UKZN

STATISTICAL ANALYSIS OF THE ATTITUDES TOWARDS BLOOD DONATION AND TRANSFUSION IN MALI

Farzana Osman

Project submitted to the University of KwaZulu-Natal in fulfilment of the requirements for the Master's degree in Statistics.

Thesis advisor: Professor Delia North

Professor Temesgan Zewotir

UNIVERSITY OF KWAZULU-NATAL

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Declaration of Authorship

This thesis and the research work done therein was carried out in the School of Mathematics, Statistics and Computer Science, University of Kwa Zulu-Natal, Westville Campus. The research work contains the original work of the author and has not been submitted in any form for any degree or diploma to any other University. Where the work of others has been quoted, it is duly acknowledged and referenced in the bibliography.

Miss Farzana Osman

Signed: . . . . . .

Date . . . . . .

Prof Delia North

Signed: . . . . . .

Date . . . . . .

Prof. Temesgen Zewotir

Signed:……

Date……

Acknowledgements

I would like to express my sincere gratitude to my supervisors; Professor Delia North and Professor Temesgan Zewotir for their guidance, encouragement and advice provided to me throughout my time as a student.

I would also like to thank SACEMA (South African Centre for Epidemiological Modelling and Analysis) for all the financial and academic support. Their research days were inspiring and motivating and also provided me the opportunity to meet many wonderful people.

I am fortunate to have been blessed with a very supportive and loving family and I would like to thank each and every one of them, especially my parents and my husband for their continuous love and support throughout my studies.

Moreover, I would like to thank all members of the Department of Statistics at UKZN and at CAPRISA.

Abstract

The demand for blood transfusion in Mali is high, because of the high prevalence of anemia, which is mostly caused by malaria, malnutrition and pregnancy-related complications. In this study a classic KNOWLEDGE, ATTITUDE AND PRACTICE (KAP) SURVEY was conducted on 323 individuals in Mali. Questions asked were aimed at finding what people in the study know about blood donation, how they feel about donating and receiving blood, and how they behave when asked to donate blood. The objective of this study is to develop a theoretical framework to better understand the attitudes toward blood donation and transfusion in Mali, thereby identifying factors that motivate and deter blood donation, and also to identify interventions to improve the supply of blood transfusion.

A main effect logistic regression model was carried out to the model the relationship between willingness status of blood donating and thirteen explanatory variables. Multiple correspondence analysis was used to confirm the results obtained. Due to the nonresponse in the survey, techniques used to handle missing data values were also explored.

More than 50% of individuals in the study responded as non-donors, however a vast majority of respondents reported their intent to become future donors. Also, the male population responded as majority donors at 58.8%. Results found, indicate that females were less likely to be donors in the Mali population and individuals that had knowledge about the different type of blood groups were more inclined to be donors. Overall results produced from the statistical methods employed in this study were consistent across the methods.

Key Words: Blood donation, transfusion, blood donors, voluntary non-remunerated blood donors, replacement donors, logistic regression, multiple correspondence analysis, missing data, subset correspondence analysis.

Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Blood Donation

For any country, a safe and adequate supply of blood for transfusion is essential. In many developing countries it is often found that the amount of blood available is insufficient or far less than what is required or blood donated may not be safe enough for transfusion. Only 39% of the world's blood supply is donated in developing countries although they have 82% of the global population (World Health Organization, 2004b). In these countries, there is therefore a tendency to rely on family blood donors. These family blood donors are referred to as family replacement donors and they give blood when it is required by a member of the donor's family or community.

Bates et al. (2008) have reported that overall, 80% of blood for transfusion in sub-Saharan Africa comes from replacement donors. These donors are often not the perfect choice for a reliable supply of blood due to their association with an increased risk of transfusion-transmitted infections (TTI's).

According to Ahmed et al. (2007) since the seventeenth century, it has been known that the transfusion of blood between individuals could have rapid and fatal consequences. They have reported that to ensure safety, the blood is tested to determine its blood group and to check that the blood is not contaminated with harmful microorganisms or infectious diseases.

TTI's are of a major concern and present vital challenges to blood transfusion services in developing countries and infections include human immunodeficiency virus (HIV), hepatitis B and C, syphilis and malaria (Please refer to Appendix A for challenges associated with TTIs).

Volberding et al. (2008) have suggested the importance of public health systems to increase and promote the use of voluntary blood donors since they are least likely to transmit TTI's whilst reducing and eventually stop paying for blood donated.

In order to ensure a safe blood supply in Africa, the Safe Blood for Africa Foundation (SBFA) was established as a multi-year, stepped implementation plan to establish the facilities and train the professionals needed to manage, track and test the millions of blood transfusions performed in sub-Saharan Africa (Volberding et al., 2008). According to Volberding et al. (2008), the SBFA foundation programs are currently being implemented in 18 African countries and is used to provide training for over 500 blood banking technicians a year, supplying test kits and supplies and also furnishing technical assistance.

1.2 Mali

Mali is a landlocked, predominantly Muslim country in West Africa that gained independence from French colonial rule in 1960. In the Africa Survey, Endres (2013) reported that the total estimated population as of mid-2012 was approximately 16 014 000. The official language of the country is French, however there are over 30 other languages spoken in the country as reported by Velton (2009). The country ranks amongst the poorest countries in the world and is greatly affected by malnutrition and insufficient sanitation. According to Velton (2009), the economy of the country is mostly based on agriculture which accounts for 45% of the country's GDP and due to the dependence on the agricultural sector, the country is vulnerable to environmental shocks. Velton (2009) reported that over 60% of the Malian population still lives below the poverty line with the majority living in rural areas. He further reported that access to medical supplies is fairly limited and the country depends heavily on foreign aid. Endres (2013) reported that the estimated total health expenditure per capita in Mali is US$45 and as of 2005 – 2011 there were 8 physicians per 100 000 inhabitants with about 10 hospital beds per 100 000 people in that time period.

Malaria and other arthropod-borne diseases are prevalent in Mali, as are a number of infectious diseases such as syphilis and HIV/AIDS. Endres (2013) claimed that in the year 2009, there were 1, 633, 423 cases of malaria reported in Mali and in the year 2010 there were 138 malaria deaths per 100 000.

Figure 1.1: Map of Mali

According to Physicians for Peace (PFP, 2015), those patients in referral hospitals outside the capital of Bamako in Mali, must rely on family members for blood donation. There is no capability to collect, screen or process blood, so these blood services are usually carried out at a basic level using rapid diagnostic testing. The blood may be tested for blood type and major infectious diseases but is generally performed from one individual directly to the next individual, which is known as 'vein to vein' transfusion. Also, due to the lack of blood supply in the country, transfusions occur in an emergency state which frequently leads to mistakes and a lack of quality control with regard to blood transfusion services. Individuals that are desperate for the need of blood and cannot rely on family replacement blood donors for some reason or the other, often tend to pay for blood donation. It is widely known that individuals that are willing to sell their blood are generally from high risk populations and are potentially at risk to lead to exposure to transfusion transmissible infections.

According to Erhabor et al. (2013), a previous study to determine the risk of transfusion-transmittable syphilis infection among Malian blood donors has shown a seroprevalence rate of 0.3% and a higher risk among donations from first time and replacement donors compared to voluntary and repeated donors.

Erhabor et al. (2013) have also reported that two studies to investigate the risk of transfusion transmissible HIV infection among Malian blood donors have indicated a prevalence of 2.6% and 4.5%, respectively. They have further reported that a cross-sectional study conducted to assess the prevalence of hepatitis B virus (HBV), and its co-infection among blood donors at the National Blood Transfusion Center in Bamako, Mali, have indicated a prevalence of 14.9% and a HIV/HBV co-infection rate of 1.13% among 11 592 blood donors.

With a maternal mortality ratio of 1200 deaths per 100 000 live births, Mali ranks among the top 10 countries in which women face the highest risk of death during pregnancy and childbirth (WHO, 2004a). Due to the high maternal mortality rates in the country, the people of Mali approached the PFP to help reduce these rates of which the root cause was lack of access to safe blood. This resulted in a collaboration between the American Red Cross, Millennium Cities Initiative, Safe Blood for Africa and the Mali Ministry of Health.

Before the collaboration with these organizations, the country had only one poorly equipped blood bank in the capital city Bamako. This partnership led to a signed memorandum of agreement with the Mali Ministry of Health, to fully equip and provide training for a highly capable blood bank in Ségou, which is the capital of Mali's fourth largest administrative region.

SBFA commenced its intervention in the country in 2012 but was interrupted by the civil strife until mid-2013. Although there were low levels of activity during this time, SBFA reported that a complete Blood Safety Assessment and planning events had taken place in Bamako, Ségou and Kita. The data used in this study was collected by the team across these three different regions in Mali. Data collection methods will be discussed in the next chapter.

Identifying the motivational factors that may affect blood donation and the recruitment of safe low-risk donors in developing countries, particularly Mali, is needed. A variety of factors may influence an individual's willingness to donate blood. Many studies have reported that there is a shortfall in recruitment of blood donors which is rooted in

culture, education and marketing. Bloch et al. (2012), have stated that education and literacy are also notable obstacles to recruitment of blood donors and have reported that in a study in Burkina Faso, 30.8% of blood donors were illiterate or of primary school level.

Ignorance and being unaware of the need for blood or other aspects of the donation process has been consistently identified as a negative factor in potential donor decision making (Gillepsie & Hillyer, 2002). Also, according to Aldamiz-echevarria & Aquirre-Garcia (2014), a representative sample of 1,350 among the population of Spanish people, found that 40% of the respondents said they had not seen or heard anything about blood donation in the last month and, if they do not hear or do not remember hearing anything, it cannot influence them.

In cultures which have little practice and knowledge of blood donation, there may be many concerns, myths and misconceptions ranging from fear of needles or fainting to beliefs that blood donation results in a loss of strength or that a disease can be contracted by donating blood. It is therefore essential to identify public perceptions and address them directly, working in partnership with the media and the community which can reach out to large numbers of people (WHO, 2010).

The method each country adopts to attract blood donors and to cover its needs in blood supplies varies. Hence the basic idea behind this study is to investigate which variables offer the best explanatory power that can predict blood donation in Mali.

The main objective of this study is to develop a theoretical framework to better understand the attitudes toward blood donation and transfusion in Mali. It also aims to identify factors that motivate and deter blood donation in Mali, as well as to identify interventions to improve the supply of blood transfusion.

Chapter 2 : Data Collection and Descriptive Report

2.1 Data Collection

A total of 323 individuals were interviewed across three regions in Mali (Kita, Bamako and SGO). This sample size was determined in order to optimize the resource usage and design of the study, i.e. improving the chance of conclusive results with maximum efficiency. The aim was to set the sample size to have an at least 80% chance of establishing differences (between blood donors and non-donors proportions) with an effective difference of 7.5% from the hypothetical proportion of 50% (no difference in blood donation likelihood as compared to non-donation) at a nominal significance level of 5%. Under this set up, of power of test, the sample size needed is at least 347. With the cost and time factor taken into consideration for chance of percent type I error and a 20% chance of type II error (i.e., 80% power of test), a sample size of 323 was found to be a reasonable sample size to have a descriptive report giving overview and insight.

Questions posed were aimed at finding out what people in the study knew about blood donation, how they felt about donating and receiving blood, and also how they behaved when asked to donate blood, i.e., KNOWLEDGE, ATTITUDE AND PRACTICE (KAP) SURVEY.

The questionnaire was organized in seven sections. Section one of the questionnaire consisted of interviewee characteristics, while section two was to be completed only if a respondent was a donor. Section three was to be completed by respondents that had never donated blood before (non-donors), whilst section four focused on a respondent's knowledge about blood donation. Section five was based on respondents' attitudes towards blood donation, section six was based on communication channels and behavior of respondents and section seven was aimed at establishing some socio-demographic factors related to the respondents.

The questionnaire designers strived to have exhaustive and comprehensive information from the respondents. Reportedly the interviewers were trained to administer the questionnaire efficiently. The questions posed by the designers are quite splendid to have a bird's eye view of KAP about blood donation and to plan for an improvement of the blood supply.

The data shows that the classic problem of missing and non-sense responses, which is common in all surveys, was present, but was minimal. The fact that oral interviews were conducted was a bonus to probe more deeply, but no doubt contributed to the missing data and inappropriate responses as interviewees might have been less reluctant at times to disclose personal information. Techniques used to handle missing data will be explored in later chapters and applied to this study.

2.2 Descriptive Report

Let π_{donor} be the proportion of the blood donors and similarly $\pi_{\text{non-donor}}$ be the proportion of non-donors in the entire population. The aim is to test the null hypothesis (H_0) that there is *no* difference between the proportion of blood donors and the proportion of non-donors of the population of the same characteristics. This hypothesis is tested against an alternative hypothesis H_1 that there is a difference between these two population proportions. In other words,

$$H_0: \pi_{\text{donor}} = \pi_{\text{non-donor}}$$

or similarly,

$$H_0: \pi_{\text{donor}} = 0.50$$

against

$$H_1: \pi_{\text{donor}} \neq \pi_{\text{non-donor}}$$

or

$$H_0: \pi_{\text{donor}} \neq 0.50.$$

The bases for the test are the sample proportions, P_{donor} and $P_{\text{non-donor}}$ and to basically test if the sample supports the claim that being a donor or a non-donor in the population of Mali with a particular characteristic is purely a 50-50 chance. The sample supports the claim if the p–value is less than 0.025. Otherwise reject H_0 at the 5% level of significance and conclude with 95% certainty that for the particular item in the survey tool under investigation, the population proportion of donors is different from 50%. Note that the test is valid if the sample size is large (large in the sense that sample size*proportion > 5). In other words, if the sample size in a particular category is small, then making an inferential statement is not advisable and no conclusions will be drawn for that particular category.

To further conclude whether being donors are more prevalent than non-donors amongst those that responded in a particular way to an item in the survey tool, the alternative hypothesis would be

$$H_1: \pi_{donor} > \pi_{non-donor}$$

or

$$H_1: \pi_{donor} > 0.50.$$

Similarly for testing whether donors are less prevalent than non-donors in the study group, the test would be

$$H_1: \pi_{donor} < \pi_{non-donor}$$

or

$$H_1: \pi_{donor} < 0.50.$$

Data collected from the Mali population reveal that 47% of individuals responded as donors whilst 53% responded as non-donors, i.e., more than 50% of individuals responded as non-donors (Figure 2.1). Donor categories are presented in Figure 2.2, and it can be observed that approximately 19% of individuals responded as family replacement donors, about 43% as lapsed donors (i.e., donation before 2012), 22% as voluntary non-remunerated blood donors and approximately 16% as a regular donor for several years. Also, from Figure 2.3 it can be observed

that more than 50% of non-donors (i.e., approximately 90%) reported their intention to donate blood in the future whilst approximately 10% reported that they had no intention of becoming future blood donors.

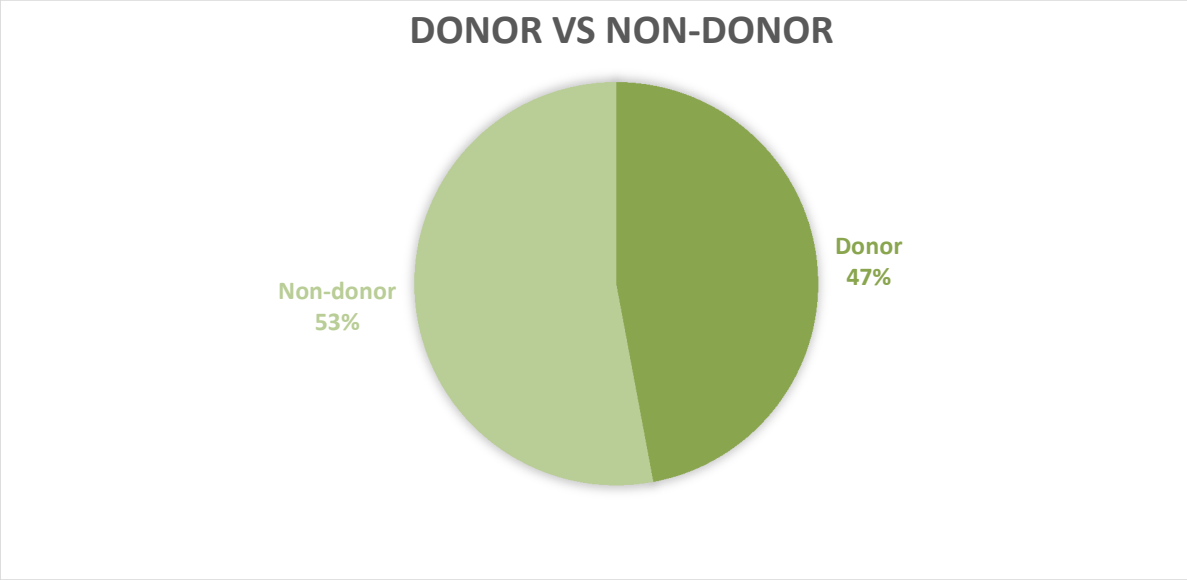


Figure 2.1: Donor vs Non-Donor

The KAP study implemented was really extensive with seven sections to the questionnaire and therefore only a few selective variables will be described and presented hereafter. It is necessary to point out that questions in Section 2 of the survey tool were only posed to donors, with the two study groups being compared for this section, being current donors and previous donors. Questions in Section 3 of the survey tool were only posed to non-donors, with the two study groups being compared for this section, being non-donors that intend to donate blood in future and those that do not intend to ever become donors in future. Note that there will be no tables with p-values for indicating the percentage of donors as compared to non-donors for these sections as questions were separately asked to donors and non-donors respectively.

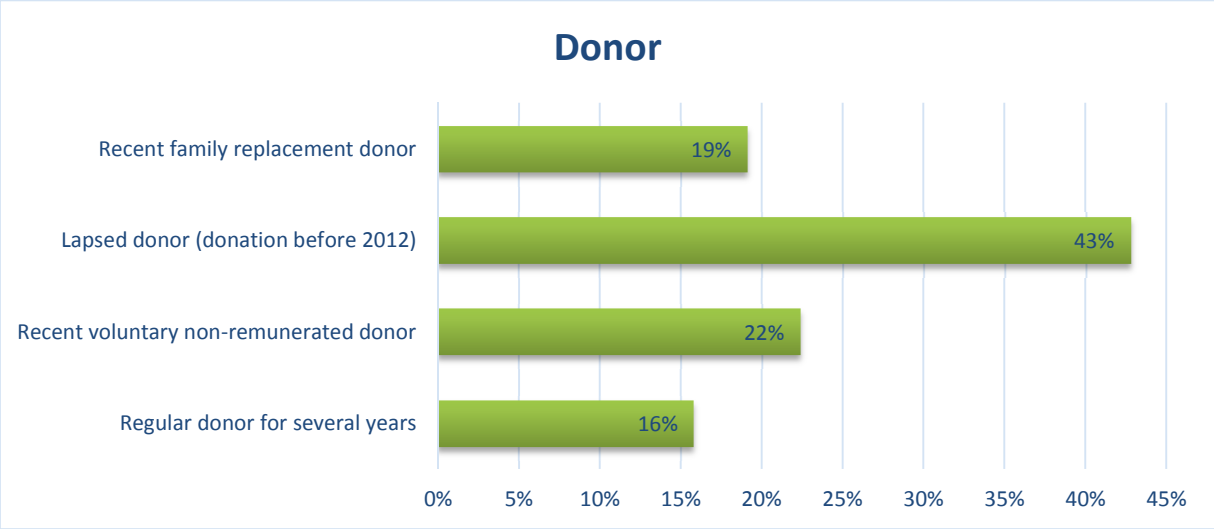


Figure 2.2: Donor categories

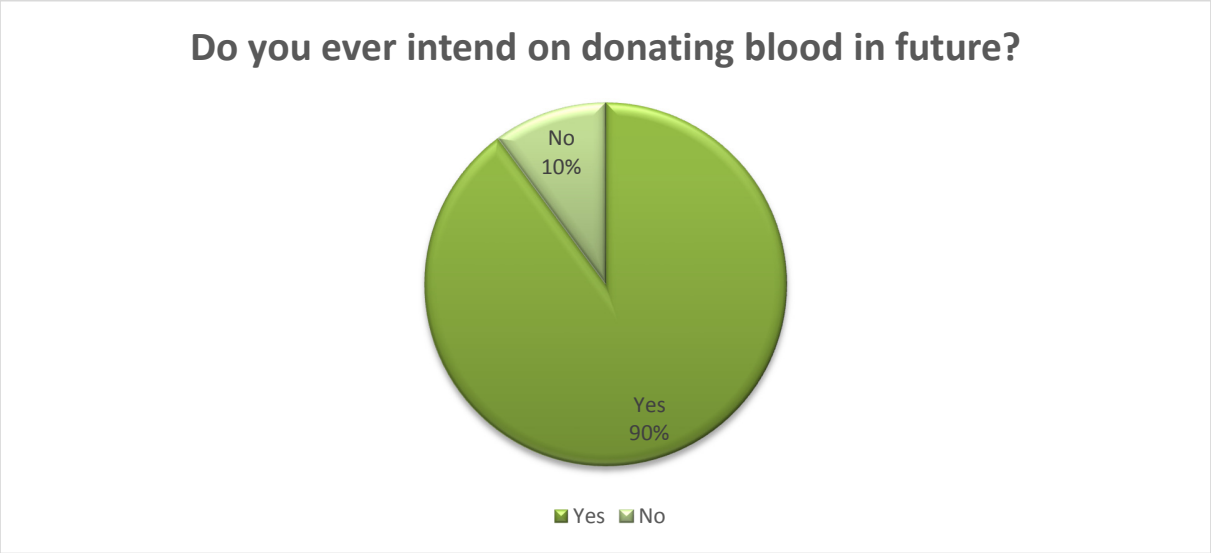


Figure 2.3: Non-donor intention on future blood donation

With regard to age and donor status, for individuals younger than 18 years old, the proportion of donors is significantly different from 50%. Equivalently one may conclude that for the age group younger than 18 years, the chance of being a donor or a non-donor is not 50-50. From Table 2.1 it can be seen that P_{donor} = 0.368 with p-value = 0.004 supports that the youngsters are less likely to be blood donors. Also, it has been found that about 36% of the young age group (<18 years old) responded that they were donors, which is statistically different from the speculated national proportion of 50% (p-value 0.004), indicating that there is a difference between the proportions of donors and non-donors for the age group 18 years or younger.

In terms of blood donation and gender, female respondents were less likely to be donors. The male population responded as majority donors at 58.8%, i.e. of all people in the study who were male, a proportion of 58.8% indicated that they were donors, indicating that across all three sites there is a significant difference between males being a donor as compared to a non-donor amongst males. A similar result for females, indicating that there are significant differences between the proportion of donors as compared to non-donors and this is true for both genders.

As seen in Table 2.1, in terms of a respondent having received blood, those respondents who said no to having ever received blood before, a proportion of 45.2% indicated that they were donors. Also, in Table 2.1 it can be observed that the majority of respondents appear to have knowledge about the different blood groups. It can thus be concluded that respondents that had knowledge about the different types of blood groups were more inclined to be donors.

In terms of blood donation and the possibility of contracting an infection by receiving blood, the proportion of donors that responded to the different questions' options are given in Table 2.1. It is observed that the majority of blood donors reported that a person can be infected with a disease by receiving blood, whilst a small percentage stated otherwise. Nationwide it is observed that about 45% of respondents that stated a person can be infected with a disease by receiving blood, responded as donors. However, this is not statistically different from the national speculated proportion of 50% (p-value 0.104), indicating that there is no significant difference between donors and non-donors with regard to a person getting infected with a disease by receiving blood.

Table 2.1: Table of association between characteristics and donor status.

		Total	Donors		P-value
			Number	Percentage	
Age	18	95	35	36.8%	0.004
	19 – 25	76	42	55.3%	0.178
	26 – 30	76	42	55.3%	0.178
	31 – 40	35	17	48.6%	0.433
	41 – 50	17	6	35.3%	0.102
	51 – 60	3	2	66.7%	0.270
Gender	Female	128	37	28.9%	0.000
	Male	194	114	58.8%	0.007
Have you received blood (RB)	Yes	15	10	66.7%	0.085
	No	290	131	45.2%	0.049
Do you know the different blood groups (KDBG)?	Yes	218	122	56.0%	0.038
	No	103	30	29.1%	0.000
Can a person get infected with a disease by receiving blood?	Yes	200	91	45.5%	0.101
	No	92	51	55.4%	0.147
	Do not know	23	6	26.1%	0.005
What do you think about blood donation?					
It is a good practice	Yes	305	146	47.9%	0.228
	No	16	5	31.3%	0.053
It is a dangerous process	Yes	3	1	33.3%	0.270
	No	318	150	47.2%	0.156
I have no strong feeling	Yes	2	1	50.0%	0.500
	No	319	150	47.0%	0.143
It is important and everyone should donate	Yes	95	49	51.6%	0.379
	No	226	102	45.1%	0.071

		Total	Donors		P-value
			Number	Percentage	
Other	Yes	7	5	71.4%	0.105
	No	313	146	46.6%	0.117
Can something bad happen to a person who donates blood?	Yes	140	65	46.4%	0.198
	No	151	77	51.0%	0.404
	Do not know	26	6	23.1%	0.001
Do people who donate blood receive something in return?	Yes	110	58	52.7%	0.283
	Yes in some cases	11	8	72.7%	0.045
	No	137	77	56.2%	0.072
	Do not know	59	6	10.2%	0.000
Have you ever seen or heard messages about blood donation?(HSMBD)	Yes	280	139	49.6%	0.452
	No	29	9	31.0%	0.014
	I cannot remember	9	2	22.2%	0.023

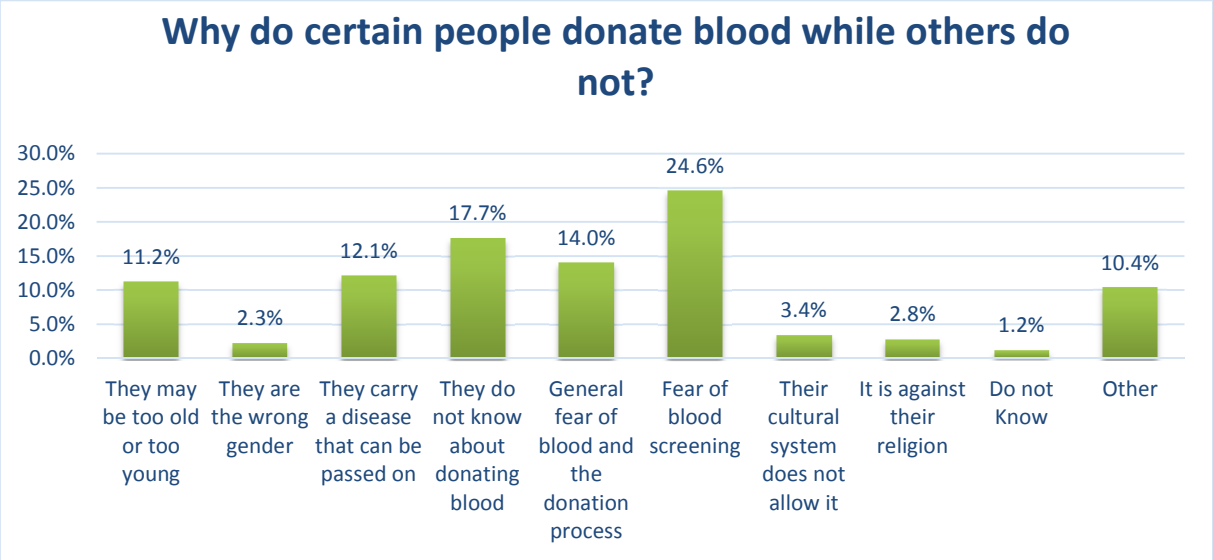


Figure 2.4: Reasons for some people donating blood while others do not.

With regard to blood donation and a respondent's opinion on blood donation, it can be noted that about 47% of those that thought that blood donation is a good practice, and 51% of respondents that thought that it is important and everyone should donate blood, were donors.

In terms of blood donation and the reasons given when asked why some individuals donate blood, while others do not, refer to Figure 2.4, it can be observed that the majority of respondents reported that the fear of blood screening is the reason that some individuals do not donate blood while others do. This was closely followed by respondents noting that they did not know about donating blood.

When exploring blood donation and an appropriate way to give blood, it is observed from Figure 2.5, that the majority of respondents reported that voluntary non-remunerated (unpaid) donation is the appropriate way to give blood. While approximately 19% of respondents reported that paid donation was the appropriate way to give blood. Respondents were asked about blood donation and whether something bad could happen to a person that donates blood, results are summarized in Table 2.1, and it can be noted that the majority of respondents in the Mali population reported that people donating blood feel weak, but this was not a significant effect.

