AdvertisementDV,ref="2")

Brand_nameDV<-C(factor(Brand_name),contr=treatment)

Brand_nameDV<-relevel(Brand_nameDV,ref="2")

Easy_useDV<-C(factor(Easy_use),contr=treatment)

Easy_useDV<-relevel(Easy_useDV,ref="2")
```

Binary Logistic Regression Model with all predictor variables (Fulmodel)

```
Fullmodel<-glm(milk_type ~ Age + EducationDV + H_EducationDV + Monthly_Income
               DV + QualityDV + PriceDV + TasteDV + NutritionDV + ThicknessDV +
               Easy_meltDV + SmellDV + Easy_buyDV + Brand_nameDV + Easy_useDV
              +Arf_ingredientDV + other_influenceDV + AdvertisementDV,data=new_data,
              family=(binomial("logit))

Summary (Fulmodel)
logLik(Fulmodel)

exp(cbind("Odds ratio" = coef(Fullmodel), confint.default(Fullmodel, level = 0.95)))
```

Binary Logistic Regression with backward elimination method

```
Reducedmodel<-glm( milk_type ~ Age + EducationDV + Monthly_IncomeDV + PriceDV
                   +ThicknessDV + Easy_meltDV + Arf_ingredientDV + AdvertisementDV
                    , data=new_data,family(binomial("logit"))
summary (reducedmodel)
logLik(Reducedmodel)

exp(cbind("Odds ratio" = coef(Reducedmodel), confint.default(Reducedmodel, level =
0.95)))
```

Hosmer and Lemeshow Goodness of fit Test

```
library(ResourceSelection)
h1<-hoslem.test(Reducedmodel$y,fitted(),g=10)
h1
```

Plot ROC Curve