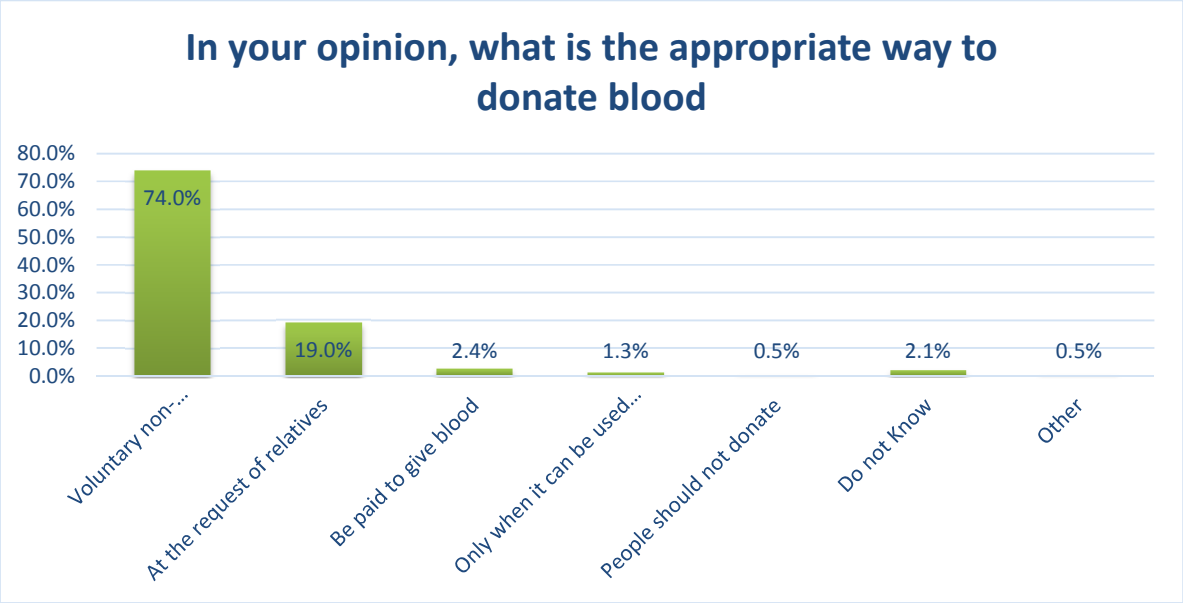


Figure 2.5: Opinion on the appropriate way to donate blood

When asked whether respondents had seen or heard messages about blood donation, it is observed that the majority of respondents in all three sites reported to have heard/seen messages on blood donation. From Table 2.1 it can be observed that of those that had seen and heard messages about blood donation, 49.6% were donors.

Only variables applicable to the outcome variable will be selected for further inferences and included in the analyses for later chapters. These variables were chosen with regard to literature review and led to collapsing and merging of variables which will be discussed in Section 3.9.

Chapter 3: Logistic Regression

3.1 The Logistic Regression Model

The logistic regression model is widely used to fit a categorical dependent variable that is dichotomous or binary, while the dependent variables can be either categorical or continuous. The primary distinction between the logistic regression model and linear regression model is that the outcome variable in logistic regression is binary or dichotomous. However, logistic regression is not limited to a simply dichotomy dependent variable, and can be generalized to dependent variables that have more than two categories that could be ordered or unordered.

Der & Everitt (2002) described modeling the expected value of the response variable Y in linear regression, as a linear function of the explanatory variables:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p \, . \tag{3.1}$$

They argue that there are two problems with using the above linear regression model when the outcome or response variable is dichotomous. Firstly, the expected value, which is simply the predicted probability, denoted by π, must satisfy $0 \leq \pi \leq 1$, while there is no limit for the linear predictor which can yield any value from $-\infty$ to $+\infty$.

Also, binary data does not follow a normal distribution but instead a Bernoulli or binomial distribution with the probability of a success given by π and the probability of a failure given by $1 - \pi$. Hence, if Y follows a Bernoulli distribution with probability of success $P(Y = 1) = \pi$, the probability function of Y is

$$f(y) = \exp\left\{y \ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right\},$$

(Molenberghs & Verbeke, 2005). Hence, the Bernoulli distribution belongs to the exponential family, with natural parameter θ equal to the logit, i.e., $\ln[\pi/(1-\pi)]$, of π, scale parameter $\phi = 1$, with mean $\mathrm{E}(Y) = \pi$ and $\mathrm{var}(Y) = \pi(1-\pi)$.

Since estimation is made for the unknown probability π for any given linear combination of the independent variables, a function is needed to link together the independent variables to π. The associated linear model (3.1) can then be generalized to

$$\begin{aligned} g\{E(Y_i)\} &= g(\pi) \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p, \end{aligned} \tag{3.2}$$

or simply $g(\pi) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p$ for some function $g(\cdot)$. Since it links the random and systematic components of the linear model, g is known as the link function (McCullagh & Nelder, 1989). In this case the link is called a logit and it follows

$$\mathrm{g}(\pi) = \mathrm{Logit}(\pi) = \log\left[\frac{\pi}{1-\pi}\right],$$

where the logit of the probability π is basically the log of the odds of the event of interest.

Once the dichotomous outcome is transformed by the logit link, it can be seen that the logistic regression model is essentially just a standard linear regression model. The transformation changes the range of the probability π from 0 to 1 to $-\infty$ to $+\infty$.

Der & Everitt (2002) suggest setting $\boldsymbol{\beta}' = [\beta_0, \beta_1, \ldots, \beta_p]$ and the augmented vector of scores for the ith individual as $\mathbf{x}_i' = [1, x_{i1}, x_{i2}, \ldots, x_{ip}]$, and so it follows that the predicted probabilities as a function of the linear predictor are:

$$\pi(\boldsymbol{\beta}'\boldsymbol{x_i}) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x_i})}{1 + \exp(\boldsymbol{\beta}'\mathbf{x_i})}.$$

This probability always satisfies $0 \leq \pi(\boldsymbol{\beta}'\mathbf{x_i}) \leq 1$.

The logit link function is a special case of generalized linear modelling and can sometimes be replaced by the probit link or complementary log-log link, which will be discussed in later sections.

3.2 Odds Ratio

Suppose the probability of a success is π, then the odds can be defined as

$$\Omega = \frac{\pi}{(1-\pi)},$$

(Agresti, 2007).

In its simplest form, this ratio can be interpreted as the ratio of the probability of occurrence of an event to the probability of the event not occurring. If $\Omega > 1$, then a success is more likely than a failure.

With reference to a 2×2 table, Hosmer & Lemshow (2000) explain that the odds of the outcome being present among individuals with $x = 1$ is defined as

$$\frac{\pi(1)}{[1-\pi(1)]},$$

and the odds of an outcome being present among individuals with $x = 0$ is defined as

$$\frac{\pi(0)}{[1-\pi(0)]}.$$

The ratio of the odds for $x = 1$ to the odds for $x = 0$ is called the odds ratio and is given by

$$\mathrm{OR} = \frac{\frac{\pi(1)}{[1-\pi(1)]}}{\frac{\pi(0)}{[1-\pi(0)]}}.$$

An odds ratio is the ratio of the probability that some event will occur over the probability that the same event will not occur (Kleinbaum & Klein, 2002).

Now consider the relationship between an outcome variable and one explanatory variable, X

$$\text{logit}(\pi) = \beta_0 + \beta_1 X_1,$$

where π is the probability of the occurrence of an event $(\text{Y} = 1)$. The logit function on the left is defined as

$$\text{logit}(\pi) = \log \frac{\pi}{1-\pi} = \log \text{odds} \rightarrow \pi = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}.$$

The odds ratio can then be simplified as follows

$$\text{OR}(X = 1 \text{ vs } X = 0) = \frac{\frac{e^{\beta_0+\beta_1}/(1+e^{\beta_0+\beta_1})}{1/(1+e^{\beta_0+\beta_1})}}{\frac{e^{\beta_0}/(1+e^{\beta_0})}{1/(1+e^{\beta_0})}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

Hence a single unit increase in X will change the odds of observing $(X = 1)$ versus $(X = 0)$ by a multiplicative factor of e^{β_1}. The odds ratio can therefore be interpreted as the effect of a single unit of change in X in the predicted odds ratio with the other variables in the model held constant, which can be generalized to include any predictor variable X, as follows:

$$\frac{P(\pi|X+1)/1-P(\pi|X+1)}{P(\pi|X)/1-P(\pi|X)}.$$

Hosmer & Lemeshow (2002) have reported that a $100 \times (1-\propto)\%$ confidence interval estimate for the odds ratio can be obtained by calculating the endpoints of a confidence interval of the coefficient, β, and then computing the exponentials of these values.

In general, the endpoints are given as follows:

$$\exp\left[\hat{\beta} \pm z_{1-\alpha/2} \times \widehat{\text{SE}}(\hat{\beta})\right].$$

The foundation of the interpretation for all logistic regression results is provided by the relationship between the logistic regression coefficient and the odds ratio.

3.3 Parameter Estimation

The maximum likelihood (ML) method is used to obtain estimates from the logistic regression model. According to McCulloch et al. (2008), since the y_i are independent and Bernoulli distributed, the likelihood can be evaluated as follows

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

$$= \prod_{i=1}^{n} \{\pi(x_i)/[1 - \pi(x_i)]\}^{y_i} [1 - \pi(x_i)]. \tag{3.3.1}$$

By the use of

$$\pi(x_i)/[1 - \pi(x_i)] = e^{\alpha+\beta x_i},$$

and

$$1 - \pi(x_i) = \left(1 + e^{\alpha+\beta x_i}\right)^{-1},$$

L is given as

$$L = \prod_{i=1}^{n} e^{y_i(\alpha+\beta x_i)} \left(1 + e^{\alpha+\beta x_i}\right)^{-1},$$

and the log likelihood as

$$l = \log L = \sum_{i=1}^{n} \left[y_i(\alpha + \beta x_i) - \log\left(1 + e^{\alpha+\beta x_i}\right)\right].$$

Differentiating the log likelihood l with respect to α and β gives

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} \left[y_i - \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}\right]$$

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

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

and

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \left[x_i y_i - \frac{x_i e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right]$$

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

where, $\pi(x_i) = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} = \frac{1}{1+e^{-(\alpha+\beta x_i)}}$.

Once equating the two derivatives to zero, the following equations need to be solved

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{1}{1 + e^{-(\hat{\alpha}+\hat{\beta} x_i)}}$$

$$\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} \frac{x_i}{1 + e^{-(\hat{\alpha}+\hat{\beta} x_i)}},$$

where and $\hat{\alpha}$ and $\hat{\beta}$ are the maximum likelihood estimators (MLEs) of parameters α and β.

To get the estimators of α and β, McCulloch et al. (2008) interpret the first equation as follows: the ML solutions are chosen so that the total predicted number of successes is equal to $\sum_i y_i$, which is the total observed number of successes. They have also showed that the second derivatives of l take on the form:

$$\frac{\partial^2 l}{\partial \alpha^2} = -\sum_{i=1}^{n} \frac{\partial \pi(x_i)}{\partial \alpha}$$

$$= -\sum_{i=1}^{n} \frac{e^{-(\alpha+\beta x_i)}}{(1 + e^{-(\alpha+\beta x_i)})^2},$$

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

(3.3.2)

$$\frac{\partial^2 l}{\partial \alpha \partial \beta} = -\sum_{i=1}^{n} \frac{x_i e^{-(\alpha+\beta x_i)}}{(1+e^{-(\alpha+\beta x_i)})^2},$$

$$\frac{\partial^2 l}{\partial \beta^2} = -\sum_{i=1}^{n} x_i^2 \pi(x_i)\,[1-\pi(x_i)].$$

(3.3.3)

For the multivariate case, estimates need to be obtained for the vector $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$. The likelihood function is nearly identical to that given in equation (3.3.1) with the only change being that $\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1+e^{g(\mathbf{x})}}$. There will be $(p+1)$ likelihood equations obtained by differentiating the log likelihood function with respect to the $p+1$ coefficients. The likelihood equations may be expressed as

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

and

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0,$$

for $j = 1,2, \ldots, p$.

The solution $(\widehat{\boldsymbol{\beta}})$ to the likelihood equations require special iterative procedures or techniques. Hosmer & Lemeshow (2000) discuss a method to estimate the variances and covariances of the estimated coefficients which involves obtaining the estimators from the matrix of second partial derivatives of the log likelihood function. The partial derivatives for the univariate case can be found in equations (3.3.2) and (3.3.3) whilst the partial derivatives for the multivariate case are of the form

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i)$$

(3.3.4)

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i),$$

(3.3.5)

for $j = 1,2, \ldots, p$ and where π_i denotes $\pi(\mathbf{x}_i)$. According to Hosmer & Lemeshow (2000), let the $(p+1) \times (p+1)$ matrix containing the negative of the terms given in equations (3.3.4) and (3.3.5) be denoted as $\mathbf{I}(\boldsymbol{\beta})$. This matrix is known as the observed information matrix. The variances and covariances of the estimated coefficients can be obtained from the inverse of the observed information matrix which is denoted as $\mathrm{Var}(\boldsymbol{\beta}) = \mathbf{I}^{-1}(\beta)$. Hosmer & Lemeshow (2000) confer that except in very special cases, the explicit expression for the elements in this matrix is not possible to write down. Therefore $\mathrm{Var}(\beta_j)$ will be used to denote the j^{th} diagonal element of the matrix, which is the variance of $\hat{\beta}_j$ and $\mathrm{Cov}(\beta_j, \beta_l)$ to denote and arbitrary off-diagonal element, which is the covariance of $\hat{\beta}_j$ and $\hat{\beta}_l$. The estimators of the variances and covariances are obtained by evaluating $\mathrm{Var}(\boldsymbol{\beta})$at $(\widehat{\boldsymbol{\beta}})$ and is denoted by $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})$. The $\widehat{\mathrm{Var}}(\hat{\beta}_j)$ and $\widehat{\mathrm{Cov}}(\hat{\beta}_j, \hat{\beta}_l), j, l = 1,2, \ldots, p$ denotes the values in this matrix.

The Information matrix is $\hat{\boldsymbol{I}}(\widehat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{V}\mathbf{X}$ where $\mathbf{X}$ is an $n \times (p+1)$ matrix containing the data for each subject as follows,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} \cdots & x_{np} \end{bmatrix},$$

and the matrix $\boldsymbol{V}$ is an $n \times n$ diagonal matrix with general element $\hat{\pi}_i(1 - \hat{\pi}_i)$.

$$\widehat{\boldsymbol{V}} = \begin{bmatrix} \hat{\pi}_{x_1}(1 - \hat{\pi}_{x_1}) & 0 \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots 0 & \hat{\pi}_{x_n}(1 - \hat{\pi}_{x_n}) \end{bmatrix}.$$

For the univariate case, the observed information matrix is given as

$$I = -E\begin{bmatrix} \dfrac{\partial^2 l}{\partial^2 \alpha} & \dfrac{\partial^2 l}{\partial \alpha \partial \beta} \\ \dfrac{\partial^2 l}{\partial \alpha \partial \beta} & \dfrac{\partial^2 l}{\partial^2 \beta} \end{bmatrix} = -\mathrm{E}\begin{bmatrix} l_{\alpha\alpha} & l_{\alpha\beta} \\ l_{\beta\alpha} & l_{\beta\beta} \end{bmatrix} = \mathrm{E}[\mathbf{X}'\mathbf{V}\mathbf{X}] = \mathbf{X}'\mathbf{V}\mathbf{X},$$

It then follows that the variance-covariance matrix of $\hat{\alpha}$ and $\hat{\beta}$ in the univariate case is $(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$, which is the inverse of the information matrix.

Since the Hessian, which is a square matrix of second-order partial derivatives of a scalar-valued function, is negative definite, the log likelihood is concave and the log likelihood function is maximized numerically using iterative procedures.

McCulloch et al. (2008) confer that large-sample tests and confidence intervals can be based on the asymptotic normality (AN) of for example, $\hat{\alpha}$ and $\widehat{\beta}$, as follows

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim AN\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\right].$$

To test for example,

$$H_0\colon \beta > 0 \text{ versus } H_\mathrm{A}\colon \beta \geq 0,$$

reject H_0 if

$$\frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{var}}(\hat{\beta})}} > z_\alpha,$$

where z_α is the $100\alpha\%$ percentile of the standard normal distribution, which is, if $Z \sim N(0,1)$, then $P\{Z > z_\alpha\} = \alpha$ and $\widehat{\mathrm{var}}(\hat{\beta})$ comes from inserting the MLEs into the lower-right-hand entry of $(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$.

The large-sample confidence interval for β is as follows

$$\hat{\beta} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\beta})}.$$

Also, a large-sample confidence interval for the odds ratio, e^{β}, would be calculated as

$$\left(e^{\hat{\beta}-z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\beta})}}, e^{\hat{\beta}+z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\beta})}}\right).$$

Alternatively, the likelihood ratio test can be used to test the two sided hypothesis

$$H_0: \beta = 0 \text{ versus } H_A: \beta \neq 0.$$

The likelihood under H_0 becomes

$$L = \prod_{i=1}^{n} e^{y_i \alpha}(1 + e^{\alpha}),$$

with maximum $\hat{\alpha}_0 = \log[\bar{y}/1 - \bar{y}]$. The maximized value of $l = \log L$ under H_0 is

$\sum y_i \log\bar{y} + \sum(1 - y_i)\log(1 - \bar{y})$. The likelihood ratio statistic is then given by

$$-2\log\Lambda = -2\left[\sum y_i \log\bar{y} + \sum(1 - y_i)\log(1 - \bar{y}) - \sum y_i\left(\hat{\alpha} + \hat{\beta}x_i\right) + \sum \log\left(1 + e^{\hat{\alpha}+\hat{\beta}x_i}\right)\right],$$

and H_0 is rejected whenever $-2\log\Lambda$ exceeds the chi-square distribution with critical point, $X^2_{1,1-\alpha}$.

3.4 Goodness of Fit Test

To assess the fit of an estimated logistic model, goodness-of-fit test statistics are used. These tests involve investigating how close predicted values are to the observed values in the model. In logistic regression there are a number of different possible ways to assess the difference between the observed values and the fitted values.

3.4.1 Pearson Chi-Square Statistic and Deviance

Suppose the fitted model contains p independent variables, $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ and let J denote the number of distinct values that $\mathbf{x}$ observed. If some subjects have the same value $\boldsymbol{x}$ then

consider $J < n$. Denote the number of subjects with $\mathbf{x} = \mathbf{x}_j$ by m_j $j = 1,2,, \ldots, J$. It then follows that $\sum_{j=1}^{J} m_j = n$. Further, let y_j denote the number of positive responses, $y = 1$, among m_j subjects with $\mathbf{x} = \mathbf{x}_j$. Denote by $\sum_{j=1}^{J} y_j = n_1$, the total number of subjects with $y = 1$. To emphasize that the fitted values in logistic regression are calculated for each covariate pattern and depend on the estimated probability for that covariate pattern, Hosmer & Lemeshow (2002) suggest denoting the fitted value for the *j*th covariate pattern as $\hat{y}_j$ where

$$\hat{y}_j = m_j \hat{\pi}_j = m_j\left(\exp\left[\hat{g}(\mathbf{x}_j)/\{1 + \exp\left[\hat{g}(\mathbf{x}_j)\right]\}\right]\right),$$

where $\hat{g}(\mathbf{x}_j)$ is the estimated logit.

Hosmer & Lemeshow (2002) further suggest considering two measures of the difference between the observed and the fitted values: the Pearson residual and the deviance residual. The Pearson residual for a particular covariate pattern can be defined as

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}}. \tag{3.4.1}$$

The summary statistic based on these residuals is the Pearson chi-square statistic given by

$$X^2 = \sum_{j=1}^{J} r\left(y_j, \hat{\pi}_j\right)^2. \tag{3.4.2}$$

The deviance residual can be defined as

$$d(y_j, \hat{\pi}_j) = \pm\left\{2\left[y_j \ln\left(\frac{y_j}{m_j \hat{\pi}_j}\right) + (m_j - y_j)\ln\left(\frac{(m_j - y_j)}{m_j(1 - \hat{\pi}_j)}\right)\right]\right\}^{1/2}. \tag{3.4.3}$$

The deviance residual for covariate patterns with $y_j = 0$ is

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j \left|\ln(1-\hat{\pi}_j)\right|}\,.$$

The deviance residual when $y_j = m_j$, is as follows

$$d(y_j, \hat{\pi}_j) = \sqrt{2m_j \left|\ln(\hat{\pi}_j)\right|}\,.$$

The summary statistic based on the deviance residuals is the deviance

$$D = \sum_{j=1}^{J} d\left(y_j, \hat{\pi}_j\right)^2 \,. \tag{3.4.4}$$

If the model is correct, the statistics X^2 and D have approximately a chi-square distribution with degrees of freedom equal to $J-(p+1)$. Further, for the deviance it follows that D is the likelihood ratio test statistic of a saturated model with J parameters versus the fitted model with $p+1$ parameters.

3.4.2 The Hosmer-Lemeshow Test

Hosmer and Lemeshow's goodness-of-fit test is another method commonly used to assess the fit of a model. Hosmer and Lemeshow (1980) and Lemeshow and Hosmer (1982) suggested grouping based on the values of the estimated probabilities. The idea behind the test is that the predicted and observed probabilities should match closely and that the closer they match, the better the fit. Groups are created using predicted probabilities, and then observed and fitted counts of successes and failures are compared on those groups using a chi-squared statistic.

According to Hosmer & Lemeshow (2002), suppose $J = n$ and think of the n columns as corresponding to the n values of the estimated probabilities, with the first column corresponding to the smallest value and the *n*th column to the largest value. The first grouping strategy proposed involves collapsing the table based on percentiles of the estimated probabilities whilst

the second strategy involves collapsing the table based on fixed values of the estimated probability. The first method uses $g = 10$ groups and results in the first group containing the $n'_1 = \frac{n}{10}$ subjects having the smallest estimated probabilities and the largest group containing $n'_{10} = \frac{n}{10}$ having the largest estimated probabilities.

The second method with $g = 10$ groups, results in cutoff points defined at the values $\frac{k}{10}$, $k = 1,2, \ldots, 9$, and the groups contain all subjects with estimated probabilities between adjacent cutoff points.

The first group contains all subjects whose estimated probabilities are less than and equal to 0.1 while the tenth group contains all subjects whose estimated probabilities is greater than 0.9.

For either grouping strategy the Hosmer -Lemeshow test statistic is defined as follows

$$H = \sum_{k=1}^{g} \frac{(O_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)},$$

where n'_k is the total number of subjects in the k^{th} group, c_k denotes the number of covariate patterns in the k^{th} decile,

$$O_k = \sum_{j=1}^{n_k} y_j$$

is the number of responses among the c_k covariate patterns and

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k}$$

is the average estimated probability.

Hosmer and Lemeshow (1989) reported simulations showing that the statistic has approximately a chi-square distribution under the null hypothesis that the model fitted is correct, with $g - 2$ degrees of freedom. A large H value, that is, an H value larger than the 100α percentage point

of the chi-square distribution (or a p-value less than α) indicates that the model is inadequate. Lee & Wang (2003) have reported that as with other chi-square goodness-of-fit tests, the approximation depends on the estimated frequencies being reasonably large and if a large number (say, more than 20%) of the expected frequencies are less than 5, the approximation may not be appropriate and the p-value should be interpreted with caution. If this is the case, the proposed solution is to combine adjacent groups to increase the estimated expected frequencies. However, Hosmer & Lemeshow warn that if fewer than six groups are used to calculate H, the test would be insensitive and would almost always indicate that the model is adequate (Lee & Wang, 2003).

3.4.3 Area under the ROC Curve

Agresti (2007) has reported that the accuracy of a diagnostic test is often assessed with two conditional probabilities, namely, sensitivity and specificity. Chen et al (2008), define sensitivity as a measure of accuracy for event prediction:

$$\text{Sensitivity} = \text{P}\big(\hat{y} = 1 \,\big|\, y = 1\big),$$

where "$\hat{y} = 1$" is the number of default individuals who screen the same, "$y = 1$" is the total number of default individuals.

$$1 - \text{Specificity} = 1 - \text{P}\big(\hat{y} = 0 \,\big|\, y = 0\big),$$

where "$\hat{y} = 0$"is the number of normal-performing individuals who screen the same; "$y = 0$" is the total number of normal-performing individuals.

In other words, sensitivity is the probability of correctly classifying an observation with an outcome of an event and specificity is the probability of correctly classifying an observation with the outcome of a nonevent. Also, the positive predictive value (PPV) is the proportion of observations classified as events that are correctly classified and the negative predictive value (NPV) is the proportion of observations classified as nonevents that are correctly classified.

Sensitivity and specificity rely on a single cutoff point to classify a test result as positive (Hosmer & Lemeshow, 2002). The area under the ROC (Receiver Operating Characteristic) curve

originated from a single detection theory and shows how the receiver operates the existence of signal in the presence of noise.

Hosmer & Lemeshow (2002) explain that it plots the probability of detecting true signal (sensitivity) and false signal (1 – specificity) for an entire range of cutoff points. The area under the ROC curve ranges from zero to one and provides a measure of the models ability to discriminate between subjects that experience the outcome of interest against those subjects that do not.

Hosmer & Lemeshow (2002) propose as a general rule:

If $\mathrm{ROC} = 0.5$: this suggests no discrimination.
If $0.7 \leq \mathrm{ROC} \leq 0.8$: this is considered acceptable discrimination.
If $0.8 \leq \mathrm{ROC} \leq 0.9$: this is considered excellent discrimination.
If $\mathrm{ROC} \geq 0.9$: this is considered outstanding discrimination.

Agresti (2007) has reported that for a given specificity, better predictive power corresponds to higher sensitivity therefore the better the predictive power, the higher the ROC curve. In essence, the higher the area under the curve the better the prediction power of the model.

3.4.4 Information Criteria

The Akaike's information criteria (AIC) and the Schwartz criterion (SC) can be used to test the goodness of fit of two nested models. These methods are used to adjust the likelihood ratio statistic which measures the deviation of the log-likelihood of the fitted model from the log-likelihood of the maximal possible model (Vittinghoff et al., 2005). The AIC was introduced by (Akaike, 1974) and judges a model by how close its fitted values tend to be to the true expected values, as summarized by a certain expected distance between the two. Agresti (2007) reports that the optimal model is the one that tends to have its fitted values closest to the true outcome probabilities.

The AIC is calculated as:

$$\mathrm{AIC} = -2\mathrm{logL} + 2p \,,$$

where p is the number of parameters used in the model. Usually, the model with the smallest AIC value is the best (Agresti, 2002).

The SC introduced by (Schwatrz, 1978) is calculated as:

$$-2\log\mathrm{L} + p\log(n),$$

where p is the number of parameters and *n* is the size of the sample.

3.4.5 Measure of Association

Allison (1999) discusses the four measures of association that measure how well one can predict the dependent variable based on the independent variables, namely, Kendall's Tau-a, Goodman-Kruskal's Gamma, Somers's D statistic, and the c statistic. The idea behind these measures is to pair the observations in different ways without pairing an observation with itself. Pairs that have either both 1's on the dependent variable or both 0's are ignored while pairs in which one case has a 1 and the other case has a 0, are retained. For each pair, a pair is considered concordant if the case with a 1 has a higher predicted value (based on the model) than the case with a 0 and discordant otherwise. However, if the two cases have the same predicted value, it is then considered a tie.

Now, let C be the number of concordant pairs, D the number of discordant pairs, T the number of ties, and N the total number of pairs (before eliminating any). The four measures of association are then calculated as follows:

$$\text{Tau-a} = \frac{C - D}{N}.$$

$$\text{Gamma} = \frac{C - D}{C + D}.$$

$$\text{Somer's D} = \frac{C - D}{C + D + T}.$$

$$c = 0.5(1 + \text{Somer's D}).$$

The Tau-a statistic is Kendall's rank order correlation coefficient without the adjustments for ties and the Gamma statistic is based on Kendall's coefficient but with the adjustments for ties. Allison (1999) further details that the four measures of association described above vary between 0 and 1, with larger values corresponding to stronger associations between the predicted and observed values and Tau-a being closest to the generalized R^2.

3.5 Overdispersion

In the analysis of discrete data, overdispersion is a very crucial concept where overdispersion is generally described as the lack of fit of a model.

For a binomial response y_i the mean is given by $\mu_i = n_i\pi_i$ and the variance is given by $\mu_i(n_i - \mu_i)/n_i$. If the variance of the response $\mathrm{y_i}$ is greater than $\mu_\mathrm{i}(\mathrm{n_i} - \mu_\mathrm{i})/\mathrm{n_i}$, then it could be an indication of overdispersion in the data.

For a model to be identified correctly, the Pearson chi-square statistic and the deviance, when divided by their degrees of freedom, should be approximately equal to one. If their values are much larger than one, then the assumption of binomial variability may be invalid and the data is thought to exhibit overdispersion.

Allison (1999) argues that overdispersion has two possible causes:

- An incorrectly specified model where more interactions and/or nonlinearities are needed in the model, and
- Lack of independence of observations which can arise from heterogeneity that operates at group levels rather than individuals.

Overdispersion can be modeled by introducing the scale parameter ϕ into the variance function. There are two solutions for overdispersion:

- The data can be remodelled by imposing $\mathrm{var}(\mu) = \phi(1 - \mu)$ for the binomial distribution, or $\mathrm{var}(\mu) = \phi\mu$ for the Poisson distribution.

- If $\widehat{\phi}$ is different from 1, then the distribution of the data is neither binomial nor Poisson thus an alternative distribution can be used.

Generally, overdispersion may occur due to a lack of homogeneity in the data. This lack of homogeneity may occur between groups of individual or within individual observations (Olsson, 2002).

3.6 Logistic Regression Diagnostics

Once a logistic model has been fitted to the data, it is crucial to check if the assumed model is a valid model. The appropriateness of the model can be studied using diagnostic testing. To check the adequacy of the fitted model, the analysis of residuals and the identification of outliers need to be investigated. Sarkar et al. (2011) describe the three ways that an observation can be considered as unusual, namely, outliers, influence and leverage. They describe outliers as a set of observations whose values deviate from the expected range and produce extremely large residuals that may indicate a sample peculiarity. They also describe an observation as being influential if the deletion of that observation substantially changes the estimate of coefficients and argue that influence can be thought of as the product of leverage and outliers. Further defining leverage as a measure of how far an independent variable deviates from its mean. These leverage points can have an unusually large effect on the estimate of logistic regression coefficients (Cook, 1998).

Influence statistics can determine how much some feature of the model changes when a particular observation is deleted from the model fit. The larger the value is for each diagnostic, the greater the influence. Ideally it is expected that each observation should have equal influence on the model. The failure to detect such influential cases could have severe distortion on the validity of inferences drawn from the model.

Pregibon (1981) provided the theoretical framework that extended linear regression diagnostics to logistic regression. The residual vector and a projection matrix are used as building blocks for the identification of outlying and influential points for the logistic regression model. Consider a

setup where the fitted model contains p covariates and that they form J covariate patterns indexed by $j = 1,2, \dots, J$. In logistic regression the errors are binomial hence the error variance is a function of the conditional mean as follows that

$$\text{var}(Y_j|\mathbf{x}_j) = m_j E(Y_j|\mathbf{x}_j) \times [1 - E(Y_j|\mathbf{x}_j)] = m_j \pi(\mathbf{x}_j)[1 - \pi(\mathbf{x}_j)].$$

Beginning with residuals as defined equations (3.41) and $(3.4.3)$ which have been "divided" by estimates of their standard errors, and letting r_j and d_j denote the expressions given in these equations respectively, for covariate pattern $\mathbf{x}_j$. Each residual is divided by an approximate of its standard error hence it is expected that if the logistic regression model is correct, these quantities should have a mean approximately equal to zero and a variance approximately equal to one.

Let $\boldsymbol{X}$ denote the $J \times (p + 1)$ matrix containing the values for all J covariate patterns formed from the observed values of the p covariates, with the first column being the one to reflect the presence of an intercept in the model. Pregibon (1981) used the weighted least squares linear regression as a model to derive a linear approximation to the fitted values which yields a hat matrix for logistic regression as follows

$$\boldsymbol{H} = \boldsymbol{V}^{1/2}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{1/2}, \tag{3.6.1}$$

where $\boldsymbol{V}$ is a $J \times J$ diagonal matrix with general element $v_j = m_j \hat{\pi}(\mathbf{x}_j)[1 - \hat{\pi}(\mathbf{x}_j)]$.

Let h_j denote the jth diagonal element of the matrix $\boldsymbol{H}$ defined in equation $(3.6.1)$, then it follows that

$$h_j = m_j \hat{\pi}(\mathbf{x}_j)[1 - \hat{\pi}(\mathbf{x}_j)]\mathbf{x}_j'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\mathbf{x}_j' = v_j \times b_j, \tag{3.6.2}$$

where $b_j = \mathbf{x}_j'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}$, $\mathbf{x}_j' = (1, x_{1j}, x_{2j}, \dots, x_{pj})$ is a vector of covariate values and $\sum h_j = (p + 1)$ is the number of parameters in the model.

Hosmer & Lemeshow (1989) discuss the formulation and bounds of any diagonal element in the hat matrix and point out the importance of keeping this distinction in mind as diagnostic information is computed differently in various programs. For more information on this see Hosmer & Lemeshow (1989, p.151).

According to Hosmer & Lemeshow (1989) consider the residual for the jth covariate pattern as $y_j - m_j\hat{\pi}(\mathbf{x}_j) \approx (1 - h_j)y_j$, then the variance of the residual is

$$m_j\hat{\pi}(\mathbf{x}_j)\left(1 - \hat{\pi}(\mathbf{x}_j)\right)(1 - h_j)^2,$$

where $(1 - h_j)^2 \approx (1 - h_j)$ for small h_j. This suggests that the Pearson residuals will not have variance equal to 1 unless they are further standardized. Recalling that r_j denotes the Pearson residual given in equation $(3.4.1)$, the standardized Pearson residual for covariate pattern $\mathbf{x}_j$ is

$$r_{sj} = r_j / \sqrt{1 - h_j}.$$

Hosmer & Lemeshow (1989) further discuss another useful diagnostic statistic, which is one that examines the effect that deleting all subjects with a particular covariate pattern has on the value of the estimated coefficients and the overall summary measures of fit, X^2 and D. They argue that the change in the value of the estimated coefficients is analogous to the measure proposed by Cook (1977, 1979) for linear regression. This is obtained as the standardized difference between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(-j)}$, which represent the respective maximum likelihood estimates computed using all J covariate patterns, excluding the m_j subjects with pattern $\mathbf{x}_j$, thus standardized via the covariance matrix of $\widehat{\boldsymbol{\beta}}$. Pregibon (1981) has shown that, to a linear approximation, this quantity for logistic regression is

$$\begin{aligned}\Delta\widehat{\boldsymbol{\beta}}_j &= \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-j)}\right)'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-j)}\right) \\ &= \frac{r_j^2 h_j}{(1 - h_j)^2} \\ &= r_{sj}^2 h_j.\end{aligned}$$

Similar linear approximation can be used to show that the decrease in the value of the Pearson chi-square statistic due to deletion of the subjects with covariate pattern $\mathbf{x}_j$ is

$$\Delta X_j^2 = \frac{r_j^2}{(1-h_j)}$$

$$= r_{sj}^2.$$

(3.6.3)

A similar quantity for the change in deviance may be obtained,

$$\Delta D_j = d_j^2 + \frac{r_j^2 h_j}{(1-h_j)}.$$

If r_j^2 is replaced by d_j^2, it yields the following approximation

$$\Delta D_j = \frac{d_j^2}{(1-h_j)},$$

which is similar in form to the expression in equation (3.6.3).

Hosmer & Lemeshow (1989) argue that these diagnostic statistics are conceptually appealing, as they allow one to identify the covariate patterns that are poorly fit (large values of ΔX_j^2 and/or ΔD_j), and also those that have great influence on the values of the estimated parameters (large values of $\Delta\widehat{\boldsymbol{\beta}}_j$).

3.7 Probit and Complementary log-log Models

The logistic regression model described in Section 3.1 is not the only approach available for modelling a dichotomous outcome. Under the assumption of a binary response, two other alternative methods exist, namely, probit model and complementary log log model.

3.7.1 Probit Regression Model

As mentioned, the probit model can also be used to model binary response data. The logit model makes use of the cumulative logistic function whereas the probit model uses a cumulative standard normal distribution functional form instead.

Let y_j be the i^{th} observation of a binary response variables Y_i with probability of success π_i, that is Y_i follows a Bernoulli distribution with parameter π_i for $i = 1,2,\dots,n$. The probit model is given by

$$\pi_i = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}),$$

where $\boldsymbol{x}_i'$ denotes the ith row of a matrix of predictors and $\Phi(\cdot)$ is the standard normal cumulative distribution function (McCulloch 2008). The above equation can be rewritten in vector form as

$$\boldsymbol{\pi} = \Phi(\mathbf{X}\boldsymbol{\beta}),$$

or equivalently

$$\Phi^{-1}(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{X}$ is the model matrix. As probabilities range between 0 and 1, the probit function ranges between $-\infty$ and $+\infty$.

Finney (1952) suggested calculating the estimate of $\boldsymbol{\beta}$ using an iteratively least squares algorithm by working probits which he defined as follows:

$$t_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \frac{y_i - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})}{\phi(\boldsymbol{x}_i'\boldsymbol{\beta})},$$

where $\phi(\cdot)$ is the standard normal probability density function (p.d.f). For a current value of $\boldsymbol{\beta}$, the working probits were regressed on the predictors using weights given by $\frac{\phi(\pi_i)^2}{\Phi(\pi_i)[1-\Phi(\pi_i)]}$ so as to get the new value of $\boldsymbol{\beta}$. The algorithm is iterated until convergence or until the estimates of $\boldsymbol{\beta}$ stabilize.

The probit model is generally estimated by maximum likelihood as in the case with the logit model. The chief difference between the two models is that the logistic model has slightly flatter tails. What this actually means is that the normal or probit curve approaches the axes more quickly than the logistic curve. Also, qualitatively both models give similar results.

3.7.2 The Complementary Log-Log Model

Dobson (2002) argues that the complementary log-log model is similar to the logistic and probit models for values of π near 0.5 but differs from them for π near 0 or 1. The model is asymmetrical and the link function in given in terms of

$$g(\pi) = \log[-\log(1-\pi)].$$

3.8 Application

The dependent variable is of the type which elicits a binary response, i.e., being a donor or a non-donor in the Mali population. The logistic regression model is generally the most common method used to fit binary response data however there are two other models that can also be used to fit binary response data, namely, the probit model and the complementary log-log model. All three models have been fit to the Mali data using the appropriate link functions. Data collection methods are described in Chapter 2.

Recall that the KAP survey was implemented and as is with all classical KAP surveys the information obtained from individuals in Mali were really extensive and exhaustive. Hence literature review was used to select only variables that were of importance to measure donor status in the Mali population. Also, section two and section three of the questionnaire were only to be answered by donors and non-donors respectively. Therefore, the variables related to these sections were not included in the final analyses but selective descriptive statistics for these sections can be found in Chapter 2.

Only variables applicable to the outcome variable were selected for further inferences and the following explanatory variables were included in the analysis, age, gender and educational level of the respondent, whether the respondent received blood or not, respondents knowledge about

the different blood groups, whether the respondent saw or heard messages about blood donation, respondents opinion on the best way to spread messages about blood donation, whether respondents opinion on the appropriate way to give blood is to be paid for blood donation or whether respondents thought the appropriate way to give blood is voluntary non-remunerated blood donation. Other explanatory variables included in the analysis were whether the respondent thought that blood transfusion is required to treat malaria or to treat other diseases, whether blood transfusion is required for emergencies/disasters (to help patients recover from accidents; in order to undergo surgery; for mothers in childbirth), whether the respondent thought that blood transfusion is required to correct malnutrition, replace lost fluids of any type or to make up blood volume, and whether they thought that blood donation is a good practice, important and everyone should donate.

All explanatory variables were recorded as categorical variables. Age was categorized into two categories, i.e., thirty years and under, and over thirty years. Gender was used to denote the sex of the respondent as female or male. A variable RB was used to denote whether a respondent received blood or not while the variable KDBG was used to represent whether a respondent had knowledge on the different blood groups or not. The variable edu_level comprised four levels: never went to school, primary school, secondary school, and tertiary education. The variable representing what would be the best way to spread messages about blood donation (SmsgsBD), comprised three levels: Media (Television; Radio; Written media; other written media; Banners), Organizations (Church; Colleges/Schools/University; Hospitals/clinics), and Direct Contact (Telephone; SMS; Word of mouth). If a respondent ever saw or heard messages about blood donation, it was recorded as a "Yes" in the variable HSMBD and no otherwise. The categorical variable BEmerg was used to denote if a respondent thought that the blood required for transfusion was used for emergencies/disasters while the variable BMal was used to denote whether a respondent thought blood transfusion is required to correct malnutrition, replace lost fluids of any type or to make up blood volume. The variable Btrt was recorded as "Yes" if a respondent knew that the blood donated for transfusion is required to treat malaria or to treat other diseases and "No" otherwise. The variables AppWay_VNRBD and AppWay_PD were used to represent a respondents' opinion on the appropriate way to denote blood as being voluntary

non-remunerated blood donation and paid donation respectively. The variable GP was used to denote whether a respondent thought blood donation is a good practice, it is important and everyone should donate and was recorded as "Yes" if a respondent thought it was a good practice and "No" otherwise.

Analyses were done using SAS version 9.3. The PROC LOGISTIC procedure was used to fit the logistic regression model to the donor data. The main model with all thirteen explanatory variables were fitted. All two-way interaction terms of the variables were fitted in the model and investigated one at a time. The main effects and the possible combinations of up to two-way interaction terms were then fitted. The models fitted were accompanied by summary statistics and goodness of fit tests describing how well the model fits the data, the amount of variation in the outcome accounted for by the model, and a basis for comparing the existing model to the other possible models. The predictive accuracy was assessed using statistics such as the concordance index (c), Somers' D (SD), Goodman-Kruskal Gamma (GKG), and Kendall's Tau-a (KT); details of which can be found in Section 3.4. After assessing the above criteria it was found that the inclusion of any or all of the possible interaction terms did not improve the fit of the model. Hence, the final model comprised of all thirteen main effects and no interaction effects.

The effect of applying the different link functions to the data were investigated. The logit, probit and cloglog links were used to fit the data. McCullagh & Nelder (1989) suggest checking the correctness of the link function for binary data by using formal methods. One such formal method suggested by Hinkley (1985) involves squaring the estimated linear predictor and adding it as an extra covariate. Significance of this test could imply the use of the wrong link function. The complementary log-log link is appropriate for modelling data with extreme values, and from Figure 3.2 and Figure 3.3 it can be seen that there are no extreme values hence the complementary log-log link is dropped. Liao (1994) argued that given the similarities between the logit and probit models, either model will give identical substantive conclusions in most applications. Also, Dobson (2002) and Agresti (2002) confirm this by arguing that if there are no extreme values, then the logistic and the probit regression models provide similar results. Many researchers however, prefer using the logit function because the odds ratio can be calculated

and the coefficients can be interpreted easily. As can be seen in Table 3.1, the estimated linear predictor is significant $(\mathrm{p}-\mathrm{value}<.0001)$ whilst the square of the linear predictor is insignificant $(\mathrm{p}-\mathrm{value}=0.4902)$ thereby suggesting that the logit link function is reasonable.

Table 3.1: Checking for Correctness of Link Function

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	P-Value
Intercept	1	0.0833	0.1849	0.20	0.6523
LPred	1	0.9686	0.1549	39.08	<.0001
SLPred	1	-0.0954	0.1383	0.48	0.4902

The overall fit of the model is statistically significant as can be seen from Table 3.2.

Table 3.2: Overall Model Significance Test

Test	Chi-Square	DF	P-Value
Likelihood Ratio	59.6581	17	<.0001
Score	54.1601	17	<.0001
Wald	45.4839	17	0.0002

From Table 3.3 it can be observed that both the Pearson X^2 (p-value = 0.7455) and deviance (p-value = 0.1981) are insignificant thus indicating that the model fits the data reasonably well.

Table 3.3: Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	P-Value
Deviance	138.1924	125		0.1981
Pearson	114.2125	125		0.7455

For a model to be identified correctly, the Pearson chi-square statistic and the deviance, when divided by their degrees of freedom, should be approximately equal to one. If their values are much larger than one, then the assumption of binomial variability may be invalid and the data is thought to exhibit overdispersion. From Table 3.4, it can be observed that both the Pearson chi-square and deviance satisfy this criteria indicating that the model does not display overdispersion.

Table 3.4: Criteria For Assessing Goodness Of Fit

Criterion	**DF**	**Value**	**Value/DF**
Deviance	125	138.1924	1.1055
Scaled Deviance	125	125.0000	1.0000
Pearson Chi-Square	125	114.2186	0.9137
Scaled Pearson X2	125	103.3148	0.8265

The Hosmer and Lemeshow Goodness-of-fit test was used to test for adequacy of the model. As discussed in Section 3.5.2, if the model is a good fit to the data then the Hosmer-Lemeshow Goodness-of-Fit test should have an associated p-value greater than 0.05. The Hosmer and Lemeshow Goodness-of-Fit test presented in Table 3.5 indicates a chi-square value of 5.7451 with 8 degrees of freedom and p-value = 0.6758, hence we do not reject model adequacy at the 0.05 level, and conclude that this measure supports the adequacy of model for the data.

Table 3.5: Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	**DF**	**P-Value**
5.7451	8	0.6758

A plot of sensitivity versus $1 -$ specificity over all possible cutpoints is shown in Figure 3.1. This curve is called a ROC Curve and measures the model's ability to discriminate between subjects who are donors versus those who are non-donors. The area under ROC Curve is 0.751 and indicates that 75.1% of the probabilities of donor status is predicted correctly by the model. The ROC curve also serves to confirm that the model is a good fit.

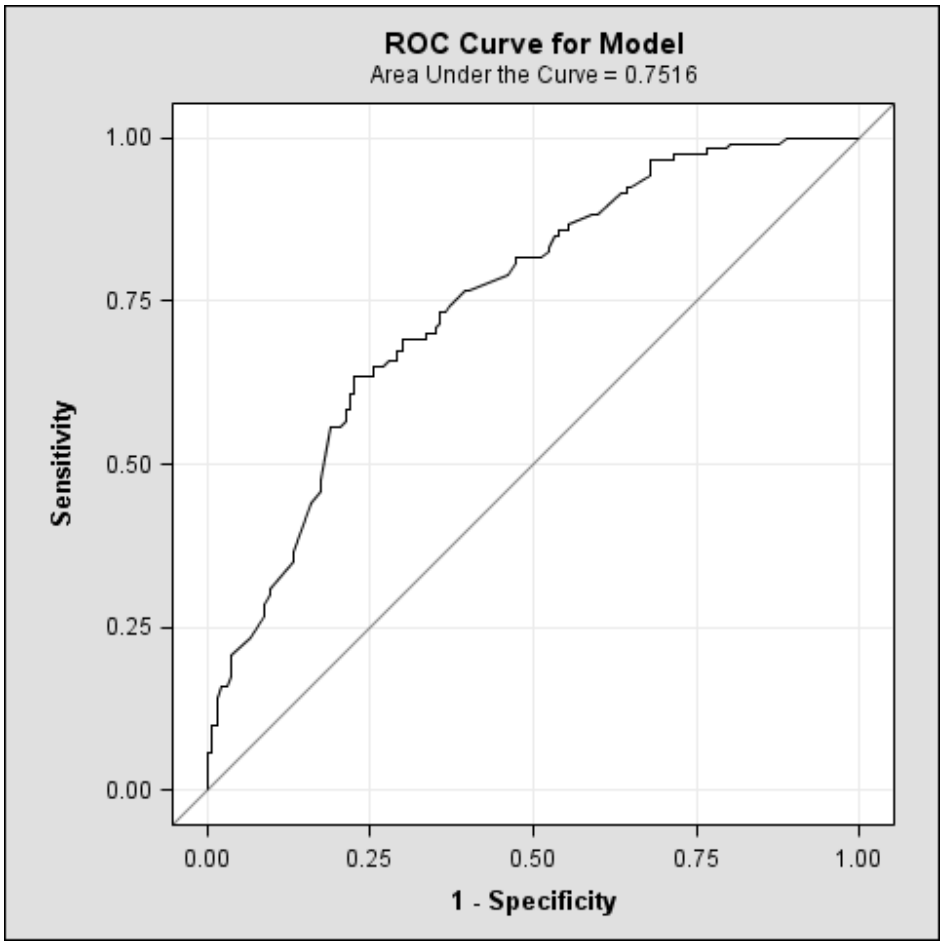


Figure 3.1 ROC Curve

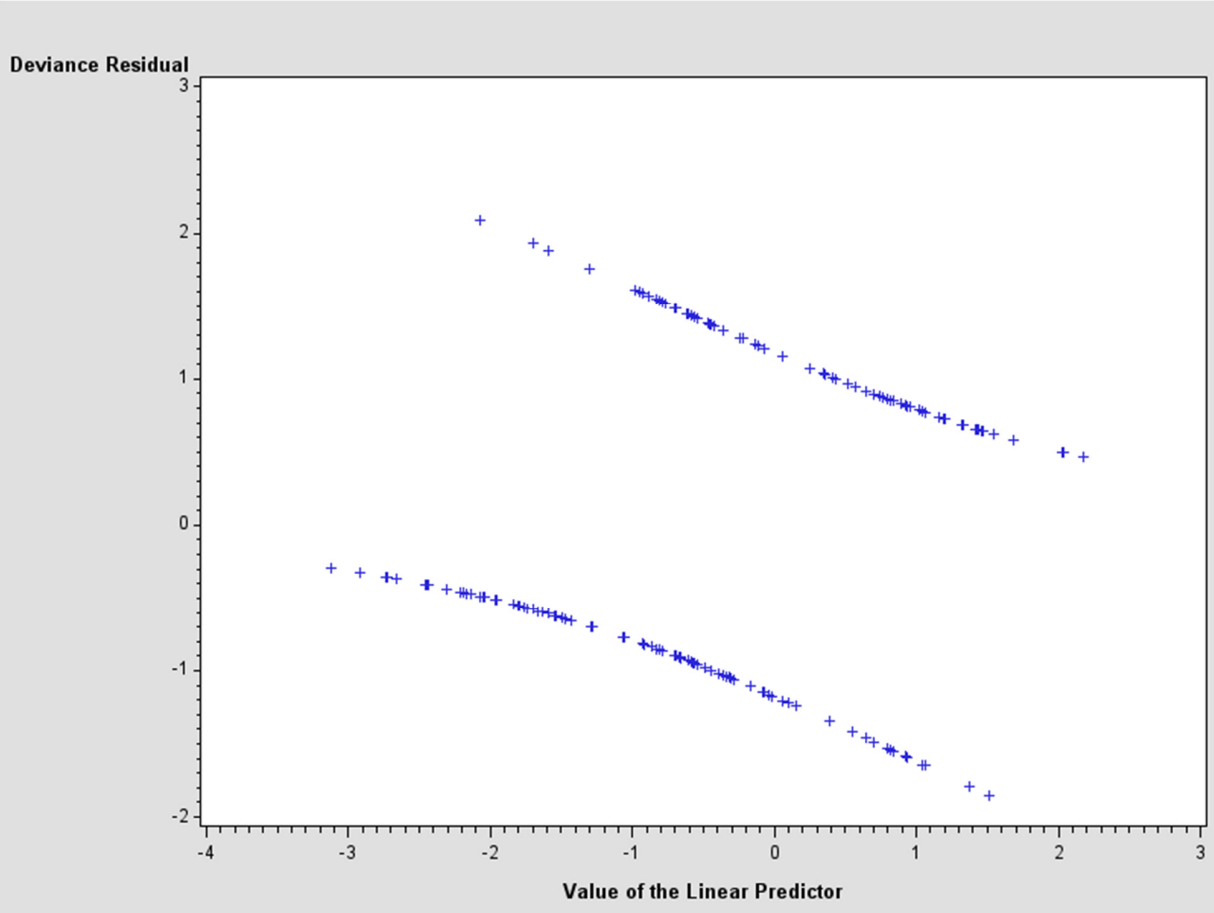


Figure 3.2: Deviance Residual Plot

The deviance residuals as can be seen from Figure 3.2 are between -2 and 2 indicating that model inadequacy is not supported. Also, from the plot on Cook's distance presented in Figure 3.3, it is observed that all the outliers of the Cook's distance do not exceed the 1.0 rule of thumb indicating that there are no influential cases present. However, further investigations on influential observations were done and are presented in Figure 3.4.

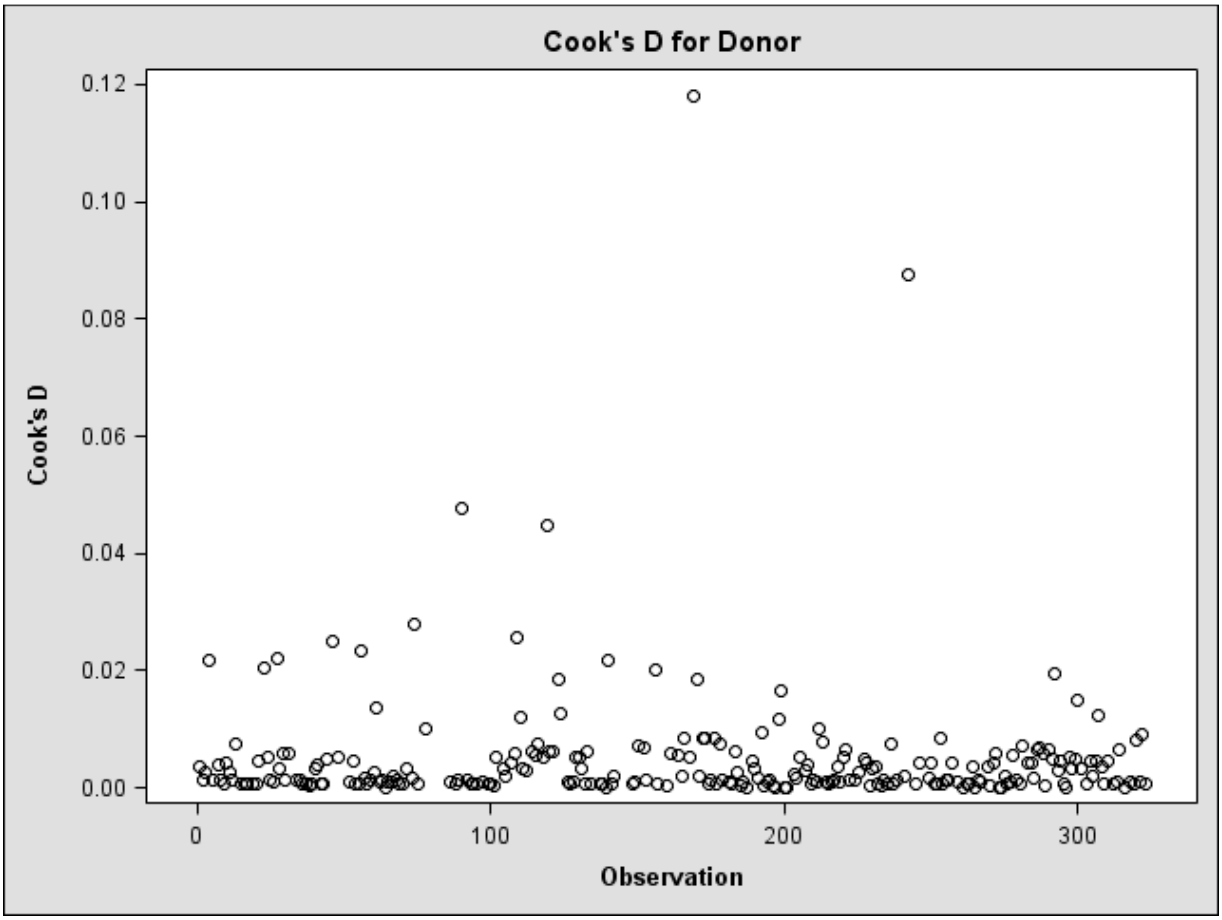


Figure 3.3: Cooks'D Plot

The observations that are the farthest away from zero are considered influential observations. Indicated from the plot below, it appears that observations 27, 56, 123, 169, 198, 242 and 300 are influential observations. How great the influence a single observation has on the coefficients of the model and on any lack of overall fit thereof can be assessed. The observations with unduly high influence were investigated together with the effect of removing the influential observation on the model and the necessary results were presented both with and without the influential observation(s). The inclusion and exclusion of the influential observation(s) did not appear to have any significant effect or influence on the estimated coefficients hence confirming that the logistic model fitted is a good fit.

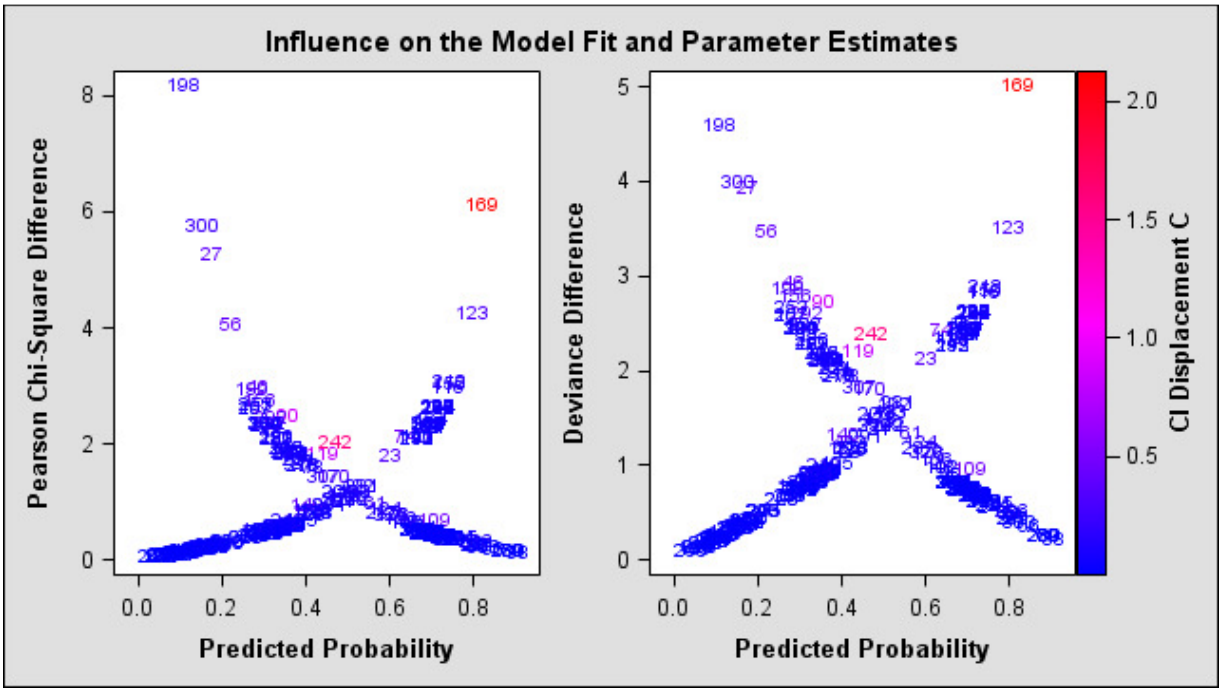


Figure 3.4: Influence Plot

The parameter estimates, as well as the odds ratio (OR) with their 95% confidence interval, and the p-vales are given in Table 3.6. It can be observed that Age has no significant effect (at 5% level of significance) on donor status in the Mali population. This implies that controlling for the other covariates, the odds of being a donor for an individual 30 years of age and under is no different from an identical individual that is over the age of 30.

Gender had a significant effect (p-value < 0.0001, CI: 0.471 –2.134) on the outcome of being a donor in this population with a confidence interval that ranges from 0.471 to 2.134. Hence a single unit increase in gender reduce the chances of being a donor. The male gender was used as the reference category or baseline and it can be seen that the odds of a female being a donor is 0.196 times that of a male, i.e. females are less likely to be blood donors in the Mali region.

The odds of being a donor for a unit increase in individuals that received blood were approximately 3.7 times (3.714, p-value = 0.1381, CI: 0.675 – 16.987) the odds of being a donor for identical individuals that did not receive blood, however, this is not a significant difference at 5% level of significance.

The odds of being a donor for an individual that had knowledge about the different type of blood groups were almost four times (3.975, p-value = 0.0001, CI: 1.956 – 8.080) the odds of being a donor for an identical individual that did not have knowledge about the different type of blood groups, with a 95% confidence interval for the odds ratio from 1.956 to as high as 8.080. This had a significant effect (at 5% level of significance) on donor status in the said population, and implies that controlling for all other covariates, the odds of being a donor for an individual that had knowledge about the different type of blood groups is significantly different from an individual that did not have the same knowledge.