```
install.packages("pROC")
library(pROC)
prob=predict(model,type=c("response"))
prob=predict(Reducedmodel)
new_data$prob=prob
h <- roc(milk_type ~ prob, data = new_data)
plot(h)
auc(h)
```

APENDIX B: Sample Questionnaire

<u>ප්‍රශ්නාවලිය : මාතර දිස්ත්‍රික්කයේ කිරි පරිභෝජනය</u>

රුහුණු විශ්ව විද්‍යාලයේ ගණිත අධ්‍යනාංශයේ ආධුනික කථිකාචාර්යවරියක් ලෙස සේවය කරන මධුෂා දිල්ශානි වන මාගේ, පශ්චාත් උපාධියේ අවසන් පරීක්ෂණ නිබන්ධන උදෙසා මාතර දිස්ත්‍රික්කයේ දේශීය කිරි හා ආනයනික කිරි පරිභෝජනය පිලිබඳව සමීක්ෂණයක් සිදු කරනු ලබයි. ඒ සඳහා පහත ප්‍රශ්නාවලියට පිළිතුරු ලබා දෙමින් සහයෝගය ලබා දෙන මෙන් කාරුණිකව ඉල්ලා සිටිමි.

පිළිතුරු ලබා දෙන ඔබගේ පෞද්ගලිකත්වය සුරකින අතර මෙම පිළිතුරු පත්‍ර , සමීක්ෂණයේ විශ්ලේෂණයට පමණක් භාවිතා කරනු ලබයි.

ප්‍රශ්නාවලිය සම්බන්ධව යම්කිසි ගැටලුවක් ඇත්නම් පහත දුරකථන අංකයට අමතන්න.

දු. අ : 071-5987172

1) වයස : ☐

2) වෘත්තීය : _____

3) ස්ත්‍රී / පුරුෂ : ☐ ස්ත්‍රී ☐ පුරුෂ

4) විවාහක /අවිවාහක : ☐ විවාහක ☐ අවිවාහක

5) පවුලේ මුළු සාමාජිකයන් ගණන : ☐

6) පවුලේ සාමාජිකයන්ගේ වයස අනුව ;

i) වයස 0 -12 අතර සාමාජිකයින් ගණන

ii) වයස 13 -19 අතර සාමාජිකයින් ගණන

iii) වයස 20 -35 අතර සාමාජිකයින් ගණන

iv) වයස 36 -59 අතර සාමාජිකයින් ගණන

v) වයස 60 ට වැඩි සාමාජිකයින් ගණන

7) ඔබගේ අධ්‍යාපන මට්ටම

i) 1 – 9 ශ්‍රේණිය දක්වා

ii) අ.පො.ස්. (සා.පෙළ) සමත්

iii) අ.පො.ස්. (උ. පෙළ) සමත්

iv) උපාධිධාරී /උපාධි අපේක්ෂක

v) පශ්චාත් උපාධිධාරී / පශ්චාත් උපාධි අපේක්ෂක

vi) වෘත්තීය සුදුසුකම් (CIMA, Chartered,...)

8) නිවසේ ගෘහ මුලිකයාගේ අධ්‍යාපන මට්ටම

i) 1 – 9 ශ්‍රේණිය දක්වා

ii) අ.පො.ස්. (සා.පෙළ) සමත්

iii) අ.පො.ස්. (උ. පෙළ) සමත්

iv) උපාධිධාරී /උපාධි අපේක්ෂක

v) පශ්චාත් උපාධිධාරී / පශ්චාත් උපාධි අපේක්ෂක

vi) වෘත්තීය සුදුසුකම් (CIMA, Chartered,...)

9) ඔබගේ මාසික ආදායම් මට්ටම

i) රු 35,000 ට අඩු

ii) රු 35,000 -50,000

iii) රු 51,000 -65,000

iv) රු 66,000 -80,000

v) රු 80,000 ට වැඩි

10) ශ්‍රී ලංකාවේ ප්‍රධාන වශයෙන් කිරි පරිභෝජනය දේශීය කිරි හා ආනයනික කිරි ලෙස වර්ග කල හැකිය. ඒ අතුරින් ඔබගේ කිරි පරිභෝජනයට පසුබිම් වූ පහත සාධක පිලිබඳව ඔබ දරන අදහස් මොනවාද?

සාධක	කිසිසේත්ම එකඟ නොවේ	එකඟ නොවේ	එකඟ වන්නේද නැත නොවන්නේද නැත	එකඟ වේ	සම්පුර්ණයෙන්ම එකඟ වේ
තන්ව සහතිකයක් ඇත.					
අනෙක් කිරි වලට වඩා අඩු මිලක් ඇත.					
ප්‍රණීත රසයක් ඇත.					
ශරීරයට අත්‍යාවශ්‍ය පෝෂණ සංඝටක ඇත.					
උකු බවින් යුක්ත වේ.					
පහසුවෙන් දිය වේ.					
ප්‍රියජනක සුවඳක් ඇත.					
වෙලඳපොලේ පහසුවෙන් මිළදී ගැනීමට හැකිය.					
පිළිගත් වෙළඳ නාමයක් ඇත.					
පහසුවෙන් භාවිතා කල හැකිය/අසුරා තබා ගත හැකිය.					
කල්තබා ගැනීමේ ද්‍රව්‍ය/කෘතීම					

රසකාරක භාවිතය සලකා බැලුවේද?					
වෙළඳ දැන්වීම් වලින් බලපෑමක් සිදුවූවාද?					

11) දිනපතා පරිභෝජනයේදී ඔබ බහුලව භාවිතා කරනු ලබන්නේ ,

☐ දේශීය කිරි ☐ ආනයනික කිරි

12) කිරි පරිභෝජනය සඳහා ඔබ මාසිකව වැය කරන මුළු මුදල,

☐ රු. 1,000 ට අඩු
☐ රු. 1,000-1,500
☐ රු. 1,500-2,000
☐ රු. 2,000-2,500
☐ රු. 2,500 ට වැඩි

13) ඔබ දේශීය කිරි පරිභෝජනය කරන්නෙක් නම් ඉහත ප්‍රශ්න අංක (10) හි සඳහන් කල සාධක වලට අමතරව ඔබ විසින් සලකා බලනු ලබන අනෙකුත් සාධක මොනවාද?

.....

.....

.....

14) ඔබ ආනයනික කිරි පරිභෝජනය කරන්නෙක් නම් ඉහත ප්‍රශ්න අංක (10) හි සඳහන් කල සාධක වලට අමතරව ඔබ විසින් සලකා බලනු ලබන අනෙකුත් සාධක මොනවාද?

.....

.....

.....