Focusing on whether an individual saw or heard messages about blood donation, where the baseline individual was one that did not remember or recall having ever heard or seen messages about blood donation, an unexpected result is apparent where, whether an individual that had heard or seen messages about blood donation or not, had no significant effect on being a donor, with a 95% confidence interval ranging from 0.097 to 6.252.

Focusing on the educational level of an individual in the Mali population where the baseline individual has tertiary education, a somewhat unexpected significant result can be seen, where, an individual that never went to school is approximately 3.8 times (3.757, p-value = 0.0443, CI: 1.034 – 13.647) the odds of being a donor compared to an identical individual that has tertiary education. While the odds of being a donor for a single unit increase for an individual that had either primary school or secondary school education as compared to an identical individual that has tertiary education, were 1.946 and 1.146 respectively.

Focusing on the reasons why people require blood transfusion, did not appear to have any significant effect on donor status in the population. An individual that thought blood transfusion is required to treat malaria or to treat other diseases is 1.862 times the odds of being a donor as opposed to an identical individual that did not hold this view, with the odds ratio not being significantly different from a unit (p-value = 0.2809, CI: 0.602 – 5.761). An individual that thought that blood transfusion is required for emergencies/disasters (to help patients recover from accidents; in order to undergo surgery; for mothers in childbirth), is surprisingly less likely to be a donor (OR: 0.70, p-value = 0.6087, CI: 0.178 – 2.745) than an identical individual that did not

hold this view and an individual that thought blood transfusion is required to correct malnutrition, replace lost fluids of any type or to make up blood volume is 2.56 times the odds of being a donor when compared to an identical individual that did not hold this view (p-value = 0.5158, CI: 0.149 – 44.182).

With regard to an individual's opinion on the best way to spread messages about blood donation, where direct contact (telephone; SMS; word of mouth) had been used as a baseline for making a comparison, using the media or using different organizations to spread messages about blood donation have had no significant effect on donor status. However, the odds for a single unit increase for an individual that thought the media should be used to spread messages about blood donation is 1.099 times the odds of being a donor when compared to an identical individual that thought direct contact should be used (OR: 1.099, p-value = 0.7895, CI: 0.549 – 2.198). Similarly, a single unit increase for an individual that thought spreading messages about blood donation through organizations, is 1.124 times the odds of being a donor when compared to an otherwise identical individual (OR: 1.124, p-value = 0.7749, CI: 0.504 – 2.506).

With regard to the practice of blood donation, an individual that thought blood donation is a good practice, important and everyone should donate, is approximately 1.25 times the odds of being a donor when compared to an identical individual that did not hold this view. This insignificant difference (OR: 1.25, p-value = 0.5370, CI: 0.616 – 2.535) implies that a single unit increase for an individual that thought blood donation is a good practice, important and everyone should donate, increases an individual's chance of being a donor.

Individuals that held the view that the appropriate way to give blood is voluntary non-remunerated blood donation were 3.179 times the odds of being a donor than otherwise identical individuals that did not hold this view. This implies that individuals that held this view were more likely to be blood donors than individuals that did not. However this is not significant (OR: 3.179, p-value = 0.0514, CI: 0.993 – 10.175) at the 5% level of significance.

Also, individuals that held the view that the appropriate way to give blood is to be paid for the blood donated, were 1.130 times the odds of being a donor than otherwise identical individuals that did not hold this view. Again, this is an insignificant difference (OR: 1.130, p-value = 0.9118,

CI: 0.616 – 2.535) at 5% level of significance and implies that a single unit increase for individuals that held this view increased their probability of being a donor with a 95% confidence interval ranging from 0.616 – 2.535. The insignificant result implies that controlling for all other covariates, the odds of being a donor for a single unit increase for an individual that held a certain view on the appropriate way to give blood, i.e. being a voluntary non-remunerated blood donor or a paid donor, is not significantly different from the odds of an identical individual that did not hold this particular view.

Table 3.6: Parameter Estimates and Odds Ratio of Main Model

Parameter	Estimate	Standard Error	Odds Ratio	95% CI	P-Value
Intercept	-1.5860	1.2986			0.2220
Age (Ref = > 30)					
≤ 30	0.00262	0.3855	1.003	0.471 –2.134	0.9946
Gender (Ref = Male)					
Female	-1.6289	0.3121	0.196	0.106 – 0.362	<.0001
RB (Ref = No)					
Yes	1.2199	0.8227	3.387	0.675 – 16.987	0.1381
KDBG (Ref = No)					
Yes	1.3801	0.3619	3.975	1.956 – 8.080	0.0001
HSMBD (Ref = Do not remember)					
Yes	-0.2487	1.0620	0.780	0.097 – 6.252	0.8148
No	-1.1216	1.1529	0.326	0.034 – 3.121	0.3306
Edu_level (Ref = Tertiary education)					
Never went to school	1.3237	0.6581	3.757	1.034 – 13.647	0.0443
Primary education	0.6658	0.6220	1.946	0.575 – 6.586	0.2845
Secondary education	0.1364	0.3653	1.146	0.560 – 2.345	0.7089
Btrt (Ref = No)					

Yes	0.6215	0.5764	1.862	0.602 – 5.761	0.2809
BEmerg (Ref = No)					
Yes	-0.3569	0.6972	0.700	0.178 – 2.745	0.6087
BMal (Ref = No)					
Yes	0.9433	1.4516	2.568	0.149 – 44.182	0.5158
SmsgsBD (Ref = Direct contact)					
Media	0.0944	0.3537	1.099	0.549 – 2.198	0.7895
Organizations	0.1170	0.4090	1.124	0.504 – 2.506	0.7749
GP (Ref = No)					
Yes	0.2228	0.3608	1.250	0.616 – 2.535	0.5370
AppWay_VNRBD (Ref = No)					
Yes	1.1565	0.5936	3.179	0.993 – 10.175	0.0514
AppWay_PD (Ref = No)					
Yes	0.1219	1.1004	1.130	0.131 – 9.763	0.9118

3.9 Summary

Logistic regression analysis of the data was carried out to model the relationship between the response variable, i.e., donor status in the Mali population and the independent variables described in section 3.8. Since logistic regression is a special case of generalized linear modelling and is not the only model that can be used to model binary response data, two other models were used, namely, the probit model and the complementary log-log model via the appropriate link functions. Formal methods can be used to check the correctness of the link function. After fitting the data to the model and investigating the effect of all three approaches, it can be seen in Table 3.1, that the estimated linear predictor is significant (p-value < 0.0001) whilst the square of the linear predictor is insignificant (p-value = 0.4902) which implies that the prediction given by the linear predictor is not improved by adding the square linear predictor term, thereby suggesting the consistency of the choice of the link function.

The deletion of unduly high influential cases were investigated and the necessary results were presented both with and without the influential observation(s). There did not appear to be any substantial changes in the model fit or estimated parameters with the deletion of such cases, hence it can be concluded that the outlying cases are not influential and were retained in the analysis. This further confirmed that the model is a good fit.

It can be seen that there are only two factors that have a significant effect on donor status in the Mali population. One of which is gender and the other being knowledge about the different blood groups. The inclusion of any or all of the possible interaction terms did not improve the fit of the model and hence was not included in further analyses. In terms of blood donation and gender, the male population responded as majority donors at 58.8% and it follows that the odds of a female being a donor is 0.196 times that of a male, i.e. females are less likely to be blood donors in the Mali region. In terms of blood donation and knowledge about the different blood groups, a single unit increase for an individual that had knowledge about the different type of blood groups increases the chance of being a donor. This significant difference implies that the odds of being a donor for an individual that had knowledge about the different blood groups were almost four times the odds of being a donor compared to an otherwise identical individual that did not have knowledge about the different type of blood groups, while controlling for all other covariate. All other explanatory variables did not appear to have a significant effect on the outcome of being a donor in this region.

Chapter 4: Correspondence Analysis

4.1 Introduction to Correspondence Analysis

Correspondence analysis (CA) is a technique for displaying the rows and columns of a data matrix (primarily, a two-way contingency table) as points in dual low-dimensional vector spaces (Greenacre, 1984). This exploratory multivariate technique is the brainchild of Jean-Paul Benzècri which originated in France in the early 1960s, and is used for the graphical and numerical analysis of almost any data matrix with nonnegative entries, but primarily involves tables and counts. Greenacre & Blasius (2006) have reported that CA can be extended to analyze presence/absence data, rankings and preferences, paired comparison data, multiresponse tables, multiway tables and square transition tables amongst others. They have further reported that since it is oriented toward categorical data, it can be used to analyze almost any type of tabular data after suitable data transformation and recording. In correspondence analysis it is claimed that no underlying distribution has to be assumed and no model has to be hypothesized, but a decomposition of the data is obtained in order to study their "structure" (Panagiotakos & Pitsavos, 2004).

Similar to PCA, the rows or columns of the data matrix are assumed to be points in a high-dimensional Euclidean space, and the method aims to redefine the dimensions of the space so that the principal dimensions capture the most variance possible, allowing for lower-dimensional descriptions of the data (Greenacre & Blasius, 2006).

The basic method underlying CA is discussed in Greenacre & Blasius (2006) and is detailed below.

Greenacre (1984) presents the theory of CA in terms of the singular – value decomposition (SVD) of a suitably transformed matrix. Greenacre & Blasius (2006) present the theory in the same context and describes the CA algorithm by use of a single cross-table, or a two-way

contingency table, with I rows and J columns, and the $(i,j)^{\text{th}}$ element **N** is denoted n_{ij}. The correspondence matrix $\boldsymbol{P}$ is calculated as a first step, with elements

$$p_{ij} = \frac{n_{ij}}{n},$$

where n is the sample size.

Corresponding to each element p_{ij} of $\boldsymbol{P}$ is a row sum

$$p_{i.} = \frac{n_{i.}}{n},$$

and a column sum

$$p_{.j} = \frac{n_{.j}}{n},$$

denoted by r_i and c_j respectively.

Greenacre & Blasius (2006) explain that these marginal relative frequencies which are called masses, play dual roles in CA by serving to center and to normalize the correspondence matrix and under the null hypothesis of independence, the expected values of the relative frequencies p_{ij} are the products of $r_i c_j$ of the masses. Further, the process of centering involves calculating differences $\left(p_{ij} - r_i c_j\right)$ between observed and expected relative frequencies, and normalization involves dividing the said differences by the square roots of $r_i c_j$, leading to a matrix of standardized residuals as follows

$$s_{ij} = \frac{\left(p_{ij} - r_i c_j\right)}{\sqrt{r_i c_j}}.$$

In matrix notation this is written as:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{C}^{\mathrm{T}})\mathbf{D}_c^{-1/2}$$

where $\mathbf{r}$ and $\mathbf{c}$ are vectors of row and column masses, and $\mathbf{D}_r$ and $\boldsymbol{D}_c$ are diagonal matrices with the masses on respective diagonals.

The sum of squared elements of the matrix of standard residuals given by

$$\sum_i \sum_j s_{ij}^2 = \text{trace}(\mathbf{S}\mathbf{S}^{\mathrm{T}}) ,$$

is called the total inertia and is defined as the amount that quantifies the total variance in the cross-table. The standardized residuals in $\mathbf{S}$ resemble those in the calculation of the chi-square statistic, X^2 , apart from the division by n to convert the original frequencies to relative ones. The following relationship exists:

$$\text{total inertia} = \frac{X^2}{n}.$$

By the use of the SVD, the association structure in the matrix $\mathbf{S}$ is revealed

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}},$$

where $\boldsymbol{\Sigma}$ is the diagonal matrix with singular values in descending order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_S > 0$, and S is the rank of the matrix $\mathbf{S}$. The columns of $\mathbf{U}$ and $\mathbf{V}$, called left singular vectors and right singular vectors respectively, are orthonormal: $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$.

The connection between the SVD and the eigenvalue decomposition can be understood as follows

$$\mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}} = \mathbf{V}\boldsymbol{\Sigma}^{\mathbf{2}}\mathbf{V}^{\mathrm{T}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$$

$$\mathbf{S}\mathbf{S}^{\mathrm{T}} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^{\mathrm{T}} = \mathbf{U}\boldsymbol{\Sigma}^{\mathbf{2}}\mathbf{U}^{\mathrm{T}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}},$$

showing that the right singular vectors of $\mathbf{S}$ correspond to the eigenvectors of $\mathbf{S}^{\mathrm{T}}\mathbf{S}$, the left singular vectors correspond to the eigenvectors of $\mathbf{S}\mathbf{S}^{\mathrm{T}}$, and the squared singular values σ^2 in $\boldsymbol{\Sigma}^2$ correspond to the eigenvalues λ of $\mathbf{S}^{\mathrm{T}}\mathbf{S}$ or $\mathbf{S}\mathbf{S}^{\mathrm{T}}$, where $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. These eigenvalues are termed principal inertias and the sum $\sum_S \lambda_S$ is equal to the total inertia since:

$$\text{trace}(\mathbf{S}\mathbf{S}^{\mathrm{T}}) = \text{trace}(\mathbf{S}^{\mathrm{T}}\mathbf{S}) = \text{trace}(\boldsymbol{\Sigma}^2) = \text{trace}(\boldsymbol{\Lambda}).$$

Hendry et al. (2014) have reported that from the result of the SVD, the principle coordinates of the points, i.e., coordinates with respect to their principal axes, can be defined. The row principle coordinates are calculated as

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Sigma},$$

whilst the column principle coordinates are calculated as

$$\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\boldsymbol{\Sigma}.$$

The row standard coordinates can be calculated as follows

$$\mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U},$$

and the standard coordinates as

$$\mathbf{B} = \mathbf{D}_c^{-1/2}\mathbf{V}.$$

These row and principle coordinates are used to produce the graphical displays of the points on the CA maps and the amount of inertia explained by each principal axis is given by the square of the corresponding single value.

The most common method used to test for significant associations between rows and columns in a contingency table is the chi-square statistic. Greenacre & Blasius (2006) define the chi-square statistic as the sum of squared deviations between observed and expected frequencies, where the expected frequencies are those calculated under the independence model

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}},$$

where $\hat{n}_{ij} = n_i \times n_j/n$.

This calculation can be repeated for relative frequencies p_{ij} , to obtain

$$\frac{X^2}{n} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij} - \hat{p}_{ij}\right)^2}{\hat{p}_{ij}},$$

which is the chi-square statistic divided by the grand total n of the table and where $p_{ij} = r_i \times c_j$.

Now, using the above equation, the total inertia can be rewritten as

$$\frac{X^2}{n} = \sum_{i=1}^{I}\sum_{j=1}^{J} c_j \frac{\left(p_{ij}/c_j - r_i\right)^2}{r_i},$$

where p_{ij}/c_j is an element of the jth column profile: $p_{ij}/c_j = n_{ij}/n_{.j}$; and r_i is the corresponding element of the average column profile: $r_i = n_{i.}/n$. The squared distance is a Euclidean – type distance where each squared difference is divided by the corresponding average value r_i , and the weight of the column profile is in its mass c_j . The X^2 distances between profiles in CA are visualized as ordinary Euclidean distances.

4.2 Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) is an extension of correspondence analysis (CA) and allows the analysis of the pattern of the relationships of many variables that are categorical. MCA is generally used to analyze a set of observations that are described by a set of nominal variables. The nominal variables may comprise of two or more levels and each level is coded as binary.

Consider the multivariate case where there are Q categorical variables which are coded as indicator matrices $\mathbf{Z}_1, \mathbf{Z}_2, \ldots \boldsymbol{Z}_{\mathrm{Q}}$, a Burt matrix which is the matrix of all two-way cross - tabulations of the categorical variables, denoted by $\mathbf{C}$ and an Indicator matrix denoted by $\mathbf{Z}$. The method of applying the CA algorithm described in Section 4.1 to the Indicator matrix or to a Burt matrix is called Multiple Correspondence Analysis (MCA).

According to Greenacre & Blasius (2006), let J_q denote the number of categories for the qth categorical variable and let $J = \sum_q J_q$ be the total number of categories. $\mathbf{Z}$ is then of order $n \times J$ and $\mathbf{C}$ is of order $J \times J$. Given that $\mathbf{Z}$ has a total sum nQ, with row sums equal to a constant Q and column sums equal to the marginal frequencies of each variable, the correspondence matrix is thus $\left({}^{1}/_{Qn}\right)\mathbf{Z}$, the row mass matrix is $\left({}^{1}/_{n}\right)\mathbf{I}$, and the column mass matrix is $\mathbf{D}$. The SVD to compute the CA of $\mathbf{Z}$ in its uncentred form is hence

$$\sqrt{n}\frac{\mathbf{Z}}{Qn}\mathbf{D}^{1/2} = \mathbf{U\Gamma V}^{\mathrm{T}},$$

where $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$, and $\boldsymbol{\Gamma}$ is the diagonal matrix of positive numbers in descending order $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_s > 0$, which are the singular values.

The trivial solution is eliminated by decomposing the following matrix

$$\sqrt{n}\left(\frac{\mathbf{Z}}{Qn} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{D}\right)\mathbf{D}^{-1/2},$$

where $(1/n)\mathbf{1}$ is the vector of row masses and $\mathbf{1}^{\mathrm{T}}\mathbf{D}$ is the vector of column masses of the indicator matrix.

The SVD for the CA of the Burt matrix $\mathbf{C}$ is as follows:

$$\mathbf{D}^{-1/2}\frac{\mathbf{C}}{Q^2 n}\mathbf{D}^{-1/2} = \mathbf{V}\boldsymbol{\Gamma}^2\mathbf{V}^{\mathrm{T}} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}},$$

where $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}$ and $\mathbf{C} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}$.

The trivial solution is removed in the form of the expected relative frequencies

$$\mathbf{D}^{-1/2}\left(\frac{\mathbf{C}}{Q^2 n} - \mathbf{D}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{D}\right)\mathbf{D}^{-1/2}.$$

4.3 Adjustments to Inertias in MCA

According to Greenacre (1984), the usual computation of the explained inertia for each dimension in MCA underestimates the quality of fit and he suggests another calculation which leads to a more precise estimate.

Greenacre & Blasius (2006) propose the possible partial remedying of the percentage-of-inertia problem in regular MCA by using simple scale readjustments of the MCA solution. Using this approach, the total inertia is measured by the average inertia of all off-diagonal blocks of $\mathbf{C}$ by removing the fixed contributions of the diagonal blocks shown below

$$\text{average off} - \text{diagonal inertia } = \frac{Q}{Q-1}\left(\text{inertia}(\mathbf{C}) - \frac{J-Q}{Q^2}\right).$$

Further, parts of the inertia are calculated from the principal inertias λ_s^2 of $\mathbf{C}$ (or of $\mathbf{Z}$); hence for each $\lambda_s \geq 1/Q$, the adjusted inertias are calculated as follows

$$\lambda_s^{\text{adj}} = \left(\frac{Q}{Q-1}\right)^2 \left(\lambda_s - \frac{1}{Q}\right)^2.$$

Greenacre & Blasius (2006) propose that the adjusted solution should be routinely reported as it considerably improves the measure of fit as well as removes the inconsistency about which of the two matrices to analyze i.e., indicator or Burt matrix.

Adjusted inertias have also been proposed by Benzècri (1979) and are expressed as percentages of their own sum over the dimension s for which $\lambda_s \geq 1/Q$. Greenacre (1994) however, claims that this adjustment is too optimistic.

Because MCA has attractive properties of optimality of scale values (thanks to achieving maximum inter-correlation and thus maximum reliability in terms of Cronbach's alpha), the compromise offered by the adjusted MCA solution is the most sensible one and the one that we recommend (Greenacre & Blasius, 2006).

4.4 Application

MCA locates all the categories in Euclidean space and aims to produce a solution where objects within the same category are plotted close together whereas objects in different categories are plotted far apart. The plotting of the variables are useful for detecting the clustering of attributes. MCA is used to represent and model datasets as "clouds" of points in a multidimensional Euclidean space; this means that it is distinctive in describing the patterns geometrically by locating each variable/unit of analysis as a point in a low-dimensional space (Costa et al., 2013). Each object will be as close together as possible to the category points of categories that apply to the object, thus the categories divide the object into homogeneous subgroups. So if a certain variable discriminates well, the objects will be close to the categories to which they belong.

The map in Figure 4.1 was generated from calculating χ^2 distances of points represented in the form of two-way contingency tables and the first two dimensions plotted, are used to examine the associations among the categories.

The variables appear to be clustered together making it difficult to differentiate between the points and those variables situated about the origin are not well represented in the Map and do not add to the interpretation of the display. It can be seen that only 13.3 percent of the data is explained by the MCA map which is relatively low. Also, the two dimensions account for 21.07 percent of the total association indicating that there is 78.93 percent error in the display. This implies that the two-dimensional figure accounts for 21.07 percent of the variability in the data, which leaves 78.93 percent unaccounted for.

Inertia and Chi-square decomposition for the MCA is presented in Table 4.1. The total inertia indicates the accuracy of the display and the total Chi-square statistic, which measures the association between the rows and columns in the full dimension of the table is 14573.6 with 1681 degrees of freedom. From Table 4.1 it can be seen that the percentage of inertia accounted for by the first ten dimensions are, 13.13 percent, 7.95 percent, 6.91 percent, 6.06 percent, 4.97 percent, 4.75 percent, 4.08 percent, 4.05 percent, 4.02 percent and 3.70 percent, respectively. Also, 59.2 percent of the total variation is accounted for by the first ten dimensions.

MCA was also carried out using Greenacre and Blasius (2006) proposed method to the adjustment of inertias described in Section 3.4. This adjustment changes the scale of each dimension of the map to best approximate the two tables of association between pairs of variables and everything else in the solution remains intact.

From Table 4.2 it can be seen that when the principal inertias are adjusted, the percentage explained by the two dimensions is 63.56 percent, which is much higher than the 21.07 percent accounted for in the MCA map without the adjustment to inertias. Also, more than 70 percent of the total variation is accounted for by the first three dimensions. The adjustment led to estimates of the explained inertias that are much closer to the true values than the values obtained in MCA.

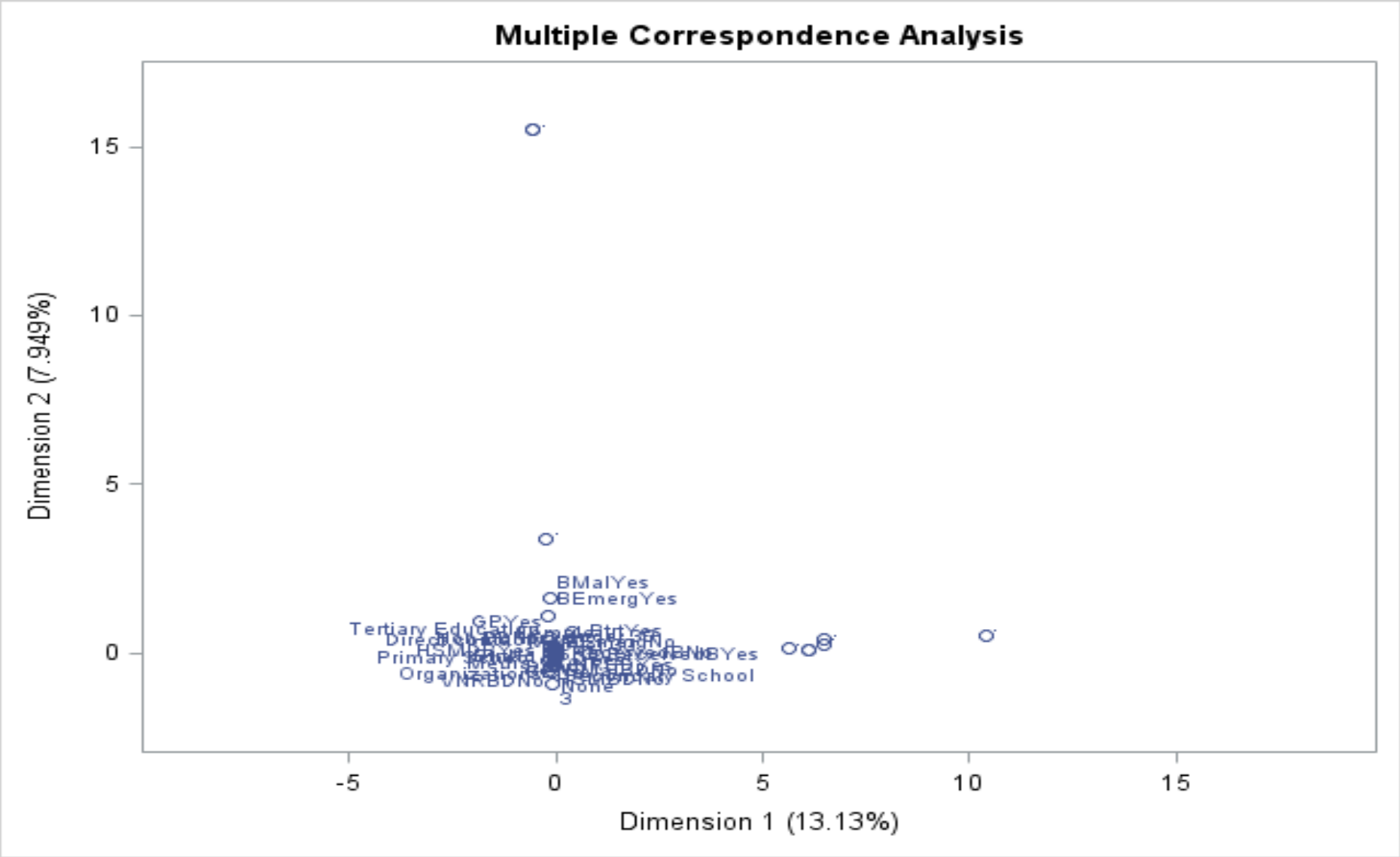


Figure 4.1: Two-dimensional MCA MAP

Table 4.1: Inertia and Chi-Square Decomposition

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.54112	0.29281	1912.9	13.13	13.13
0.42109	0.17732	1158.4	7.95	21.07
0.39256	0.15410	1006.8	6.91	27.98
0.36755	0.13509	882.5	6.06	34.04
0.33286	0.11079	723.8	4.97	39.01
0.32545	0.10592	692.0	4.75	43.75
0.30182	0.09110	595.1	4.08	47.84
0.30041	0.09025	589.6	4.05	51.88
0.29928	0.08957	585.1	4.02	55.90
0.28712	0.08244	538.6	3.70	59.59
0.27861	0.07763	507.1	3.48	63.07
0.27370	0.07491	489.4	3.36	66.43
0.27120	0.07355	480.5	3.30	69.73
0.26446	0.06994	456.9	3.14	72.86
0.26228	0.06879	449.4	3.08	75.95
0.25333	0.06418	419.3	2.88	78.82
0.24715	0.06108	399.0	2.74	81.56
0.23887	0.05706	372.8	2.56	84.12
0.23576	0.05558	363.1	2.49	86.61
0.23042	0.05310	346.9	2.38	88.99
0.22744	0.05173	337.9	2.32	91.31
0.20876	0.04358	284.7	1.95	93.26
0.19716	0.03887	253.9	1.74	95.01
0.18532	0.03434	224.4	1.54	96.55
0.18358	0.03370	220.2	1.51	98.06
0.15740	0.02477	161.9	1.11	99.17
0.13626	0.01857	121.3	0.83	100.00
	2.23077	14573.6	100.00	

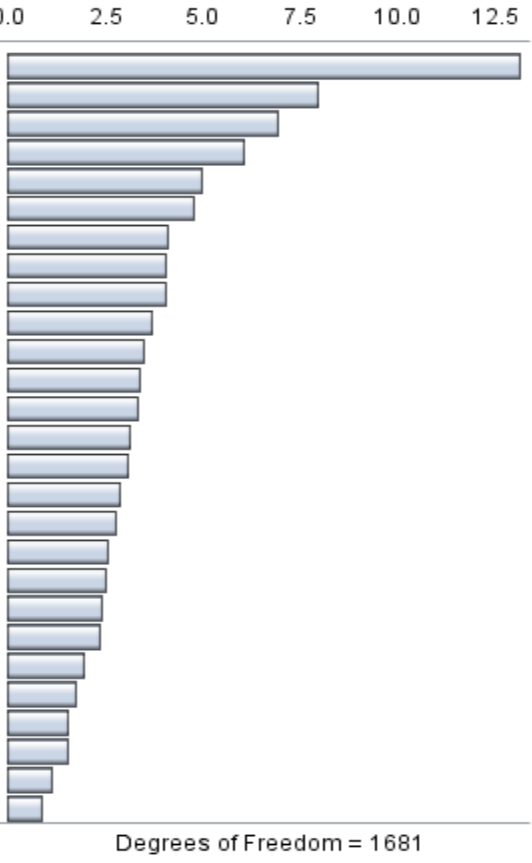
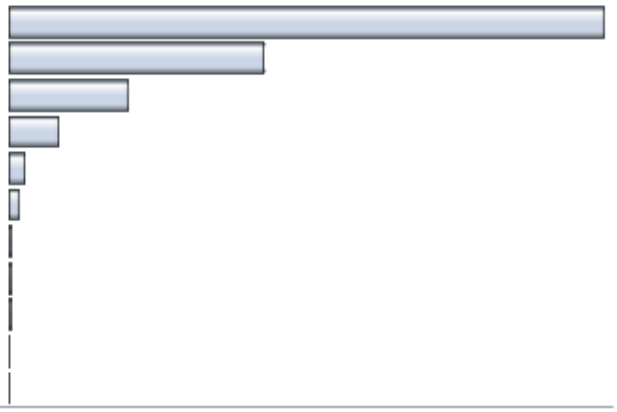


Table 4.2: Greenacre Adjustment to Inertia Decomposition

Greenacre Adjusted Inertia Decomposition

Principal Inertia	Adjusted Inertia	Percent	Cumulative Percent
0.27356	0.04738	44.48	44.48
0.20383	0.02033	19.09	63.56
0.16176	0.00946	8.88	72.44
0.13020	0.00401	3.76	76.21
0.10444	0.00126	1.19	77.39
0.09918	0.00089	0.84	78.23
0.08569	0.00024	0.22	78.45
0.08464	0.00020	0.19	78.64
0.08359	0.00017	0.16	78.80
0.07677	0.00003	0.03	78.83
0.07309	0.00000	0.00	78.84
	0.08399	78.84	



The figures presented in the following three pages describe the MCA map with adjustment of principal inertias along the different dimensions using Greenacres (1994) method. In Figure 4.2, there is a cluster of variables making it difficult to differentiate between the points. A similar conclusion could be seen in Figure 4.1, which represents the two-dimensional MCA map without the adjustment to inertias. The variables situated about the origin are not well represented in the Map and do not add to the interpretation of the display.

Figure 4.3 presents the MCA map with adjustment of principal inertias along Dimension 1 and Dimension 3. There appears to be some clustering of variables about the origin, however, it can be noted that categories corresponding to the positive response of variables, BEmerg, Btrt and BMal, can be found on the bottom left of the MCA map. These variables were used to assess if the individual had knowledge on the usage of the blood required for donation.

Also, the categories for lower educational level (i.e., none and primary school), no knowledge about the different type of blood groups (KDBGNo), having not seen or heard or having not remembered to have seen or heard the messages about blood donation (HSMBDNo, 3) and having a strong opinion that the appropriate way to give blood should not be voluntary non-remunerated blood donation (VNRBDNo), are situated toward the top left of the map.

A similar pattern is displayed in Figure 4.4, which presents the MCA map with adjustment of principal inertias along Dimension 2 and Dimension 3. There appears to be a clustering of variables about the origin making it difficult to differentiate between the points on the map. However, the category of responses to the variables described above for the lower and upper left regions of the MCA map in Figure 4.3, is again evident in the map in Figure 4.4.

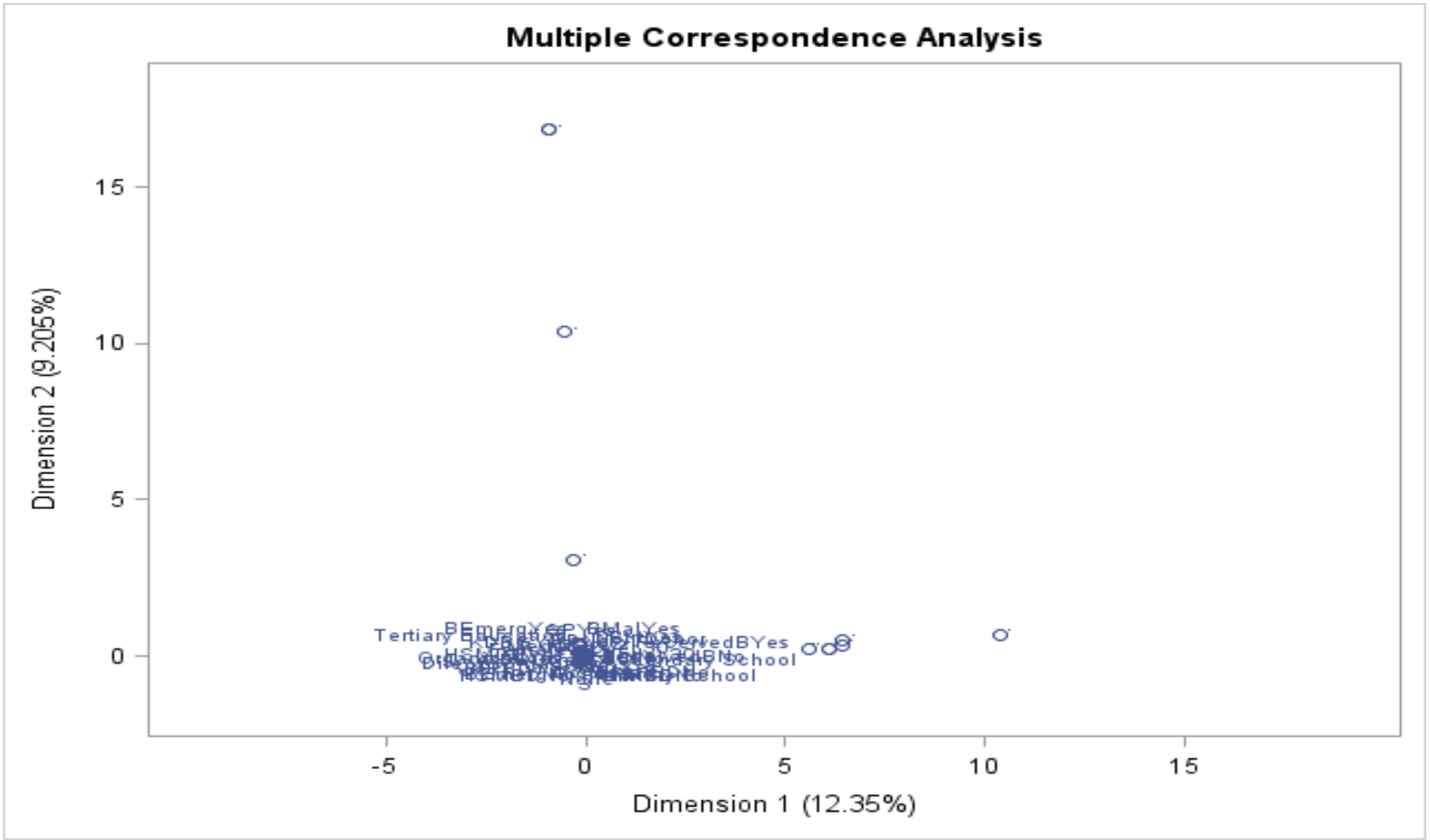


Figure 4.2: MCA MAP with adjustment of principal inertias along Dimension 1 and Dimension 2

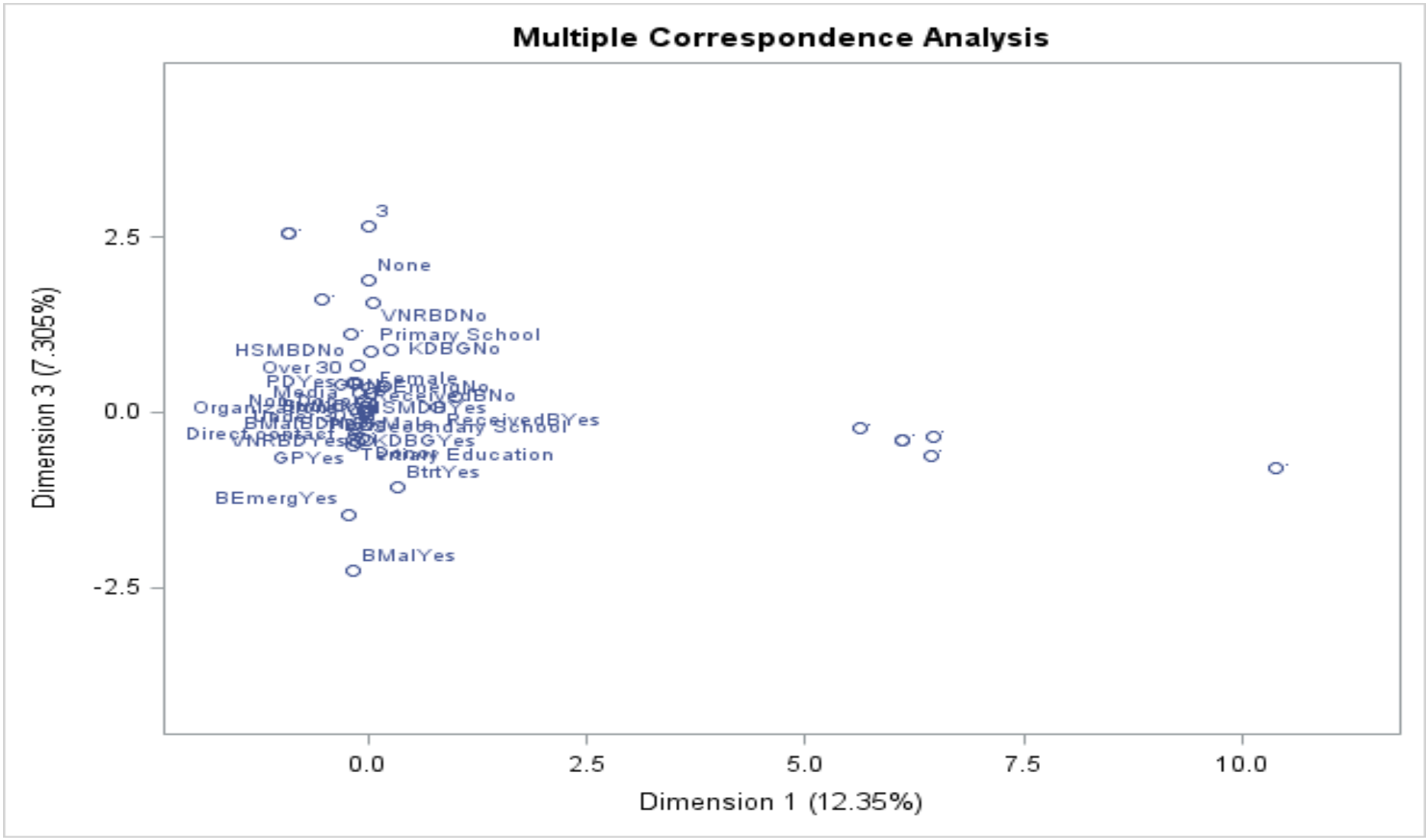


Figure 4.3: MCA MAP with adjustment of principal inertias along Dimension 1 and Dimension 3

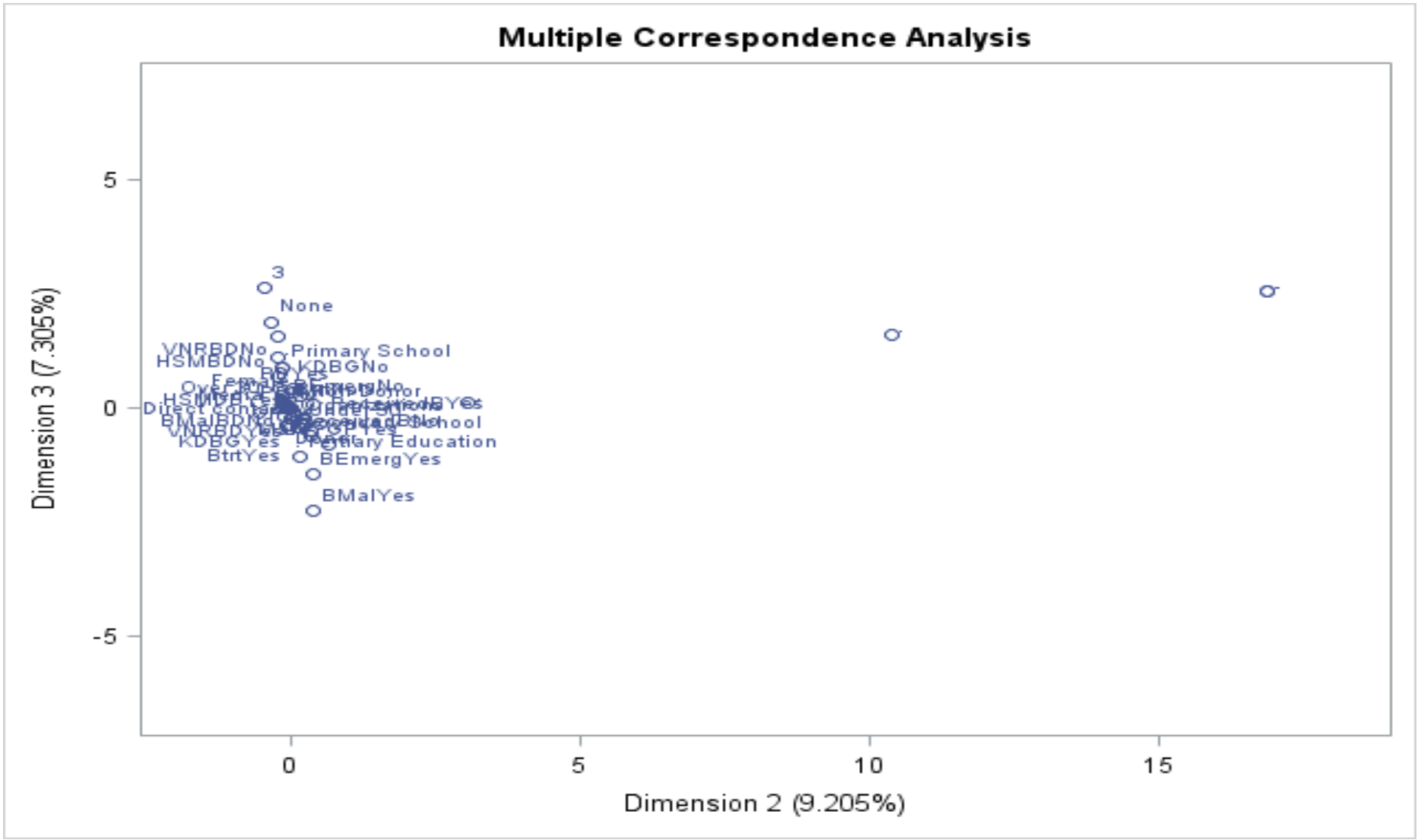


Figure 4.4: MCA MAP with adjustment of principal inertias along Dimension 2 and Dimension 3

4.5 Summary

CA is a rather useful technique as outlined by Costa et al. (2013) where it is claimed that no underlying distribution has to be assumed hence the technique can accommodate any type of categorical variable whether binary, ordinal or nominal; it provides key exploratory insights into the relationship between the data by representing the associations between variables in a low-dimensional space and it can also be used in pair with other methods such as multidimensional scaling, biplots and principal components analysis (PCA).

MCA generally looks at the associations among a set of two or more categorical variables. By the application of MCA, a decomposition of the collected data is obtained in order to study their structure and to visualize the association between the explanatory variables and donor status in the Mali population.

The variables in the MCA map appear to be clustered thus making it difficult to differentiate between the points, and those variables situated about the origin are not well represented in the map and do not add to the interpretation of the display.

The total Chi-square statistic is 14573.6 with 1681 degrees of freedom. Only 13.3 percent of the data is explained by the MCA map and the two dimensions account for 21.07 percent of the total association. The percentages explained by the MCA map are relatively small and tend to give a pessimistic view of the value of the MCA analysis.

By the use of the adjustment of inertias described in Section 4.3, estimates of the explained inertia are obtained which are much closer to the true values than the values obtained in MCA. The adjustment changes the scale on each dimension of the map, so as to best approximate the two tables of association between pairs of variables whilst leaving everything else in the solution intact. It can be seen that 44.48 percent of the data is explained by the adjusted MCA and the two dimensions account for 63.56 percent of the total association.

Chapter 5: Missing Data

In collecting survey data, partial responses are relatively common which often leads to incomplete data sets that include arbitrary patterns of missing data. There are different methods that can be used to handle incomplete cases in statistical analysis. The inadequate handling of these incomplete cases could possibly lead to biased and/or inefficient parameter estimates such as means or regression coefficients, and biased standard errors which result in incorrect confidence intervals and significance tests.

In evaluating a missing data method, a more robust method would be one that minimizes the bias caused by the missing data and making that bias as small as possible, maximizes the use of available information avoiding the discarding of any data and producing estimates that are efficient, and yielding good estimates of uncertainty.

There are three different types of missing data mechanisms that can arise when the data are being collected. It is important to distinguish between these different types of missing data mechanisms and according to Little & Rubin (1987), the implementation of any technique depends heavily on the mechanisms that lead to the missing values.

5.1. Missing Data Mechanism

Rubin (1976) discussed the different missing value processes and reasoned that they can be distinguished and termed in what follows. Consider subject i and let y_i be the univariate outcome of interest and let $\boldsymbol{x}_i$ be a $p \times 1$ vector of covariates corresponding to y_i.

5.1.1 Missing completely at random (MCAR)

The MCAR mechanism potentially depends on observed covariates, but not on observed or unobserved outcomes. In a logistic regression, for example, suppose that for subject i, y_i is completely observed and that some components of $\boldsymbol{x}_i$ are missing. If the probability of observing $\boldsymbol{x}_i$ is independent of y_i and is also independent of the values of $\boldsymbol{x}_i$ that are observed or would have been observed, then the missing values of $\boldsymbol{x}_i$ are MCAR. This means that the

missing values is independent on both observed and unobserved data and has no systematic cause for missing.

5.1.2 Missing at random (MAR)

The MAR mechanism depends on the observed outcomes and perhaps also on the covariates, but not further on unobserved outcomes, suggesting that the missing value could depend on the observed data but not on the unobserved data. Consider for example, the setting described above where y_i is completely observed whereas some components of $\boldsymbol{x}_i$ may be missing. If conditional on the observed data, the probability of observing $\boldsymbol{x}_i$ is independent of the values of $\boldsymbol{x}_i$ that would have been observed, but this probability is not necessarily independent of y_i and the observed values of $\boldsymbol{x}_i$, then the missing values of $\boldsymbol{x}_i$ are MAR. Further, the unconditional probability of observing $\boldsymbol{x}_i$ may depend on $\boldsymbol{x}_i$.

5.1.3 Not missing at random (NMAR)

The MNAR mechanism is operating, missingness does depend on unobserved outcomes, perhaps in addition to dependencies on covariates and/or on observed outcomes. Essentially the cause for missing values may depend on the observed data as well as the unobserved data.

Rubin (1976) further explains that under precise conditions, the missing data mechanism can be ignored when interest lies in inferences about the measurement process. This concept is termed ignorability.

According to Wu (2010), if the data is MCAR then the individuals in the sample with completely observed data can be viewed as a random subsample of the population and the complete case (CC) method, which discards all observations with missing values, is still valid. The disadvantage of this approach is the loss of efficiency due to discarding some of the data and this also results in a smaller sample size. Wu (2010) further explains that this method cannot be used for data that is not MCAR since the individuals with complete data cannot be treated as a random subsample of the population and the said approach could lead to biased results in such cases.

5.2 Ad Hoc Techniques to Deal with Missing Data

5.1.1 Deletion Procedures

Listwise deletion or complete case deletion

One of the more common methods employed as a default method by most statistical software packages, is to completely ignore observations with missing variable values and base the analysis only on those cases for which all measurements were recorded, i.e. a complete data set. The chief advantage of listwise deletion is convenience. This method, also known as complete case (CC) method, could lead to loss of efficiency as much information is likely to be lost as a result of excluding incomplete observations and could lead to biased estimates. The deletion of the incomplete data records can dramatically reduce the sample size and this reduction can reduce statistical power especially with small to moderate samples. This is the most appealing method due to its simplicity in application and interpretation of results, however, the dramatic attenuation in sample size and loss of information in the data set leads to possible bias and reduction in statistical power.

Pairwise deletion

Tsikriktsis (2005) explains that the pairwise deletion method (also known as available-case analysis) deletes cases only from the statistical analyses that require the information. Acock (2005) explains that with pairwise deletion, all available information is used whereby all participants that answered a pair of variables are used to estimate the covariance between those variables regardless of whether they answered other variables. One of the drawbacks of using such a method is that correlations or covariances may be biased since different parts of the sample are used for each statistic. Acock (2005) further points out that selecting a sample size using the correlation that has the most observations or that has fewer observations would be a mistake as it would either exaggerate statistical power or reduce statistical power respectively. According to Roth (1994), when compared to listwise deletion, pairwise deletion preserves much more information that would have likely been lost if listwise deletion were employed. Enders (2010) alludes that the primary problem with listwise deletion is that the data should be MCAR and if this assumption is violated or does not hold, it can produce distorted parameter estimates.

5.2.2 Single Imputation Methods

Single imputation methods impute data prior to the analysis. One of the key differences between single imputation and multiple imputation is that while single imputation generates a single replacement value for each missing data point, multiple imputation creates several copies of the data set and imputes each copy with different plausible estimates of the missing value.

Last Observation carried forward (LOCF)

Molenberghs & Verbeke (2005) explain that with the LOCF method, whenever a value is missing, the last observation value is substituted for that missing value. Further, this method can be applied to both monotone and non-monotone missing data patterns. The advantage of LOCF approach is one of convenience as it generates a complete data set. This technique of handling missing data is known to be specific to longitudinal designs. Despite its frequent use in medical studies and clinical trials, a growing number of empirical studies suggest that this approach is a poor strategy for dealing with longitudinal missing data (Cook et al., 2004; Liu & Gould, 2002; Mallinckrodt et al., 2003; Molenberghs et al., 2004; Shao & Zhong, 2004).

Mean substitution

The average value for the sample is imputed for missing observations of a particular variable. This simple method is known to perform well, especially if the data is normally distributed. Acock (2005) advises that mean imputation could possibly be the worst choice of handling missing data as it attenuates variance and can produce inconsistent bias when there is great inequality in the number of missing values for different variables.

Regression methods

The missing observation is imputed using the prediction taken from a multiple regression analysis. A detailed description of this technique is available in Enders (2010). This method can be biased as it overstates the correlation between variables and underestimates the variability of the data.

Hot-deck imputation

Enders (2010) have described the hot-deck imputation as a collection of techniques that impute the missing values with scores from "similar" respondents. This means that the missing value is replaced with an observed value taken from a matched observation based on the non-missing variables. Enders (2010) reports that this imputation technique generally preserves the univariate distributions of the data and does not attenuate the variability of the filled-in data to the same extent as other imputation techniques. However, hot-deck approaches are not well suited for estimating measures of association and can produce substantially biased estimates of correlations and regression coefficients (Brown, 1994; Schafer & Graham, 2002).

Each of these single imputation methods have been found to be inadequate in terms of accurately reproducing known population parameters and standard errors (Schafer & Graham, 2002). Schafer (1999) has reported that without special corrective measures, single imputation inference tends to overstate the precision because it omits the between-imputation component of variability and for joint inferences of multiple parameters, even small rates of missing information may seriously impair the said procedure.

Maximum likelihood also plays a central role in missing data analyses and is one of two approaches that methodologists currently regard as state of the art (Schafer & Graham, 2002).

5.3 Maximum Likelihood (ML)

According to SAS Global Forum (2012), with or without missing data, the first step in ML estimation is to construct a likelihood function. Suppose that there are n independent observations $(i = 1, \dots, n)$ on k variables $(y_{i1}, y_{i2}, \dots, y_{ik})$ and no missing data. The likelihood function is as follows

$$L = \prod_{i=1}^{n} f_i(y_{i1}, y_{i2}, \dots, y_{ik};\ \theta),$$

where $f_i(\cdot)$ is the joint probability or density probability function for observation i, and θ is the set of parameters to be estimated. They further suggest that, suppose for a particular observation i, the first two variables y_1 and y_2 have missing data that satisfy the MAR

assumption which is assumed to be ignorable. The joint probability for the observation is the probability of observing the remaining variables, y_{i3} through y_{ik}. If y_1 and y_2 are discrete, the joint probability summed over all possible values of the two variables with missing data is:

$$f_i^*(y_{i3}, \dots, y_{ik};\ \theta) = \sum_{y_1} \sum_{y_2} f_i(y_{i1}, \dots, y_{ik};\ \theta).$$

If the missing variables are continuous, integrals are used as follows

$$f_i^*(y_{i3}, \dots, y_{ik};\ \theta) = \int_{y_1} \int_{y_2} f_i(y_{i1}, \dots, y_{ik};\ \theta)\, dy_2\, dy_1.$$

To search for each observation's contribution to the likelihood function, sum or integrate over the variables that have missing data, to obtain the marginal probability of observing those variables that have actually been observed.

The overall likelihood is the product of the likelihoods for all observations. As an example, SAS Global Forum (2012), considers m observations with complete data and $n - m$ observations with missing data on y_1 and y_2. It follows that the likelihood function for the full data set becomes

$$L = \prod_{i=1}^{m} f_i(y_{i1}, \dots, y_{ik};\ \theta) \prod_{m+1}^{n} f_i^*(y_{i3}, \dots, y_{ik};\ \theta),$$

where the observations are ordered such that the first m have no missing data and the last $n - m$ have missing data. The likelihood can then be maximized to get ML estimates of θ using the usual applicable techniques.

5.4 Multiple Imputation

The fundamental aim of multiple imputation (MI) is to yield valid inferences for the statistical estimates of interest from the data imputed.

MI was formally introduced by Rubin (1978) and using his terminology, MI can be expressed in three distinct stages. In stage one, the values that are missing, are filled in m times to

generate M complete data sets. White et al. (2010) report that the unknown missing data are replaced by m independent simulated sets of values drawn from the posterior distribution of the missing data conditional on the observed data. In stage two the M complete data sets are analyzed using the standard procedures. What this actually means is that once the multiple imputations have been generated, each of the imputed data sets are analyzed separately using complete data methods with the retention of parameter estimates and standard errors from each analysis. The final stage involves combining the results from the m analyses into a single inference using Rubin's rules (Rubin, 1987), which are based on asymptotic theory in a Bayesian framework. White et al. (2010) explain that the combining of the variance-covariance matrix incorporates both within-imputation variability (uncertain about the results from one imputed data set) and between-imputation variability (reflecting the uncertainty due to the missing information).

In summary, the idea behind the MI procedure is to use the distribution of the observed data to estimate a set of plausible values for the missing data. Essentially, multiple data sets are created and analyzed independently but identically so that a set of parameter estimates are obtained and the estimates are finally combined into a single inference to obtain overall estimates, variances and confidence intervals.

An advantage of the MI technique is that it can be applied to virtually any kind of data or model and the analysis can be carried out using any conventional software. The imputed values are random draws rather than deterministic quantities, hence, a major shortcoming to this procedure is that it produces different results every time MI is implemented.

When the missing data is categorical, the appropriate methodology to impute the missing data is not clear. Multiple imputation by chained equations (MICE) has a desirable feature in its ability to handle different types of variables (continuous, binary, unordered categorical and ordered categorical). It generates imputations based on a set of imputation models and each variable is imputed using its own imputation model. MICE is also known as fully conditional specification (FCS) and (SRMI), sequential regression multivariate imputation (Raghunathan et al. 2001).

The FCS method does not rely on the assumption of multivariate normality. In general, conditional distributions are specified for each variable with missing values which is

conditional on all of the other variables in the imputation model. This method as applied by SAS contains two phases for each imputation: the preliminary filled-in phase followed by the imputation phase. At the first step, which is the filled-in phase, the missing values for all variables are imputed sequentially over the variables taken one at a time. This initial step provides starting values for the missing values at the imputation phase. At the next step, which is the imputation phase, the missing values for each variable are imputed sequentially for a number of burn-in iterations before imputation.

It is a flexible method that does not restrict the conditional distributions to being normal, hence, univariate regression models can be tailored appropriately to accommodate different types of variables, be it binary or ordinal. Schafer (1997) suggests using the MI approach, with the rounding of the imputed values to fit with the possible values of the variables and argues that the MI approach should work in most situations. However, Buuren et al. (2006) argue that a major advantage of the FCS approach is increased flexibility in model building. They explain that it is easy to incorporate constraints on the imputed values, work with different transformations of the same variable, account for skip patterns, rounding and so on.

Rubin (1987) shows that the relative efficiency of an estimate on m imputations to one based on an infinite number of imputations, is approximately $(1 + \lambda/m)^{-1}$, where λ is the rate of missing information. The percent efficiency achieved for various rates of missing information and values of m can be seen in Table 5.1.

Table 5.1: Percent efficiency for various rates of missing information and values of m

	λ				
m	0.1	0.3	0.5	0.7	0.9
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92

20	100	99	98	97	96

Another commonly used method for missing data is the expectation maximization algorithm. The condition for the algorithm to be valid is ignorabilty and hence MAR.

5.5 The Expectation-Maximization Algorithm

Expectation-maximization (EM) is a numerical algorithm that can be used to maximize the likelihood under a wide variety of missing-data models (Dempster et al., 1977). There are three steps involved in the EM algorithm, the initial step, the expectation or E step and the maximization or M step. An initial parameter is found using, for example, a complete case analysis, an available case analysis or a simple method of imputation. Molenberghs & Kenward (2007) advise that this estimate is possibly a biased estimate but sets the commencement of the algorithm. The E step finds the distribution of the data based on unknown values for the observed variables and the current estimate of the parameters while the M step replaces the missing data with the expected value. The E-step and M-step are iterated until the iterations converge.

One of the advantages of the EM algorithm is that it is guaranteed to reach convergence to a perhaps local maximum, however, this convergence is slow and the precision estimates are not automatically provided.

5.6 Subset Correspondence Analysis

CA as discussed in Chapter 4, is an exploratory tool that deals with the analysis of multivariate categorical data by representing associations between two or more categorical data in a low dimensional Euclidean space. The analysis of subsets of response categories is relevant to handling of missing data where the focus is either on the analysis of substantive responses in the presence of missing data or in the analysis of the missing data themselves.

Hendry et al. (2014) have shown that subset correspondence analysis (s-CA) can be applied to manage non-response whilst simultaneously retaining all observed data.

Greenacre & Pardo (2006b) propose a methodology that allows a direct analysis and interpretation of the non-response items, how they interrelate, how they relate to other response categories and to demographic covariates. They explain that this approach allows subset of categories to be analyzed and visualized by focusing the map on relationships within a chosen subset or between a subset and another subset.

An appealing feature of s-CA is that, as the full data matrix, ***N***, can be partitioned into a number of separate non-overlapping and all-inclusive matrices, so too is the inertia of the full matrix equal to the sum of the inertias of the separate matrices (Greenacre and Pardo, 2006a).

The description of s-CA involves applying it to a matrix ***N***, in the form of a contingency table. Further details can be found in Greenacre (1984), Greenacre & Parbo (2006a) and Greenacre & Blasius (2006).

The suggested methods to obtain the corresponding matrix $\mathbf{P}$, together with the marginal densities and diagonal matrices were discussed earlier and presented in Section 4.1. Using a variant of the same concept described in the said section, Greenacre & Parbo (2006a) describe s-CA as an adaptation to CA. They argue that the said theory can be applied to a subset of the table, maintaining the same row and column weighting as in classical CA but applied to a subset of profiles rather than a subset of the original table. This approach is said to avoid the recalculation of profiles for the selected subset. They further explain the theory from the row profile point of view which is presented next.

Suppose that $\mathbf{H}$ is a selected subset of the columns of $\mathbf{D}_{\mathrm{r}}^{-1}\mathbf{P}$. Further, suppose that the corresponding subset of the mean vector $\mathbf{c}$ is denoted by $\mathbf{h}$ where $\mathbf{h}$ is the weighted average of the rows of $\mathbf{H}$: $\mathbf{H}^T\mathbf{r} = \mathbf{h}$. Subset CA can then be defined as the weighted principal component analysis of $\mathbf{H}$ with row masses $\mathbf{r}$ in $\mathbf{D}_r$ and metric can be defined by $\mathbf{D}_h^{-1}$, where $\mathbf{D}_h$ is the diagonal matrix of $\mathbf{h}$. The subset CA solution is then obtained using steps 1 through 4 summarized as follows:

Step 1: $$\mathbf{S} = \mathbf{D}_r^{1/2}\mathbf{Y}\mathbf{D}_w^{1/2} \tag{5.1}$$

Step 2: $$\mathbf{S} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^T \tag{5.2}$$

Step 3: Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Delta}$

(5.3)

Step 4: Principal coordinates of columns: $\mathbf{G} = \mathbf{D}_w^{1/2}\boldsymbol{\Delta}\,\mathbf{V}$,

(5.4)

with $\mathbf{Y}$ equal to $\mathbf{H} - \mathbf{1}\mathbf{h}^T = (\mathbf{I} - \mathbf{1}\mathbf{r}^T)\mathbf{H}$, equal to present $\mathbf{D}_r$, and $\mathbf{D}_w$equal to $\mathbf{D}_h^{-1}$. The decomposed matrix is thus

$$\mathbf{S} = \mathbf{D}_r^{1/2}(\mathbf{I} - \mathbf{1}\mathbf{r}^T)\mathbf{H}\mathbf{D}_h^{-1/2},$$

and the row and column coordinates from Equation 5.3 and Equation 5.4 are as follows

$$\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\alpha \qquad \mathbf{G} = \mathbf{D}_h^{-1/2}\,\mathbf{V}\boldsymbol{\Delta}.$$

These co-ordinates are used to produce the graphical display of the points.

5.7 Application

The different techniques used to handle missing data are reviewed in the previous sections, however, it is the missing data pattern that will determine the appropriate methodology applicable for the respective data set. If the missing data pattern is monotone, the parametric regression method that assumes multivariate normality or the nonparametric method that uses propensity scores can possibly be used for imputation. An arbitrary missing data pattern makes use of a Markov chain Monte Carlo (MCMC) method (Schafer, 1997) that assumes multivariate normality or the FCS method also known as the MICE method.

Logistic regression was carried out in Chapter 3, to identify predictors of donor status in Mali. From Figure 5.1, it can be observed that the thirteen explanatory variables experienced some missing data. Also, it can be seen that approximately 20% (66) of subjects have incomplete data whilst approximately 2% (95) of values have missingness. In Table 5.2, it can be observed that missingness of variables ranged between 0.3% and 11.5% and the only variable

completely measured was the outcome variable which was donor status. The variables with the highest proportion of missing information are RB and SmsgsBD with approximately 5.6% and 11.5% missingness respectively.

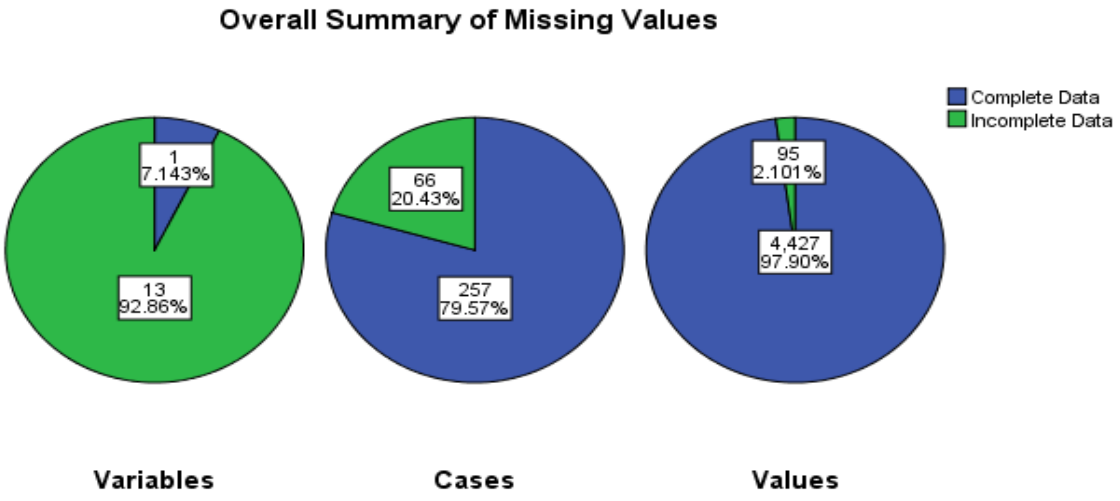


Figure 5.1: Summary of Missing Values

For each incomplete variable in excess of 2% missingness, an indicator variable was created and a Chi-square analyses was done to test if either the incomplete variable or its missingness was in relation to observed values of other variables. If the p-value between the indicator variable and an observed variable was less than 0.05, the null hypothesis was rejected and it could be concluded that the missingness was dependent on the missing variable. Since the variables SmsgsBD (11.5%), RB (5.6%) and BMal (2.5%) had missingness greater than 2%, indicator variables were created and chi-square tests carried out for each of them. The missingness for these variables is significantly related to at least one other variable in the data set and therefore can be assumed to be MAR. It cannot be ruled out, however, that there may exist some MNAR mechanism in the data. To ensure that the MAR assumption is plausible, it was necessary to include in the imputation model, the outcome variable; donor status, as well as all other possible likely predictors for the analysis.

Table 5.2: Number and Percentage of Missing Data

	Missing		
Variable	N	Percent	Valid N
SmsgsBD	37	11.5%	286

RB	18	5.6%	305
BMal	8	2.5%	315
AppWay_PD	6	1.9%	317
AppWay_VNRBD	6	1.9%	317
Edu_level	6	1.9%	317
HSMBD	5	1.5%	318
GP	2	0.6%	321
KDBG	2	0.6%	321
Age	2	0.6%	321
BEmerg	1	0.3%	322
Btrt	1	0.3%	322
Gender	1	0.3%	322

The missing data pattern grid produced by PROC MI in SAS can be viewed in Table 5.3 and indicates an arbitrary (non-monotonic) missing data pattern. In assessing the missing data patterns, each group represents a set of observations in the data set that share the same pattern of missing information. There appears to be 18 patterns for the specified variables. As can be seen in Table 5.3, 257 cases had no missing values in all variables, 28 cases had missing values in SmsgsBD (what do you think is the best way to spread messages about blood donation), 14 cases had missing values in RB (have you received blood donation), whilst 3 cases had missing values in RB and SmsgsBD. Also, it can be seen that 3 cases had missing values in AppWay_VNRBD (Do you think the appropriate way to give blood is voluntary non-remunerated blood donation). There was 1 case that had missingness in 7 of the explanatory variables i.e., Age, HSMBD (have you seen or heard messages about blood donation), SmsgsBD, Edu_level (highest level of education), AppWay_VNRBD, GP (blood donation is a good practice and everyone should donate) and AppWay_PD (Do you think the appropriate way to give blood is paid blood donation).

Results from Multiple Imputation

The MI technique, is generally used to impute missing data when the data is continuous or longitudinal. This iterative approach goes through a process of trying to find data that is missing and tries to simulate it so that it best fits the data that is available. As can be seen in Table 5.3, the data appears to have an arbitrary missing data pattern. Given the arbitrary missing data pattern and the use of categorical data, the FCS method appears to be appropriate for the imputation in this study. Analyses were carried out using SAS 9.4.

Table 5.3: Missing Data Pattern																
Group	Donor	Age	RB	HSMBD	Btrt	BEmerg	SmsgsBD	Gender	KDBG	Edu_level	AppWay_VNRBD	GP	BMal	AppWay_PD	Freq	Percent
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	257	79.57
2	X	X	X	X	X	X	X	X	X	X	X	X	.	X	6	1.86
3	X	X	X	X	X	X	X	X	X	X	.	X	X	.	3	0.93
4	X	X	X	X	X	X	X	X	X	.	X	X	X	X	1	0.31
5	X	X	X	X	X	X	X	X	.	X	X	X	X	X	1	0.31
6	X	X	X	X	X	X	X	.	X	X	X	X	X	X	1	0.31
7	X	X	X	X	X	X	.	X	X	X	X	X	X	X	28	8.67
8	X	X	X	X	X	X	.	X	X	X	X	X	.	X	1	0.31
9	X	X	X	X	X	X	.	X	X	.	X	X	X	X	1	0.31
10	X	X	X	X	.	.	X	X	.	X	X	X	.	X	1	0.31
11	X	X	X	.	X	X	.	X	X	.	X	X	X	X	1	0.31
12	X	X	X	.	X	X	.	X	X	.	.	X	X	.	1	0.31
13	X	X	X	.	X	X	.	X	X	.	.	.	X	.	1	0.31
14	X	X	.	X	X	X	X	X	X	X	X	X	X	X	14	4.33
15	X	X	.	X	X	X	.	X	X	X	X	X	X	X	3	0.93
16	X	X	.	.	X	X	X	X	X	X	X	X	X	X	1	0.31
17	X	.	X	X	X	X	X	X	X	X	X	X	X	X	1	0.31
18	X	.	X	.	X	X	.	X	X	.	.	.	X	.	1	0.31

The logistic regression carried out in Chapter 3 and the MCA method carried out in Chapter 4, employed the CC method and excluded any observations with missing values for the response or explanatory variables. The logistic regression analysis and the analysis carried out of the imputed data will be used for comparative purposes. It can be seen that only 257 cases were used in the analysis; in other words approximately 20% of the cases in the Mali data set were excluded from the analysis because of missing data. The reduction in sample size and statistical power could possibly be considered a problem and the CC analysis could also lead to biased estimates as discussed earlier. In other words, if the missing data mechanism is not MCAR, this CC method will introduce bias into the parameter estimates.

The FCS imputation method was selected in this study as it easily handles arbitrary missing data patterns with classification variables that need imputation. This procedure uses a multivariate imputation by chained equations assuming that a joint distribution exists for the data. Recall that the FCS method has the ability to handle different types of variables (continuous, binary, unordered categorical and ordered categorical) and generates imputations based on a set of imputation models and each variable is imputed using its own imputation model. Hence, the discriminant function method was used to impute all variables in this data set since the variables were all classification variables with either a binary or a nominal response. In this study, twenty sets of data were imputed.

The MIANALYZE procedure was used to generate inferences for the regression coefficients by combining the results over the twenty imputed data sets using Rubin's rules (Rubin, 1987).

Table 5.4 contains the variance information and includes the between, within and total variance for each parameter in the model. The within-imputation (W) variance is just a reflection of the normal sampling variability that is found in all analyses and was calculated as the average of the 20 squared standard error (SE) values that resulted from the analyses of the 20 imputed data sets. The between-imputation variance (B) is just a measure of uncertainty or added variability due to the data that is missing and is calculated as the sample variance of the regression parameters across the 20 imputed data sets.

This table also details the relative increase in variance due to missing data ranging from 0.008547 to 0.189131. Also, the fraction missing information is an estimate of how much

information about each coefficient is lost because of missing data and it ranges from 0.008482 to 0.161280. These reflect the impact of missing data among the variables used in this model. The total variance (T) calculated as $T = W + (1 + 1/20)B$, is the weighted sum of the within and between variances. As shown in Table 5.1 and based on the 20 imputed data sets, it can be seen that the relative efficiency which is greater than 0.99 for all effects, i.e., close to 1.0 for all effects, suggests that the 20 imputations are sufficient.

Table 5.5 contains parameter estimates that represents the averaged estimates with standard errors that are adjusted for both the sample design and the variability introduced by multiple imputation. Therefore, 95% confidence limits and t-tests are based on the fully corrected standard errors (Berglund, 2015). This table also confirms the results obtained in Chapter 3, based on logistic regression which employed the default CC method. Again we see a significant result between gender (p-value < 0.0001) and donor status as well as KDBG (p-value < 0.0001), i.e., knowledge about the different blood groups, and donor status. To reiterate the results discussed in Chapter 3, it follows that gender has a significant effect on the outcome of being a donor in the Mali population, where females were less likely to be blood donors.

Also, those individuals that had knowledge about the different blood groups were more likely to be donors as opposed to those that did not have knowledge on the different blood types.

Table 5.6 details the comparative results of estimates and standard errors for the CC method and the FCS imputation method. The comparison of the estimates reveal that parameter estimates in the FCS data are not very similar to those in the CC analysis although the overall significance and non-significance of variables in the data set remains unchanged. Additionally, the standard errors of the CC analysis are larger in all predictor variables and there are also noticeable differences in the magnitude of estimated coefficients. Notably, the standard errors from the analysis of the FCS imputed data were smaller than those from the CC analysis which resulted in greater accuracy of the estimated coefficients.

The increase in precision is an indication of superior efficiency and statistical power obtained for the analysis of the FCS imputed data.

Table 5.4: Variance Information

Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Intercept	0.040243	0.555077	0.597332	3796.9	0.076124	0.071228	0.996451
Gender	0.000217	0.018266	0.018493	125478	0.012459	0.012321	0.999384
Btrt	0.000818	0.064816	0.065675	111045	0.013254	0.013098	0.999346
BEmerg	0.001864	0.075996	0.077954	30131	0.025758	0.025176	0.998743
Age	0.000232	0.028487	0.028730	264547	0.008547	0.008482	0.999576
KDBG	0.000892	0.027691	0.028628	17743	0.033831	0.032833	0.998361
GP	0.000335	0.023028	0.023380	84034	0.015266	0.015060	0.999248
HSMBD	0.008884	0.120500	0.129829	3680.2	0.077415	0.072357	0.996395
HSMBD	0.010475	0.166574	0.177573	4952.2	0.066031	0.062319	0.996894
Edu_level	0.021443	0.167469	0.189984	1352.8	0.134442	0.119810	0.994045
Edu_level	0.008596	0.129832	0.138858	4496.9	0.069520	0.065417	0.996740
Edu_level	0.007524	0.078041	0.085941	2248.6	0.101227	0.092729	0.995385
AppWay_VNRBD	0.003054	0.055563	0.058770	6381.4	0.057715	0.054862	0.997264
AppWay_PD	0.005636	0.163050	0.168968	15488	0.036296	0.035149	0.998246
BMal	0.041592	0.230909	0.274581	751.08	0.189131	0.161280	0.992001
RB	0.006042	0.082906	0.089250	3760.5	0.076520	0.071575	0.996434
SmsgsBD	0.004521	0.033490	0.038237	1232.5	0.141761	0.125578	0.993760
SmsgsBD	0.006891	0.043274	0.050510	925.9	0.167201	0.145094	0.992798

A useful diagnostic that gives an indication of the stability of the estimates resulting from multiple imputation is the degrees of freedom (df)

Hendry et al. (2014) have reported that the df associated with multiple imputation is not the same as the df found in other statistical concepts and rather is a 'measure' of the ratio of the within-imputation variance and between-imputation variance. It can be observed from Table 5.5 that the df for the FCS imputations ranged from 751.08 to 264547 in this study. This being large compared to the number of imputed sets, is an indication that the estimates have been stabilized and can be trusted.

Results from Subset Correspondence Analysis

The analysis of MCA to the full data set in Chapter 4, employed the default CC method which led to a reduction in the sample size from 323 to 257, that is, a loss of 66 cases, hence, approximately 20 percent of the data was missing from the analysis. The category points contributing to the graphical representation was difficult to interpret as there was a clustering of variables about the origin. The analysis of a subset of response categories is thought to be of particular relevance to the handling of missing data. This approach allows the inclusion and exclusion of the missing data in a way that incurs no loss of data. It enables the analysis of the non-responses themselves to understand how they are correlated between items as well as the analysis of different subsets of categories.

Subset analysis was applied to this study as an alternative approach to handle the missing data. All analyses were done using R 3.22 (R Core Team, 2015) in R Studio 0.99.489 (RStudio, 2015) with the following packages: FactoMineR (Husson et al., 2015), ca (Nenadic & Greenacre, 2007) and graphics (R Core Team, 2015). A separate missing category was introduced for each variable with a non-response, i.e., all thirteen explanatory variables included a missing variable category. All categories representing the missing data were excluded from the subset for analysis. For the graphical displays, principal co-ordinates were used to plot the variables.

Table 5.5: Parameter Estimates and Standard Errors

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	0.184865	0.772873	-1.33042	1.70015	3796.9	-0.244490	0.437061	0	0.24	0.8110
Gender	-0.712183	0.135990	-0.97872	-0.44564	125478	-0.730566	-0.681270	0	-5.24	<.0001
Btrt	0.154816	0.256272	-0.34747	0.65710	111045	0.097639	0.197331	0	0.60	0.5458
BEmerg	-0.072319	0.279202	-0.61957	0.47493	30131	-0.135503	0.011202	0	-0.26	0.7956
Age	0.118147	0.169500	-0.21407	0.45036	264547	0.090123	0.145493	0	0.70	0.4858
KDBG	0.692856	0.169198	0.36121	1.02450	17743	0.624204	0.738447	0	4.09	<.0001
GP	0.159434	0.152905	-0.14026	0.45913	84034	0.111283	0.192392	0	1.04	0.2971
HSMBD	0.469691	0.360318	-0.23675	1.17613	3680.2	0.314368	0.627198	0	1.30	0.1925
HSMBD	-0.214739	0.421394	-1.04086	0.61138	4952.2	-0.373326	-0.057251	0	-0.51	0.6104
Edu_level	0.876771	0.435872	0.02171	1.73183	1352.8	0.614690	1.073713	0	2.01	0.0445
Edu_level	0.107296	0.372637	-0.62325	0.83785	4496.9	-0.034791	0.379567	0	0.29	0.7734
Edu_level	-0.395249	0.293157	-0.97014	0.17964	2248.6	-0.538453	-0.173872	0	-1.35	0.1777
AppWay_VNRBD	0.356177	0.242426	-0.11906	0.83141	6381.4	0.228570	0.445249	0	1.47	0.1418
AppWay_PD	0.407372	0.411057	-0.39835	1.21309	15488	0.254074	0.556546	0	0.99	0.3217
BMal	0.106674	0.524005	-0.92201	1.13536	751.08	-0.400211	0.465813	0	0.20	0.8387
RB	0.517640	0.298747	-0.06808	1.10336	3760.5	0.338068	0.603875	0	1.73	0.0832
SmsgsBD	-0.106952	0.195544	-0.49059	0.27668	1232.5	-0.229335	0.064120	0	-0.55	0.5845
SmsgsBD	0.243671	0.224743	-0.19739	0.68474	925.9	0.098110	0.423050	0	1.08	0.2785

Table 5.6: Complete Case and FCS Imputation Estimates and Standard Errors

	CC Method		FCS Imputation Model	
Parameter	**Estimate**	**Standard Error**	**Estimate**	**Standard Error**
Intercept	-1.5860	1.2986	0.184865	0.772873
Age (Ref = > 30) ≤ 30	0.00262	0.3855	0.118147	0.169500
Gender (Ref = Male) **Female**	-1.6289	0.3121	-0.712183	0.135990
RB (Ref = No) **Yes**	1.2199	0.8227	0.517640	0.298747
KDBG (Ref = No) **Yes**	1.3801	0.3619	0.692856	0.169198
HSMBD (Ref = Do not remember) **Yes**	-0.2487	1.0620	0.469691	0.360318
No	-1.1216	1.1529	-0.214739	0.421394
Edu_level (Ref = Tertiary education) **Never went to school**	1.3237	0.6581	0.876771	0.435872
Primary education	0.6658	0.6220	0.107296	0.372637

	CC Method		FCS Imputation Model	
Parameter	**Estimate**	**Standard Error**	**Estimate**	**Standard Error**
Secondary education	0.1364	0.3653	-0.395249	0.293157
Btrt (Ref = No) **Yes**	0.6215	0.5764	0.1548	0.2563
BEmerg (Ref = No) **Yes**	-0.3569	0.6972	-0.0723	0.2792
BMal (Ref = No) **Yes**	0.9433	1.4516	0.1067	0.5240
SmsgsBD (Ref = Direct contact) **Media**	0.0944	0.3537	-0.1069	0.1955
Organizations	0.1170	0.4090	0.2437	0.2247
GP (Ref = No) **Yes**	0.2228	0.3608	0.1594	0.1529
AppWay_VNRBD (Ref = No) **Yes**	1.1565	0.5936	0.3562	0.2424
AppWay_PD (Ref = No) **Yes**	0.1219	1.1004	0.4074	0.4111

Figure 5.2 shows the MCA map of the response categories, omitting the non-response categories. The total inertia as explained before, measures the variability in the data and the examination of the percentage of the total inertia represented on each axis, makes it possible to identify the relative importance of the axes and the amount of variability present in the data. The percentage of inertia accounted for by the first two dimensions is 63.9 percent as shown in Table 5.2 which is much higher than the MCA map in Figure 4.1, without taking into consideration the missing data. Those variables clustered about the origin in the graphical display of Figure 5.2, are not well represented and do not add to the interpretation of the display. To identify the strength of association between the points, the smaller the angle between the points, the closer is the association. It can be observed that the category Donor (1) is situated above the horizontal axis and is strongly associated with Gender2 and KDBG1.

This implies that there is a strong association between the outcome of a donor being male and having knowledge about the different type of blood groups in the Mali population. Further, the horizontal axis separates the categories for the best way to spread messages about blood donation with categories SmsgsBD (1) and SmsgsBD (2) above the horizontal axis and closely related to Donor (1) whilst SmsgsBD (3) is below the horizontal axis. Accordingly, thinking the best way to spread messages about blood donation is via the media or organizations, is closely related to the outcome of being a donor. The strongest association to Donor (0), i.e., non-donor, is Gender (1), KDBG (2), Age (2) and SmsgsBD (3). This implies that females, over the age of thirty without any knowledge about the different type of blood groups are more likely to be non-donors. Also, strongly associated with the non-donor category, were individuals that thought the best way to spread messages about blood donation is through direct contact. Other variables fairly related to non-donors include AppWay_PD (1) and Edu_level (2), which imply that non-donors are fairly associated to individuals with primary school education and individuals that thought that paid donation is the appropriate way to donate blood. The variables BMal1, BEmerg1, Btrt1 and GP1 are separated from the other variables and appear on the negative side of axis 1. These are positive responses to the usage of blood required for donation.

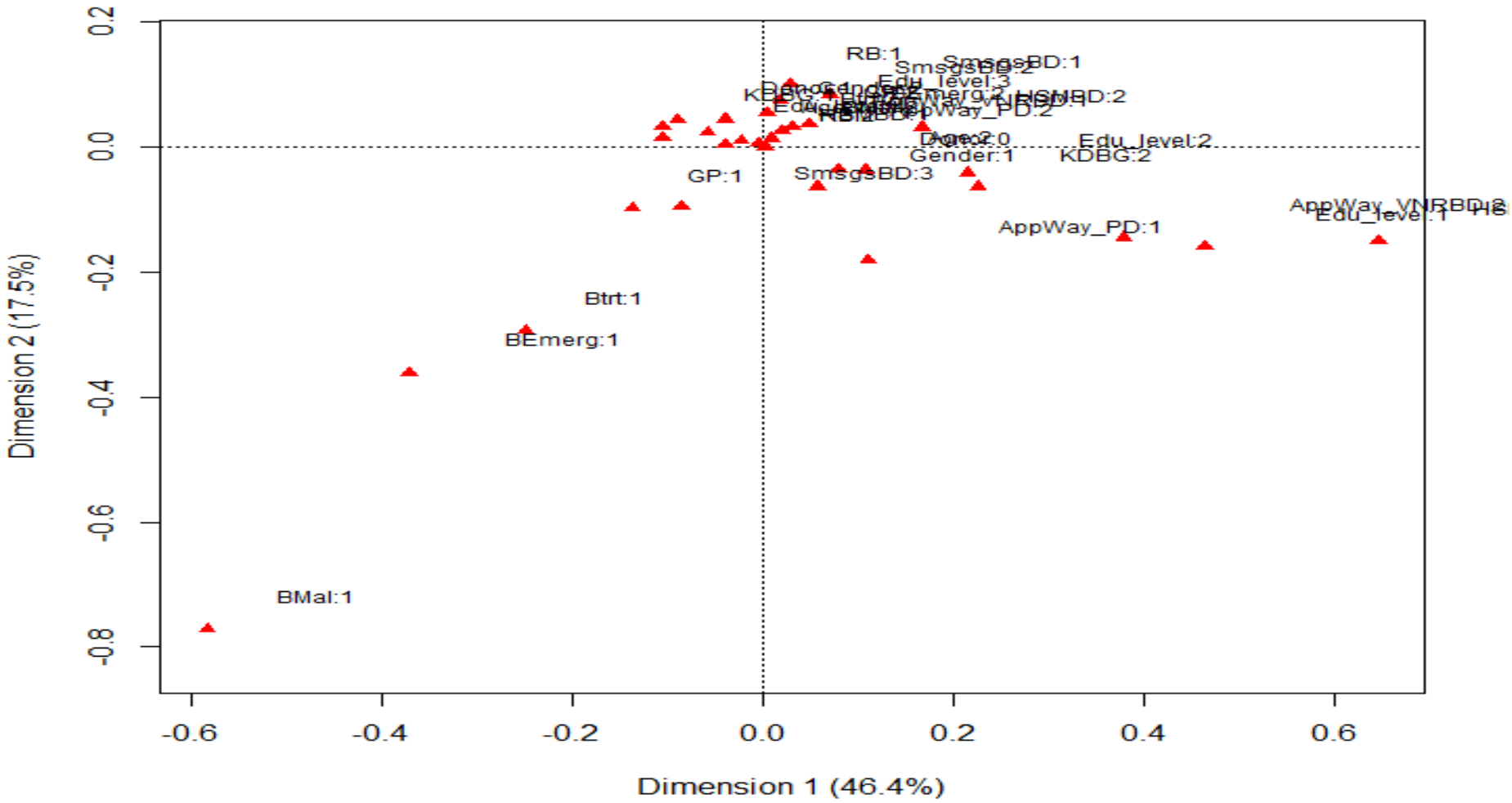


Figure 5.2: Subset MCA map of response categories omitting the non-response categories.

Table 5.7: Principal Inertias (eigenvalues) for response categories only

Dim	Value	%	cum%	scree plot
1	0.009608	46.4	46.4	**********
2	0.003635	17.5	63.9	******
3	0.000781	3.8	67.7	*
4	0.000511	2.5	70.1	*
5	0.000178	0.9	71.0	
6	4.1e-050	0.2	71.2	
7	1.2e-050	0.1	71.3	
8	6e-06000	0.0	71.3	
Total: 0.020719				

These responses suggests that the blood required for transfusion is used to correct malnutrition, for emergency cases and disasters, and to treat diseases/malaria. A subset MCA map was performed of the non-response categories alone, which was originally omitted, and the resulting map is shown in Figure 5.3. The inertia associated with this subspace is presented in Table 5.8. The first two dimensions account for 79.6 percent of the inertia. Hence, it follows that the two-dimensional figure accounts for 79.6 percent of the variability in the data and 20.4 percent is not accounted for. There appears to be a clustering of category points about the origin. The variables BMal (3), KDBG (3), Btrt (3), and BEmerg (3) lie below the horizontal axis and are separated from the other non-response items. The origin represents the average non-response point for all thirteen variables hence, categories to the right have more than average non-responses and categories to the left have fewer than average. This implies that variables BMal (3), KDBG (3), Btrt (3) and BEmerg (3) have fewer non-responses. Due to the clustering of variables, it is difficult to determine from the graphical display (Figure 5.3) as to which variables have a higher than average response.

Table 5.8: Principal Inertias (eigenvalues) for non-response categories only

Dim	Value	%	Cum%	Scree plot
1	0.071812	51.2	51.2	**********
2	0.039798	28.4	79.6	*******
3	0.005916	4.2	83.8	*
4	0.005455	3.9	87.7	*
5	0.004691	3.3	91.1	*
6	0.004403	3.1	94.2	*
7	0.003781	2.7	96.9	*
8	0.002419	1.7	98.6	
9	0.000882	0.6	99.3	
10	0.000655	0.5	99.7	
11	0.000362	0.3	100.0	
12	00000000	0.0	100.0	
13	00000000	0.0	100.0	
Total:	0.140175	100.0		

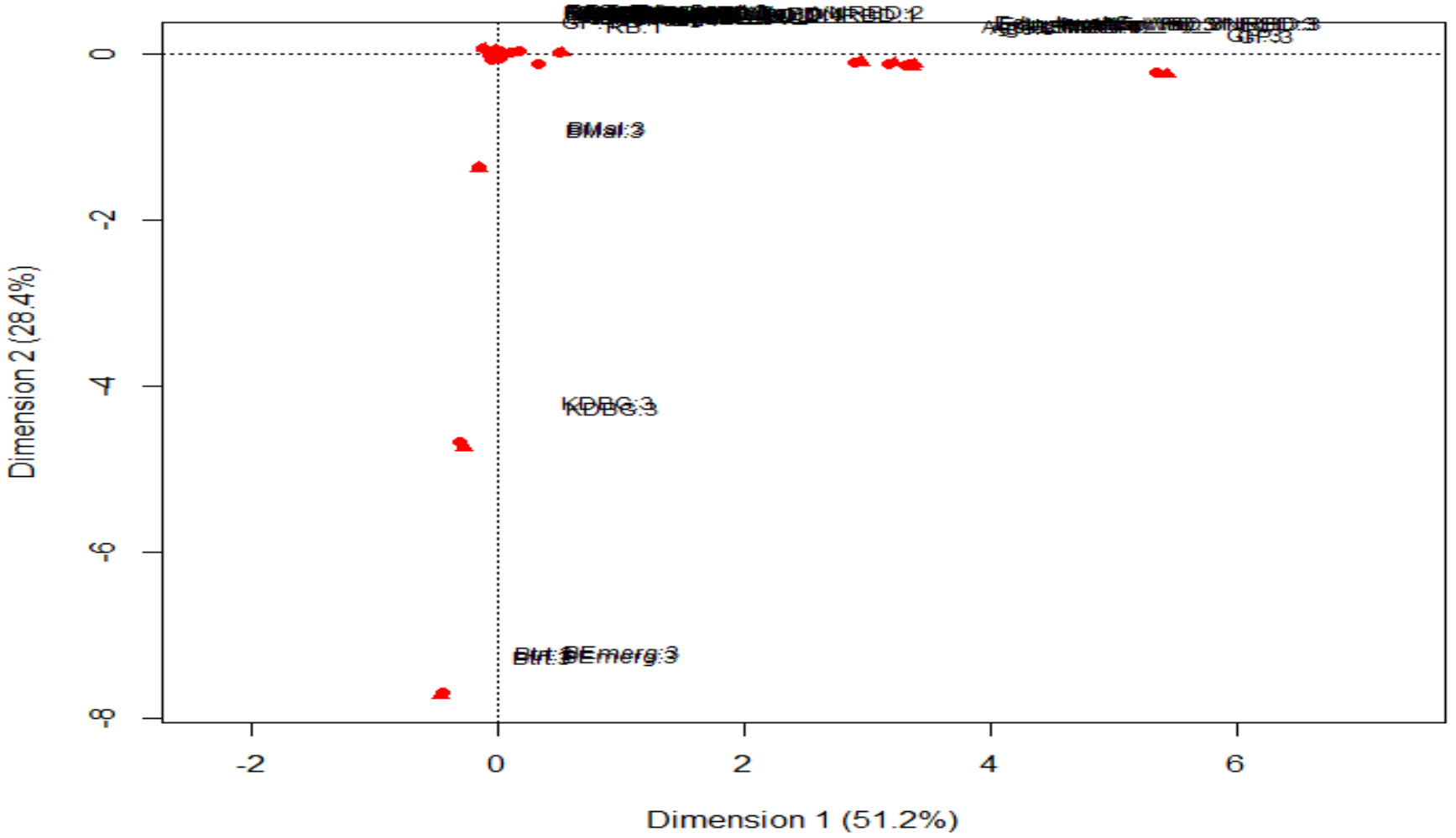


Figure 5.3: Subset MCA map of non-response categories only.

5.6 Summary

Missing data in surveys and questionnaires are quite common as sometimes survey respondents may choose to leave out one or many unanswered items, unintentionally or perhaps because they feel inhibited and are not comfortable answering items of a sensitive topic. It is important to address missing data and account for incomplete observations, as it arises in almost all real world investigations.

The importance of distinguishing between the different missing data mechanisms is highlighted since the reasons for the missing data can affect the underlying assumptions of the statistical modelling techniques employed.

The methods discussed include the removal of cases with incomplete data or the filling in of missing values via imputation. The focus here was on the issues related to missing data within the context of categorical data since the data in this study included all classification predictors.

It will be much easier to just remove those observations with incomplete data instead of going through the process of imputation as discussed in Section 5.2. This approach will seem reasonable if the deleted cases form a relatively small part of the data set in its entirety. However, recall from Section 5.1 on missing data mechanism that Wu (2010) argued that if the data is MCAR then the individuals in the sample with completely observed data can be viewed as a random subsample of the population and discarding observations with missing values will then only render valid inferences. When the discarded cases differ systematically from the rest, estimates may be seriously biased (Schafer, 1999).

In this study all thirteen explanatory/predictor variables had some missing data and the variables with the highest proportion of missing information are RB and SmsgsBD with approximately 5.6% and 11.5% respectively. The implementation of the CC analysis resulted in an omission of approximately 20% of cases with missing data values. The dropping of all these observations and fitting a model to only the complete cases could possibly be inefficient and potentially biased if the missing data mechanism is not MCAR.

The FCS approach appears to be a powerful and convenient method that easily handles arbitrary missing data patterns with classification variables that need imputations for multivariate missing data. Based on 20 imputed data sets, the relative efficiency is close to 1.0 for all effects, thereby suggesting that the 20 imputations are sufficient.

Results obtained from the FCS imputed model confirm those found in Chapter 3 that were based on logistic regression which employed the default CC method. There is a significant difference between gender (p-value < 0.0001) and the outcome of being a donor which implies that females were less likely to be blood donors than their male counterparts in the Mali population. Also, those individuals that had knowledge about the different blood groups (p-value < 0.0001), were more likely to be donors as opposed to those that did not have knowledge on the different blood types.

The comparison of the CC analysis and the FCS method reveal that parameter estimates in the FCS data differs to those in the CC analysis although the overall significance and non-significance of variables in the data set remains unchanged. The standard errors from the analysis of the FCS imputed data were smaller than those from the CC analysis which resulted in greater accuracy of the estimated coefficients. The increase in this precision is an indication of superior efficiency and statistical power obtained for the analysis of the FCS imputed data.

The traditional approaches used to handle missing data can often lead to biased estimates or either reduce or exaggerate statistical power which could result in invalid conclusions as argued by different authors reported throughout this chapter.

Another relatively new technique that can be used to explore the relationships between categorical variables that suffer from missingness is s-CA. CA is discussed in detail in Section 4.1, and can be adapted to deal with the analysis of a subset of categories to manage non-response whilst simultaneously retaining all observed data.

The s-CA was used as an exploratory tool to seek interrelationships between variable categories and to identify those variable categories that are associated with donor status. A good feature of s-CA is that by restricting the analysis to subsets of categories, a subset can be visualized separately with better quality than is the case with the

complete MCA analyzed in Chapter 4 (Figure 4.1, Figure 4.2, Figure 4.3, and Figure 4.4). This approach with the added advantage of analyzing the response categories only, allowed a clearer display of the points thus enabling the exploration of the relationships between the relevant variables. Association between these variables and donor status confirm the results found in earlier chapters which showed that gender and knowledge about the different type of blood groups are strongly associated with blood donor status.

Chapter 6: Conclusion

Blood donation is a highly relevant issue worldwide and factors that motivate individuals to donate blood makes it possible to determine which individuals are likely to remain blood donors or to become new or prospective donors. Voluntary non-remunerated blood donation is generally associated with safer blood supplies in terms of TTI's and that is one of the reasons the WHO recommends that blood and blood components be collected only from voluntary non-remunerated blood donors (Dhingra, 2002).

In poor resource countries, the safe and adequate supply of blood still remains very much of a challenge. Mali is ranked as one of the top 10 poorest countries in the world and access to medical supplies is fairly limited. The country experienced a high maternal mortality rate due to post-partum haemorrhaging and the root cause of which was lack of access to safe blood. Before the collaboration with the American Red Cross, the Milennium Cities Initiative, Safe Blood for Africa and the Mali Ministry of Health, there was only one poorly equipped blood bank in the Capital of Bamako to hold blood for the country's estimated population of 16 014 000 as of mid-2012.

The main objective of this study was to develop a theoretical framework to better understand the attitudes toward blood donation and transfusion in Mali. It also aimed to identify factors that motivate and deter blood donation in Mali, as well as to identify interventions to improve the supply of blood transfusion.

Recall that data was collected from 323 individuals across three different sites in Mali (Bamako Ségou & Kita). Descriptive statistics reveal that more than 50% of individuals responded as non-donors. The male population responded as majority donors at 58.8%. Approximately 19% of individuals responded as recent family replacement donors, 43% as lapsed donors (i.e., donation before 2012), about 22% as voluntary non-remunerated blood donors and 16% as regular donors for several years. Also, more than 50% of non-donors reported their intention to donate blood in the future whilst a small number reported a lack of intent to become future blood donors. The relationships between the

measured variables and donor status are of interest in this study and further statistical analyses were investigated.

For a dichotomous dependent variable, logistic regression proves to be a powerful analytical technique. In this case the response variable of interest is binary, indicating whether an individual is a donor or non-donor. The common method used to analyze binary response data is logistic regression which makes use of the cumulative logistic function, however, it can be noted that the probit model and the complementary log-log model can also be used to model such data via the appropriate link functions. After fitting the data to the model and investigating the effect of all three approaches, it can be seen that the logit link appears to be the most reasonable approach to fit the data.

The effectiveness of the model was supported by significance test of the model and of each predictor, measures of association for predictive accuracy, goodness of fit tests and residual diagnostics.

Results from the analysis are presented in Section 3.9. There were only two significant factors to donor status in the Mali population. Gender had a significant effect on the outcome of being a donor in this population, hence the odds of a female being a donor is 0.196 times that of a male, i.e. females are less likely to be blood donors in the Mali region. Also, the odds of being a donor for an individual that had knowledge about the different type of blood groups is significantly different from an identical individual that did not have knowledge on the different type of blood groups. There were no other significant factors to donor status and the inclusion of any or all of the possible interaction terms did not improve the fit of the model and hence was not included in further analyses.

MCA was used as an exploratory technique to describe the pattern of relationships of the explanatory variables using geometrical methods to locate each variable as a point in a low dimensional space. By the application of the MCA, the association between the investigated parameters and blood donor status was visualized. The variables were clustered, making it difficult to differentiate between the points and interpret the relationships between them. Also, the variables situated about the origin were not well represented in the MCA map and did not add to the interpretation of the display. As is

usually the case with survey data, it is not uncommon to find non-responses from individuals partaking in the survey. There are numerous reasons for non-response items however, most often than not, most individuals feel inhibited in their response to sensitive questions. The different ways to deal with missing data were explored and the two approaches applicable to the categorical data presented in this study were investigated. One such method saw the imputation of values using the FCS approach whilst the other method was s-CA which is a relatively new approach to handling of missing data. This enabled the exploration of the relationships between the relevant variables.

Selecting an imputation method to handle the missing data values will depend on the data mechanism, structure of the attributes and the given data set. The FCS approach has the added advantage of taking into account arbitrary missing data patterns and different types of variables (continuous, binary, unordered categorical and ordered categorical). The application of the FCS approach to this study, confirmed the results obtained in the logistic regression analysis. This implies that gender and the knowledge about the different type of blood groups, had a significant effect on donor status in the Mali population. All other exploratory variables did not have a significant effect on the said population. Although the overall significance and non-significance of variables in the data set remained unchanged, the standard errors from the analysis of the FCS imputed data were smaller than those from the CC analysis and this resulted in a greater accuracy of the estimated coefficients. The increase in this precision indicated the superior efficiency and statistical power obtained for the analysis of the FCS imputed data. However, this is not the only technique used to deal with missing data values and may also not be the best technique available for a given problem with categorical variables. A relatively new technique which is not yet widely adopted as a tool to handle missing data is s-CA.

In the visualization of the complete MCA presented in Section 4.4, with and without adjustment to inertias (Figure 4.1, Figure 4.2, Figure 4.3 and Figure 4.4), the points representing the missing data were strongly associated, and overall, the other categories were clustered about the origin. It was difficult to understand the points and interpret the relationships between the response categories as noted earlier. With s-

CA the full data matrix can be partitioned into smaller mutually exclusive sub-matrices, with regard to the respective decomposition of the total inertia. Greenacre & Blasius (2006) explain that this approach maintains the geometry of the masses and chi-square distances of the complete MCA, with the only difference being that there is no need to re-express the elements of the subset with respect to their own totals, but maintain their profile values with respect to the totals of the complete data set. A separate missing category was introduced for all thirteen explanatory variables, since each of these variables suffered from missingness. The overcrowding of the graphical display and domination of the missing categories can occur and makes the identification of the relevant inter-variable association difficult. This can be alleviated by analyzing the subset of data that excluded the missing categories. Hence, CA was applied to the sub-matrix of response categories only, which allowed a clearer display of the points presented in Figure 5.2. Association between variables confirm the results found in earlier chapters which showed that gender and knowledge about the different type of blood groups are strongly associated with blood donor status in the Mali population.

A subset analysis of the non-response categories alone, which was originally omitted, is presented in Figure 5.3. The variables BMal (3), KDBG (3), Btrt (3), and BEmerg (3) are separated from the other non-response items. The origin is a representative of the average non-response point for all thirteen variables, hence, the categories to the right of the origin have more than average non-responses and categories to the left have fewer than average. This implies that variables BMal (3), KDBG (3), Btrt (3) and BEmerg (3) have fewer non-responses. Due to the clustering of variables, it is difficult to determine from the graphical display (Figure 5.3) as to which variables have a higher than average response.

MI by the FCS approach and s-CA analysis are two very different methods that have been used to identify associations between donor status and exploratory variables in the presence of missingness. Handling missingness by s-CA is a novel approach which has not yet been adopted widely in dealing with missing data. This approach requires non-negative and categorical data which is easily achievable through transformation. While MI works with fitting data to a pre-assumed model, under the assumption of missing data mechanisms and distributional assumptions, s-CA has no restriction with missing

data mechanisms and distributional assumptions. The data is decomposed to reveal trends and relationships among categories where no model assumption is required.

The overall results produced from the statistical methods employed in this study did not differ substantially and the associations found between donor status and selected factors were consistent across these methods.

It is crucial to understand which factors motivate individual's to donate blood and which factors deter individuals from blood donation and transfusion. Further, it is essential to promote awareness of the need for blood, and to have educational programs, good communication from blood banks and health services, and also the endorsement of mass media. Donors need to be made aware that it is completely safe to donate blood and that these generous donations could alleviate unnecessary death from the lack of access to safe blood.

In this study there appeared to be only two significant factors to blood donor status in the region. To reiterate the results discussed, females were less likely to be donors in the said population and individuals that had knowledge about the different type of blood groups were more inclined to be donors.

Nevertheless, accurately predicting motivational factors and blood donation behaviour remains problematic, hence, continued research attempts are needed to identify which variables are necessarily the best predictors of blood donation. There is a high willingness of intent for future blood donation (approximately 90%) from the non-donor category and this should be considered as an opportunity for future mobilization initiatives in the region. However, recall from Section 1.2 that SBFA commenced its intervention in the country in 2012 but was interrupted by the civil strife until mid-2013. This could be a possible limitation to the study and could have affected an individual's willingness to become a future blood donor.

Further studies to understand the root causes among non-donors in the Mali population, as well as the reasons behind failure to retain regular blood donors are recommended. Also of interest to future research would be to include additional variables in the study or to extend the donor category of donor versus non-donor, to more than two

categories. The inclusion of the additional response categories could prove to be a viable direction of research in this area. This will entail the extension of the binomial logistic regression to the multinomial logistic regression which is used when the dependent variable has more than two nominal categories.

The understanding of these factors could be the key to unlocking the misconception and misinformation of the blood donation process, whilst factors that encourage the donation of blood can be identified and evaluated so as to improve the retaining and recruitment of blood donors in the Mali population.

Bibliography

Acock, A. C. (2005). Working with Missing Values. *Journal of Marriage and Family*, 1012-1028.

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed.). New Jersey: John Wiley & Sons.

Aguirre-Garcia, M. S., & Aldamiz-echevarria, C. (2014). A Behavior Model for Blood Donors and Marketing Strategies to Retain and Attract Them. *Rev.Latino-Am.Enfermagem, 22(3)*, 467-75.

Ahmed, N., Dawson, M., Smith, C., & Wood, D. (2007). *Biology of Disease.* New York: Taylor & Francis.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 716-723.

Allison, P. D. (1999). *Logistic Regression Using the SAS System: Theory and Application.* Cary, NC: SAS Institute Inc.

Bates, I., Chapotera, G. K., McKew, S., & van den Break, N. (2008). Maternal Mortality in sub-Saharan Africa: The Contribution of Ineffective Blood Transfusion Services. *BJOG, 115*, 1331-1339.

Benzécri, J. (1979). Sur le Calcul des taux d'inertie dans l'analyse d'un questionaire, Addendum et erratum á [BIN.MULT.]. *Cah Anal Donnees*, 377–378.

Berglund, P. A. (2015). Multiple Imputation using the Fully Conditional Specification Method: A Comparison of SAS, Stata, IVEware and R. *SAS Institute Inc*.

Bloch, E. M., Vermeulen, M., & Murphy, E. (2012). Blood Transfusion Safety in Africa: A Literature Review of Infectious Diseasesand Organizatiional Challenges. *NIH Pubic Access*, 164-180.

Brown, R. L. (1994). Efficacy of the Indirect Approach for Estimating Structural Equation Models with Missing Data: A Comparison of Five Methods. *Structural Equation Modeling: A Multidisciplinary Journal, 1*, 287-316.

Buuren, S. V., Brand , J. P., Groothuis-Oudshoorn, G. M., & Rubin, D. B. (2006). Fully Conditional Specification in Multivariate Imputation. *Journal of Statistical Computation and Simulation*, 1049-1064.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regression Through Graphics.* New York: Wiley & Sons.

Cook, R. J., Zeng, L., & Yi, G. Y. (2004). Marginal Analysis of Incomplete Longitudinal Binary Data: A Cautionary Note on LOCF Imputation. *Biometrics, 60*, 820-828.

Costa, P. S., Santos, N. C., Cunha, P., Cotter, J., & Sousa1, N. (2013). The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables in Healthy Ageing. *Journal of Aging Research, 2013*. doi:10.1155/2013/302163

Dempster, A. P., Laird, N. M., & Donald, R. B. (1977). Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 1-38.

Dennis, L. K., & Anthony, S. F. (2010). *Harrison's Infectious Diseases.* New York: McGraw-Hill.

Der, G., & Everitt, B. (2002). *A Handbook of Statistical Analyses using SAS.* Boca Raton: Chapman & Hall/CRC.

Dhingra, N. (2002). Blood Safety in Developing World and WHO Initiatives. *Vox Sanguinis*, 173-177.

Duggan, J. M., & Duggan, A. E. (2006). *The Epidemiology of Alimentary Diseases.* Dordrecht: Springer.

Enders, C. K. (2010). *Applied Missing Data Analysis.* New York: The Guilford Press.

Endres , J. (2013). *Africa Survey: Africa in Figures.* South Africa: Good Governance Africa.

Erhabor, O., Adias, T. C., & Mainasara, A. S. (2013). Provision of Safe Blood Transfusion Services in Low Income Setting in West Africa. Case Study of Nigeria. *Advances in Medicine and Biology, 59*, 1-58.

Finney, D. J. (1952). *Probit Analysis.* Cambridge: Cambridge University Press.

Gillespie, T. W., & Hillyer, C. D. (2002). Blood Donors and Factors Impacting the Blood Donation Decision. *Transfusion Medicine Reviews, 16*, 115-130.

Graham, J. W. (2012). *Missing Data: Analysis and Design.* New York: Springer.

Greenacre , M. J. (1994). Multiple and Joint Correspondence Analysis. (M. J. Blasius, Ed.) *Correspondence Analysis in the Social Sciences*, 141-61.

Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis.* London: Academic Press Inc.

Greenacre, M. J., & Pardo, R. (2006a). Multiple Correspondence Analysis of Subsets of Response Categories. In M. GREENACRE, & J. BLASIUS, *Multiple Correspondence Analysis and Related Methods.* Boca Raton: Chapman & Hall/CRC.

Greenacre, M. J., & Pardo, R. (2006b). Subset Correspondence Analysis Visualizing Relationships Among a Selected Set of Response Categories from a Questionnaire Survey. *Sociological Methods and Research, 35*, 193-218.

Greenacre, M., & Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods.* New York: Chapman & Hall/CRC.

Hendry, G. M., Naidoo, R. N., Zewotir, T., North, D., & Mentz, G. (2014). Model Development Including Interactions with Multiple Imputed Data. *BMC Medical Research Methodology, 14:136*.

Hinkley, D. V. (1985). Transformation Diagnostics for Linear Models. *Biometrika*, 487-96.

Hosmer, D. W., & Lemeshow, S. (1980). A Goodness-of-Fit Test for The Multiple Logistic Regression Model. *Communications in Statistic*, 1043-1069.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied Logistic Regression.* New York: John Wiley & Sons.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* New York: John Wiley & Sons.

Hosmer, D. W., & Lemeshow, S. (2002). *Applied Logistic Regression* (second ed.). New York: John Wiley & Sons, Inc.

Husson, F., Josse, J., Le, S., & Mazet, J. (2015). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. *R Package Version 1.31.4*. Retrieved from http://CRAN.R-project.org/package=FactoMineR

Kleinbaum, D. G., & Klein, M. (2002). *Logistic Regression A Self-Learning Text* (second ed.). New York: Springer.

Lee, E. T., & Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis.* New York: John Wiley & Sons.

Liao, T. F. (1994). *Interpreting Probability Models, Logit Probit and Other Generalized Linear Models.* USA: Sage Publications, Inc.

Little, R. J., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data .* New York: John Wiley & Sons Inc.

Liu, G., & Gould, A. L. (2002). Comparison of Alternative Strategies for Analysis of Longitudinal Trials. *Journal of Biopharmaceutical Statistics, 12*, 207-226.

Mallinckrodt, C. H., Carroll, R. J., Debrota, D. J., Dube, S., Molenberghs, G., Potter, W. Z., . . . Tollefson, G. D. (2003). Assessing and Interpreting Treatment Effects in Longitudinal Clinical Trials with Subject Dropout. *Biological Psychiatry*, 754-760.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models.* London: Chapman and Hall.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models* (Second ed.). New Jersey: John Wiley & Sons, Inc.

Molenberghs, G., & Kenward, M. G. (2007). *Missing Data in Clinical Studies.* West Sussex: John Wiley & Sons, Ltd.

Molenberghs, G., & Verbeke, e. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., & Mallinckrodt, C. (2004). Analyzing Incomplete Longitudinal Clinical Trial Data. *Biostatistics, 5*, 445-464.

Nelson, K. E., & Williams, C. M. (2007). *Infectious Disease Epidemiology: Theory and Practice* (second ed.). Boston: Jones & Bartlett.

Nenadic, O., & Greenacre, M. (2007). Correspondence Analysis in R with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software, 20(3)*.

Olsson, U. (2002). *Generalized Linear Models, An Applied Approach.* Lund: Studentlitteratur AB.

Panagiotakos, D. B., & Pitsavos, C. (2004). Interpretation of Epidemiological Data Using Multiple Correspondence Analysis and Log-Linear Models. *Journal of Data Science*, 75-86.

Physicians for Peace (PFP). (2015, 11 20). *Access to Safe Blood / Physicians for Peace/"teach one. heal many".* Retrieved from Physicians for Peace: https://physiciansforpeace.org/highlight-story/110

Pregibon, D. (1981). Logistic Regression Diagnostics. *The Annals of Statistics, 9*, 705-724.

R Core Team. (2015). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. Retrieved from https://www.R-project.org/.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology, 27*, 85-95.

Roth, P. L. (1994). Missing Data: A Conceptual Review for Applied Psychologists. *Personnel Psychology, 47*, 537-560.

RStudio Team. (2015). RStudio: Integrated Development for RStudio, Inc. Retrieved from http://www.rstudio.com/.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63*, 581-592.

Rubin, D. B. (1978). Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse. *In Imputation and Editing of Faulty or Missing Survey Data*, 1-23.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons, Inc.

Safe Blood for Africa (SBFA). (2015, 6 6). Retrieved from Safe Blood for Africa: http://www.safebloodforafrica.org/

Sarkar, S. K., Habshah, M., & Sohel, R. (2011). Detection of Outliers and Influential Observations in Binary Logistic Regression: An Empirical Study. *Journal of Applied Sciences, 11*, 26-35.

SAS Global Forum. (2012).

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data.* UK: Chapman & Hall.

Schafer, J. L. (1999). Multiple Imputation: a primer. *Statistical Methods in Medical Research, 8*, 3-15.

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*, 147-177.

Schafer, J. L., & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing Data Problems: A Data Analyst's Perspective. *ultivariate Behav Res, 33*, 545-571.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-467.

Shao, J., & Zhong, B. (2004). Last Observation carried Forward & Last Observation Analysis. *Statistics in Medicine, 22*, 2429-2441.

Sood, R. (2010). *Hematology for Students and Practitioners* (sixth ed.). New Delhi: Jaypee Brothers.

Stine, G. J. (2011). *AIDS Update 2011.* New York: McGraw-Hill.

Tsikriktsis, N. (2005). A Review of Techiques for Treating Missing Data in OM Survey Research. *Journal of Operations Management, 24*, 53-62.

Velton, R. (2009). *Mali* (Third ed.). England: Bradt Travel Guides.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch. (2005). *Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models.* New York: Springer.

Volberding, P. A., Sande, M. A., Greene, W. C., & Lange, J. M. (2008). *Global HIV/AIDS Medicine.* Philadelphia: Elsevier Inc.

White, I. R., Royston, P., & Wood, A. M. (2010). Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine*, John Wiley & Sons, Ltd.

World Health Organization. (2004a). *Maternal Mortality in 2000: Estimates Developed by WHO, UNICEF, UNFPA.* Geneva: World Health Organization.

World Health Organization. (2004b). Global Database on Blood Safety Summary Report 2001-2. Retrieved 2015, from http:www.who.int/bloodsafety/GDBS_Report_2001-2002.pdf

World Health Organization. (2010). *Towards 100% Voluntary Blood Donation: A Global Framework for Action.* Geneva: World Health Organization 2010.

World Health Organization, & UNICEF. (2003). *The Africa Malaria Report 2003.* World Health Organization/UNICEF 2003.

Wu, L. (2010). *Mixed Effects for Complex Data.* USA: Chapman & Hall/CRC.

Appendix A

Challenges of Transfusion-Transmissible Infections

Blood transfusion saves many lives, however the transfusion of infected blood may be detrimental to recipients. Duggan & Duggan (2006) have reported that following the development of blood transfusion before World War II and its rapid evolution in the 1940's, it became evident that there were patients developing jaundice sometime after blood transfusion. Nelson & Williams (2007) have reported that there is evidence to suggest that several microbial agents can be transmitted by blood transfusion if exposure occurs during the time when organisms are present in the blood stream. According to Nelson & Williams (2007), Hepatitis B virus, Hepatitis C virus and HIV are commonly transmitted by the transfusion of blood and blood products. They have further reported that although Malaria is caused by the bite of an infected mosquito, it can also be transmitted by blood transfusion and together, malaria and HIV cause more than 4 million deaths per year with more than 90% of these occurring in sub-Saharan Africa.

I. Malaria

According to Nelson & Willaims (2007), malaria affects about 400 million individuals each year which results in 1 to 2 million deaths worldwide. In the 2004 World Health Report, a total of 1,272,000 malaria deaths were estimated globally, of which 1,136,000 were in Africa (Nelson & Williams, 2007).

WHO & Unicef (2003) have reported that malaria continues to be a major impediment to health in Africa south of the Sahara, where it frequently takes its greatest toll on very young children and pregnant women, who are at highest risk for malaria morbidity and mortality. It is further reported that most children experience their first malaria during

the first or the second year of life, when they have not yet acquired adequate clinical immunity.

Nelson & Williams (2007) have reported that malaria is prevalent in regions were childhood malnutrition is common and they claimed that by improving nutrition in children, malaria morbidity and mortality could be reduced significantly. They have further reported that although malaria can affect anybody, it is mainly a disease of the poor and uninformed and the disease is much higher in the poor rural areas of Africa than in the developed urban areas.

According to Sood (2010) the transfusion of blood that contains malaria will result in transfusion-transmitted malaria. He further reported that in some chronic malaria patients, the parasite may not be seen or visible on the peripheral smear examination but a unit of their blood would pass on enough parasites to the recipient.

Nelson & Williams (2007) have reported that severe anaemia from malaria requires blood transfusion, yet there remains an additional risk to the consequences of malaria due to the dangers of TTIs.

II. Syphilis

Nelson & Williams (2007) have reported that syphilis became epidemic in the 1940's, as a highly contagious venereal disease in Spain, Italy and France. They have reported that thereafter the disease spread rapidly through Europe and America. Although the disease is predominantly transmitted sexually, Dennis & Anthony (2010) have reported that the disease could also be transmitted via blood transfusion and organ donation.

Dennis & Anthony (2010) have claimed that with the onset of penicillin therapy, the United States saw a 95% decline in the number of syphilis cases in 2000 from 1943. However, they reported that syphilis continues to be a significant health issue worldwide with about 12 million new infections each year with sub-Saharan Africa, South America and Southern Asia been being the regions mostly affected by the disease.

III. HIV/AIDS

According to Volberding et al. (2008) by the end of 2006 an estimated 39.5 million people were living with HIV and more that 20 million people had died worldwide.

According to Stine (2011) the first case of AIDS in Africa was identified in 1982 and as of 2011, about 75% of AIDS deaths have occurred in Africa. He reported that Africa consists of 10% of the world's population, accounts for about 68% of all global HIV infections and 90% of all new HIV infections.

Nelson & Williams (2007) have claimed that most countries in sub-Saharan Africa have been devastated by the HIV/AIDS pandemic with the life expectancy of populations declining by 20 years or more with a significant amount of young adults critically ill or dead from the disease. Volberding et al. (2008) have also agreed that sub-Saharan Africa remains the hardest-hit region of the disease with about 29.7 million people living with HIV and about 2.8 million new HIV infections in 2006. Stine (2011) reported that in sub-Saharan Africa, 61% of HIV positive adults are women and about 90% of children are living with HIV/AIDS.

Volberding et al. (2008) have reported that HIV can be transmitted from person to person through sexual intercourse, blood transfusion, sharing of contaminated injection equipment for intravenous drug use, from mother to child and through other forms of exposure to contaminated blood. According to Nelson & Williams (2007) the risk of transmitting HIV through the transfusion of blood and blood products was discovered very early in the AIDS epidemic. They have reported that the transfusion of HIV-contaminated blood is the most effective way to transmit the virus and over 90% of seronegative recipients are infected by the transfusion of a single contaminated unit of blood. Dennis & Anthony (2010) have claimed that the first transfusion-associated AIDS cases were reported in 1982 and by the end of 2005, more than 9300 individuals in the United States developed AIDS after having received HIV-contaminated blood transfusions, blood components or transplanted tissue. In Africa, an estimated 10% of HIV infections are caused by unsafe blood transfusion as reported by Volberding et al. (2008).

Nelson & Williams (2007) have reported that in the 1980's Mexico experienced a significant incidence of HIV infections in paid plasmapheresis donors when donors were infected during donation by contaminated blood collection equipment. These transfusion transmissible cases were acquired among women who needed blood transfusion for bleeding during delivery. It was found that there were 400 cases of AIDS among paid donors and over 2500 cases were reported among transfusion recipients in this population. Nelson & William (2007) have further reported that the epidemic in the country was controlled by closing commercial plasmapheresis centers, outlawing paid donors and establishing licensed state blood transfusion centers with adequate infection control procedures.

Volberding et al. (2008) have recounted that an estimated 5-10% of cumulative HIV infections worldwide occur through blood transfusion; however the incidence of such cases has declined due to the implementation of standard blood control procedures in most countries. These standard blood control procedures include the screening of blood doors for behavioral risks and donors are selected if they have a significant lower risk of infections. On the other hand, Nelson & Williams (2007) have reported that despite ensuring the safety of blood supply in industrialized countries, blood transfusion in many developing countries still carry a significant risk of HIV transmission. Dennis & Anthony (2010) have conferred that HIV transmission by blood and blood products is still an ongoing threat in sub-Saharan Africa where the routine screening of blood is not universally practiced. Nelson & Williams (2007) have further reported that first time donors are much more common in developing countries than repeat donors and that first time donors and paid donors carry a much higher risk of transfusion transmissible HIV, Hepatitis B Virus (HBV) and Hepatitis C Virus (HBC) infections in most populations.

IV. Hepatitis B Virus (HBV)

Duggan & Duggan (2006) have claimed that the HBV virus is present in blood and blood products such as semen and saliva and settles in the liver after transmission following an incubation period of two to six months. They have reported that the virus enters the cell nucleus and undergoes a complex series of changes leading to the appearance in the

serum of HBsAg (the outer coat of the virus which indicates the presence of the virus) and HBeAg (its presence indicates active viral replication).

Nelson & Williams (2007) reported that HBV can be transmitted by percutaneous blood exposure, sexual intercourse and from a mother to an infant. They have also reported that persons receiving pooled blood products also have high rates of HBV infection because very large pools may include a rare donor who was in the seronegative window period or had a false-negative test for HBsAg at the time of donation.

The prevalence of the HBV infection differs broadly throughout the world. Both Duggan & Duggan (2006) and Nelson & Williams (2007) have reported that in China, Southeast Asia, sub-Saharan Africa, Indonesia and several other areas including Alaska, Northern Canada and Greenland, carrier rates are high from about 8%. According to Duggan & Duggan (2006), in the undeveloped world, childhood infection is almost universal, most episodes of which are subclinical and subicteric and in one half of which there are not even abnormal liver function tests. Duggan & Duggan (2006) have argued that in about 10% of episodes, the infection does not resolve and a so called 'carrier state' follows and as a result the worldwide prevalence of HBV is about 300 million. They have also reported that in the United States (US) the carrier rate is about 1-1.25 million with serological evidence of past infection in about 6% of those less than 20 years, rising to 31% in those over 20.

Nelson & Williams (2007) have reported that the screening of blood donors for HBsAg was instituted in 1973. The introduction of screening of donors for HBsAg reduced the risk of transfusion transmitted HBV as reported by Nelson & Williams (2007). In addition, they have claimed that the risk of transfusion-transmitted HBV infection in the U.S has declined considerably in the last few decades.

According to Nelson & Williams (2007), the WHO has recommended that HBV vaccines be included with the vaccines given in the Expanded Programme of Immunization (EPI) for countries having high or moderate endemicity of HBV infection, however, many countries in sub-Saharan Africa have not yet included the HBV vaccine in their EPI programs due to economic constraints and lack of appreciation of the sequelae of chronic HBV infections in their countries.

V. Hepatitis C Virus (HVC)

Approximately 170 million people worldwide may be infected with HVC as reported by both Ahmed et al. (2007) and Nelson & Williams (2007). They have claimed that blood transfusions were the most common way for the virus to be transmitted prior to the testing of blood products for HCV. According to Duggan & Duggan (2006), a screening test was developed in 1990 and enormous strides were made in the knowledge of the virus and the role it plays in human disease. Duggan & Duggan (2006) have reported that before the adequate testing methods developed in 1990, the risk of the virus was 1 in 500 units and is now down to about 1 in 10^5 units. They claim that a significant contributor to this reduction is the rigorous donor screening and testing for HBV and HIV, which frequently coexist with HCV testing.

Nelson & Williams (2007) have reported that since 1999 all blood donors in the United States have been screened for HIV-1 and HCV RNA and among 39,721,404 donors screened between March 1999 and April 1, 2002, 170 (4.3 per million) were HCV RNA positive and 105 (2.9 per million) of these individuals donated blood that had no infectious markers and was safe for transfusion. They have further stated that RNA amplification for donor screening is used in the United States and Europe and hence the rate of transfusion-transmissible HCV infection in these countries is low. Whereas, in developing countries that only utilize enzyme immunoassays (EIAs), or do not have adequate screening tests for HCV, the infection is much more common.

Appendix B

SAS Procedures

A.1 Main-Effect Model

PROC LOGISTIC was used to fit the main-effect model as follows:

```
ods html;
Proc logistic descending data=Mali;
Class Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD GP AppWay_VNRBD AppWay_PD/ param=ref;
Model Donor = Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD GP AppWay_VNRBD AppWay_PD
 /link=logit alpha=0.05 lackfit;
run;
ods html close;
run;
```

where Donor=Donor/Non-donor, RB= Have you ever received blood, HSMBD=Have you ever heard or have seen messages about blood donation, Edu_level=highest educational level, Btrt=blood required is used to treat malaria and other diseases, BEmerg= a respondent thought that the blood required for transfusion was used for emergencies/disasters, BMal=whether a respondent thought blood transfusion is required to correct malnutrition, replace lost fluids of any type or to make up blood volume, SmsgsBD=what do you think is the best way to spread messages about blood donation, GP=Do you think blood donation is a good practice and everyone should donate, AppWay_VNRBD= Do you think the appropriate way to give blood is voluntary non-remunerated blood donation, AppWay_PD= Do you think the appropriate way to give blood is paid blood donation.

`Lackfit' request the Hosmer-Lemeshow goodness of fit test

A.1 Model Fitting

PROC GENMOD was used to do diagnostic testing, such as checking for overdispersion, calculation of the predicted probabilities, residuals and linear predictor statistics. The procedure was implemented as follows

```
ods html;
proc genmod descending data=Mali;
Class Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD GP AppWay_VNRBD AppWay_PD/ param=ref;
Model Donor = Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD GP AppWay_VNRBD AppWay_PD /dist=bin link=logit alpha=0.05 aggregate scale=deviance scale=pearson converge=1e-20 obstats type3;
run;
ods html close;
```

Where, 'aggregate' specifies the subpopulation on which the Pearson and the deviance are calculated. 'scale' specifies the scale parameter for an overdispersed model, 'converge', sets the convergence criterion. 'obstats' specifies an additional statistic including; residuals, predicted values, linear predictors and the dfbetas statistics and 'type3'=requests statistics for type3 contrast.

A.3 Plots in SAS

Plots in PROC LOGISTIC can be done using the output statement, or by directly specifying the plot options in the PROC LOGISTIC statement or the model statement. The PROC LOGISTIC plots were done directly using the PROC LOGISTIC statement and the model statement. The plots in PROC GENMOD was done directly by specifying the plot option in the PROC GENMOD statement and was used to plot the Cook's distance for influence diagnostics.

1. Plotting of the diagnostics for the leverage and influential observations and also the ROC curve for the model predictive accuracy power was done as follows:

```
ods html;
ods graphics on;
Proc logistic descending data=Mali plot (only label) = (phat leverage dpc);
Class Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD GP
AppWay_VNRBD AppWay_PD/ param=ref;
Model Donor = Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD
GP AppWay_VNRBD AppWay_PD/link=logit alpha=0.05 lackfit plcl outroc=rocl;
run;
ods graphics off;
ods html close;
```

Where 'phat leverage dpc' plots the leverage and influential observations and 'outroc=rocl' plots the ROC curve.

2. Plotting the Cooks' distance

```
ods html;
ods graphics on;
proc genmod descending data=Mali plots=cooksd;
Class Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD GP
AppWay_VNRBD AppWay_PD/ param=ref;
Model Donor = Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD
GP AppWay_VNRBD AppWay_PD /dist=bin link=logit;
run;
ods graphics off;
ods html;
```

where 'plots=cooksd' plots the Cooks' distance for the test of influential observations. The option 'plots=Predicted' was used to find the probability distribution of the predicted values for the logit, probit, and complementary log-log models respectively.

3. Plot for Deviance Residual vs Linear Predictor

```
ods html;
ods graphics on;
proc logistic descending data=Mali;
Class Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD GP AppWay_VNRBD AppWay_PD/ param=ref;
Model Donor = Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD GP AppWay_VNRBD AppWay_PD/ link=logit alpha=0.05;
output out=sasuser p=pred xbeta=logit resdev=resdev;
run;

ods html;
ods graphics on;
proc gplot data=sasuser;
plot DevianceResidual*logit;
ods graphics off;
ods html;
```

4. Plot for Pearson Residual vs Linear Predictor

```
proc logistic descending data=Mali;
Class Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD GP AppWay_VNRBD AppWay_PD/ param=ref;
Model Donor = Age Gender RB KDBG HSMBD Edu_level Btrt BEmerg BMal SmsgsBD GP AppWay_VNRBD AppWay_PD/scale=none;
output out=out1 xbeta=xb reschi=reschi;
run;
axis1 label=('Linear Predictor');
axis2 label=('Pearson Residual');
proc gplot data=out1;
plot reschi * xb / haxis=axis1 vaxis=axis2;
run;
```

A.4 Checking of Link Function

PROC GENMOD was used to test the choice of the link function in what follows:

```
ods html;
proc genmod descending data=Mali;
model Donor=LPred SLPred/dist=bin link=logit;
run;
ods html close;
```

where 'LPred'=linear predictors and 'SLPred'=Squared linear predictors.

A.5 Multiple Correspondence Analysis

1. The first two dimensions are plotted to examine the associations among the categories as follows:

```
proc corresp mca observed data = Mali outc=Coor;
tables Age Gender RB Donor HSMBD AppWay_VNRBD AppWay_PD GP Btrt BEmerg
BMal Edu_level SmsgsBD ;
run;
% plotit(data=Coor, datatype=corresp, href=0, vref=0);
```

2. The following code was used to get the mca plot using Greenacres'adjustment to inertias:

```
proc corresp mca observed data = Mali dim=3 outc=Coor greenacre;
tables Age Gender RB KDBG Donor HSMBD AppWay_VNRBD AppWay_PD GP Btrt
BEmerg BMal Edu_level SmsgsBD ;
run;
```

where 'greenacre' specifies greenacres' adjustment to inertias and 'dim' specifies the number of dimensions to use.

A.5 Missing Data

The PROC MI procedure implements methods for creating imputations under monotone and arbitrary data patterns of missing data, PROC LOGISTIC is used to generate the parameter

estimates and covariance matrix for each imputed data set stored in the dataset mi_fcs and PROC MIANALYZE analyzes results from the multiple imputed data sets.

```
proc mi data=Mali nimpute=20 out=mi_fcs ;
class Donor Gender Btrt BEmerg Age KDBG GP HSMBD Edu_level AppWay_VNRBD
AppWay_PD BMal RB SmsgsBD;
fcs plots=trace(mean std);
var Donor Gender Btrt BEmerg Age KDBG GP HSMBD Edu_level AppWay_VNRBD
AppWay_PD BMal RB SmsgsBD;
fcs discrim(Donor Gender Btrt BEmerg Age KDBG GP HSMBD Edu_level AppWay_VNRBD
AppWay_PD BMal RB SmsgsBD /classeffects=include) nbiter =100 ;
run;

proc logistic desc data=mi_fcs ;
class Donor Gender Btrt BEmerg Age KDBG GP HSMBD Edu_level AppWay_VNRBD
AppWay_PD BMal RB SmsgsBD ;
model Donor (event='1') = Gender Btrt BEmerg Age KDBG GP HSMBD Edu_level
AppWay_VNRBD AppWay_PD BMal RB SmsgsBD;
by _imputation_ ;
ods output parameterestimates = outparms ;
run ;

proc mianalyze parms (classvar=classval)=outparms;
class Donor Gender Btrt BEmerg Age KDBG GP HSMBD Edu_level AppWay_VNRBD
AppWay_PD BMal RB SmsgsBD;
modeleffects intercept Donor Gender Btrt BEmerg Age KDBG GP HSMBD Edu_level
AppWay_VNRBD AppWay_PD BMal RB SmsgsBD;
run ;
```

where , ‘nimpute=20’ specifies 20 imputations that are created for the missing data, ‘fcs discrim’ specifies discriminant function method to impute the classification variables.

Appendix C

Questionnaire

Safe Blood for Africa Foundation™

KNOWLEDGE, ATTITUDE AND PRACTICE (KAP) SURVEY: DRC

Voluntary Non- Remunerated Blood Donation

INSTRUCTIONS TO THE INTERVIEWER

You have an important role to ensure that this survey results in reliable data to assist in the development of a Voluntary Non Remunerated Blood Donation (VNRBD) strategy and programme in DRC. The purpose of collecting blood from VNRBD is to increase and ensure the adequate supply of safe blood in DRC.

The objectives of a KNOWLEDGE, ATTITUDE AND PRACTICE (KAP) SURVEY is to tell us ***what people know about blood donation, how they feel about donating and receiving blood, and also how they behave when asked to donate blood.***

Using this survey, we will ask individuals from a number of groups several questions relating to the following subjects:

1. Knowledge: These questions investigate their understanding of blood donation, of the importance of *voluntary* blood donation, and how donated blood is used;
2. Attitude: These questions gauge people's prevailing attitudes, beliefs and misconceptions about blood, blood donation and how blood is used by doctors; and
3. Practice: These questions assess people's current or potential practices and behaviours with regard to blood donation.

It is very important that you let participants answer all of the questions in this survey in their own words. Do not lead or encourage them to give any particular answer or choice. For this reason the answers to the questions should not be revealed to the interviewee. This is critical in the **PRACTICES** and **KNOWLEDGE** sections.

Your role is to ask the questions and match the answer you receive from the participant to the choices on the questionnaire. This is designed to prevent us gathering false information as a result of the interviewee guessing or selecting additional answers from the choices offered.

In the **ATTITUDE** section, once the person being interviewed has made all of his or her choices in response to the question, please ask them to rank the top three from most important by assigning a (1) to the most important through to (3) for the third most important and record this in the column provided.

Informed consent

Many of the questions in this survey are personal and so the data collected are sensitive. Therefore, you need to inform the potential participants of the purpose of the survey and to reassure them that everything that is written down will be anonymous – in other words, no one will be able to tell which answers belong to any specific person. You must obtain their permission to proceed before you start asking questions. As we are obligated to maintain the strictest confidentiality, the name of the participant should never be written down anywhere. For literate participants, please get the Interviewee to read the "**Informed consent Blood Donation KAP Survey"** that follows and get their consent to participate. When you receive this verbal permission to start the survey, **they** should sign the form to indicate that verbal consent was received by **you. For illiterate participants you sign the form after they have given** their **consent. You should never sign on behalf of another member of the interview team**. If anyone refuses to give their consent, please thank them for their time and move to the next participant. **Do not, under any circumstances, attempt to persuade them to participate**. You must, however, record unsuccessful attempts to help us understand people's willingness to participate in a survey about blood donation.

Safe Blood for Africa Foundation™

KNOWLEDGE, ATTITUDE AND PRACTICE (KAP) SURVEY

Democratic Republic of the Congo (DRC)

Task order: **Blood Safety Technical Assistance Services contract # 200-2010-36449-0001**

Interviewers name: _____ Cell:_____ Email: _____

DATE	LOCATION	INTERVIEW START TIME	INTERVIEW END TIME	QUESTIONNAIRE NUMBER
DD/MM/YYYY		hh:mm	hh:mm	**Code:** nnn
Survey group	Current VNRBD donor	Non-donor	Lapsed donor	FRD

FRD – family replacement donor

GUIDE TO INTERVIEWEES

Informed consent Blood Donation KAP Survey

ALLOW YOUR POTENTIAL INTERVIEWEE TIME TO READ AND UNDERSTAND THIS DECLARATION. IF THE INTERVIEWEE CANNOT READ, YOU MAY READ THE FOLLOWING CONSENT ALOUD.

Dear Participant,

The Centre National de Transfusion Sanguine are conducting a survey regarding knowledge attitudes and practices towards blood donation in DRC.

Introduction: My name is *… insert interviewers name…*and we are interviewing people in your area in order to assess knowledge, attitudes and practices relating to blood donation in DRC. The information will be used for recommending appropriate strategies to strengthen and improve blood transfusion practices that seek to ensure that adequate safe blood is available in DRC for the benefit of all citizens.

Confidentiality and Consent

Your answers are anonymous. Your name will not be written on this survey form, and will never be used in connection with any of the information you tell me. You do not have to answer any questions that you do not want to answer, and you may end this interview at any time you want to. However, your honest answers to these questions will help us better understand what people think, say and do in regard to blood donation. We would greatly appreciate your help in responding to this survey. The survey will take about 15-20 minutes to ask and respond to the questions. Would you be willing to participate?

If you are willing to participate, please witness and observe my signing the questionnaire below and complete the consent form if possible? In doing so I certify that I have your informed consent to proceed with the interview and fully commit to upholding confidentiality.

IF YOU HAVE ANY QUESTIONS REGARDING THE PROCESS PLEASE ASK BEFORE YOU CONSENT TO ME SIGNING THIS DOCUMENT

Interviewer's Name and Signature_____

(Signature of interviewer certifying that informed consent has been given verbally by respondent)

CHECKED BY SUPERVISOR: Signature _____ Date _____

SECTION 1: INTERVIEWEES CHARACTERISTICS

		1	2	3	4	5	6	7
Q1.1	What is your age?	18	19-25	26-30	31-40	41-50	51-60	>60
Q1.2	What is your gender?	**1** Female				**2** Male		
Q1.3	Have you ever received blood?	**1** Yes		**2** No		**3** Don't know		
Q1.4	Which of the following applies to you?	Never donated, no intention						**1**
		Never donated but would donate						**2**
		Regular donor for several years						**3**
		Recent voluntary non remunerated donor						**4**
		Lapsed Donor (donation before 2012)						**5**
		Recent family replacement donor						**6**
		Remunerated donor (by either family or by the transfusion service or blood bank)						**7**
		Before this moment, I didn't know that blood could be donated						**8**

SECTION 2: PRACTICES ABOUT BLOOD DONATION FOR CURRENT OR PREVIOUS BLOOD DONORS

Q2.1	Who asked you to donate?	A blood donor recruiter	**1**
		A friend	**2**
		A family member	**3**
		It was my decision	**4**
		I cannot remember	**5**
		Faith leader	**6**
		Other	**7**
Q2.2	How many times have you ever donated blood?	1 time (**Skip Q 2.3**)	**1**
		2 times	**2**
		Several times (estimate the number)	**Number**
		I cannot remember how many times	**?**
Q2.3	In an ordinary year how many times do you donate?	Less than once per year	**0**
		1 time	**1**
		2 times	**2**
		Many times (estimate the number)	**Number**
		I cannot remember	**?**
Q2.4	Which of the following describes the **type** of donation you have made?	Intended for a family member or friend ONLY	**Q241**
		Unspecified, for the sick in general	**Q242**
		For my own blood transfusion needs	**Q243**
		Don't know	**Q244**
		I was told I had to donate to replace blood used by a friend, family member or stranger	**Q245**
		Other	**Q245**
Q2.5	How old were you when you made your first blood donation?		**Years**
		Don't know	**?**
Q2.6	Why did you donate blood the first time?	Volunteered	**1**
		Recruited	**2**
		Relative or friend was sick	**3**
		For benefits (days off from work etc.)	**4**
		For money	**5**
		To relieve hypertension	**6**
		Higher self-appreciation	**7**
		Interest in getting test results	**8**
		Don't know	**9**
		Obliged by school or employer	**10**
		Other	**11**

Q2.7	*If you answered "Volunteered" to Q2.6 then answer this question, else move to Q2.8.* What was your reason for volunteering?	It is a family tradition	**1**
		As a service to my community	**2**
		To encourage my friends	**3**
		To help a member of my family	**4**
		To help a friend	**5**
Q2.8	Did you have any concerns when donating blood?	Yes	**1**
		No	**2**
		None that I remember	**3**
Q2.9	*If you answered "yes" to Q2.8 then answer this question, else move to Q2.10.* What were your main concerns when donating blood?	Acquiring HIV or another infection	**Q291**
		Infecting someone else	**Q292**
		Donating is against my culture	**Q293**
		Fainting	**Q294**
		Uncertain about the process	**Q295**
		Pain	**Q296**
		Fear of needles	**Q297**
		Blood donation is discouraged by my religion	**Q298**
		Do not want to know my status	**Q299**
		Don't know	**Q2910**
		I was afraid to learn my status	**Q2911**
Q2.10	After your first blood donation, how long was it until your next donation?	I never donated again	**1**
		Within 3 months	**2**
		4-6 Months	**3**
		7-12 Months	**4**
		More than 1 year	**5**
		Don't remember	**6**
		Don't know	**7**
Q2.11	When did you last donate blood?	In 2013	**1**
		In 2012	**2**
		Longer than the last year and a half to 5 Years ago	**3**
		More than 5 years ago	**4**
		Do not know	
Q2.12	How did you experience the personnel that assisted you with your last blood donation?	Friendly	**1**
		Unfriendly	**2**
		Indifferent	**3**
		Professionally dressed	**4**
		Poorly dressed	**5**
Q2.13	How did you find the equipment at the place where you last donated blood?	Very good	**1**
		Adequate	**2**
		Inadequate	**3**
		Can't remember	**4**
		Unclean	**5**
		Poorly maintained	**6**
Q2.14	How did you find the environment in which you last donated blood?	Suitable	**1**
		Unsuitable	**2**
		I don't know	**3**
		Can't remember	**4**
		Unclean	**5**
		Poorly maintained	**6**
Q2.15	To your knowledge which of your family members have ever donated blood? ***(please indicate the number of individuals that you are aware of in each category)***	Parents	**Q2151**
		Spouse	**Q2152**
		Children	**Q2153**
		Siblings	**Q2154**
		Uncles, aunts	**Q2155**
		None	**Q2156**
		Do not know	**Q2157**

Q2.16	What is donated blood used for? ***(please indicate all that you are aware of)***	To treat sick patients	**Q2161**
		To save lives	**Q2162**
		To help prevent bleeding	**Q2163**
		To treat malaria in children	**Q2164**
		To save mothers giving birth	**Q2165**
		Do not know	**Q2166**
		For the blood service to sell	**Q2166**
		For evil purposes	**Q2167**

SECTION 3: FOR INTERVIEWEES WHO HAVE NEVER DONATED BLOOD

Q3.1	Why have you never donated blood*?*	Never been asked	**1**			
		Did not know it was necessary	**2**			
		I had no information on blood donating	**3**			
		Worried about diseases	**4**			
		Cannot stand the sight of blood	**5**			
		Fear of hospitals and clinics	**6**			
		Doctor advised against donation	**7**			
		I do not want to make other people sick	**8**			
		BTS staff deferred me	**9**			
		Heard negative things about donation	**10**			
		I am afraid to know my serology status	**11**			
		Donor clinic too far	**12**			
		Donation times inconvenient	**13**			
		Staff in clinics inefficient & incompetent	**14**			
		No gifts were given for my blood	**15**			
		Do not wish to donate	**16**			
		I fear needles	**17**			
		Too busy	**18**			
		I do not know where to donate blood	**19**			
		Did not know it was possible	**20**			
		Never thought of it	**21**			
		Do not know	**22**			
		Against my culture of religious beliefs	**23**			
		I do not have enough blood to donate	**24**			
		My health does not allow me	**25**			
		I don't eat properly	**26**			
		Because I have received a transfusion	**27**			
		I am breast feeding	**28**			
Q3.2	Have you refused to donate blood when asked?	**1** Yes	**2** No	**3** Never asked	**4** Can't remember	
Q3.3	*If you answered "Yes" to Q 3.2 then answer this question, else proceed to Q3.4* Why did you refuse?	Fear of acquiring HIV or other infection	**Q331**			
		Fear of infecting someone else	**Q332**			
		Fear of damage to my health	**Q333**			
		Against my religion or culture	**Q334**			
		Fear of fainting	**Q335**			
		I do not know enough about blood donation	**Q336**			
		I do not think I should give my blood free when the recipient must pay	**Q337**			
		Do not know	**Q338**			
		The doctor advised against it	**Q339**			
		Other (please explain)	**Q3310**			
Q3.4	Do you ever intend to donate blood in the future?	**1** Yes	**2** No	**3** I am unsure		
Q3.5		Just ask me to donate	**1**			
		If someone I know needs blood	**2**			
		Provide more information on the need	**3**			

	Which of the following factors would motivate you to become a blood donor?	An appeal on TV or radio	**4**	
		Provide video or audio material	**5**	
		Have a recognition program for donors	**6**	
		More efficient and competent staff	**7**	
		Convenient donation location	**8**	
		Shorter donation time	**9**	
		Gift to show I donated	**10**	
		Chance to win a prize	**11**	
		Inspired by a leader	**12**	
		Discount vouchers from local merchants	**13**	
		Do not know	**14**	
		Nothing would motivate me	**15**	
		If I have too much blood	**16**	
		If the doctor advises me to	**17**	
		Other	**18**	
Q3.6	Why would you not donate blood in the future?	No I want to donate	**1**	
		There is no benefit to me	**2**	
		Fear of HIV infection	**3**	
		Fear of infecting someone else	**4**	
		Fear of learning my HIV status	**5**	
		Against my culture or religion	**6**	
		Fear of damaging my health	**7**	
		My health does not allow donation	**8**	
		Fear of the procedure	**9**	
		Uncertain of the process	**10**	
		Do not know	**11**	
		I am too old	**12**	
Q3.7	*If you answered "Against my culture or religion" to Q 3.6 then answer this question, else proceed to Q.4.1.* Please list the cultural beliefs that prevent you donating blood.	List		

SECTION 4: KNOWLEDGE ABOUT BLOOD TRANSFUSION

Q4.1	Do you know the different blood groups?	**1** Yes				**2** No	
Q4.2	What blood groups are there?	**Q421** A	**Q422** B	**Q423** O	**Q424** AB	**Q425** Other	**Q426** Do not know
Q4.3	Which blood group are you?	**Q431** A	**Q432** B	**Q433** O	**Q433** AB	**Q435** Other	**Q436** Do not know
Q4.4	What is blood in clinical terms? *Mark all those you think are correct*	Red liquid flowing in veins and arteries					**Q441**
		A fluid that can be manufactured for people					**Q442**
		Something we can get from animals for people					**Q443**
		Gives you life					**Q444**
		Body fluids made in the heart					**Q445**
		Do not know					**Q446**
		Other (please explain)					**Q447**
Q4.5	What are the functions of blood? *Mark all those you think are correct*	Carries oxygen to tissues to sustain life					**Q451**
		Contains red cells and white cells					**Q452**
		Carries proteins, minerals and nutrients					**Q453**
		Carries CO_2 and waste products					**Q454**
		Replaces body fluid					**Q455**
		It is poisonous					**Q456**
		It gives people power and energy					**Q457**
		Do not know					**Q458**
		Other (please explain)					**Q459**

Q4.6	Why do people require blood transfusions?	To treat diseases	**1**	
		To help them recover from accidents	**2**	
		In order to undergo surgery	**3**	
		For mothers in childbirth	**4**	
		To treat malaria	**5**	
		Emergencies/disasters	**6**	
		To give one energy	**7**	
		To gain spiritual power	**8**	
		To replace lost fluids of any type	**9**	
		Do not know	**10**	
		To correct malnutrition	**11**	
		To make up blood volume	**12**	
		Other (please explain)	**13**	
Q4.7	Can a person get infected with a disease by receiving blood?	Yes	**1**	
		No	**2**	
		Do not know	**3**	
Q4.8	*If you answered "Yes" to Q 4.7 then answer this question, else proceed to Q4.9* List these diseases?	Malaria	**Q481**	
		Tuberculosis	**Q482**	
		HIV/AIDS	**Q483**	
		Trypanosomiasis	**Q484**	
		Hepatitis	**Q485**	
		Syphilis	**Q486**	
		Yellow fever	**Q487**	
		Tetanus	**Q488**	
		Diabetes	**Q489**	
		Typhoid	**Q4810**	
		Others	**Q4811**	
Q4.9	How often can blood be donated	Weekly	**1**	
		Monthly	**2**	
		Every 7 weeks (56 days)	**3**	
		Every three months	**4**	
		Every four months	**5**	
		Every six months	**6**	
		Only once per year	**7**	
		Do not know	**8**	
		Other	**9**	

Q4.10	Who can donate blood*?*	Men	**1**	
		Women	**2**	
		Young	**3**	
		Old	**4**	
		Pregnant women	**5**	
		Vulnerable groups	**6**	
		Healthy	**7**	
		Sickly	**8**	
		Do not know	**9**	
		Those with O group	**10**	
		Those with sufficient blood	**11**	
		Other (please explain)	**12**	
Q4.11	Who cannot donate blood*?*	Men	**1**	
		Women	**2**	
		Young people under 18	**3**	
		Old people over 60	**4**	

		Pregnant women	5	
		Vulnerable groups	6	
		Healthy people	7	
		Sickly people	8	
		Do not know	9	
		Women who are Menstruating	10	
		Anaemic people	11	
		Breast feeding women	12	
		Other (please explain)	13	
Q4.12	What do you need to do to get blood?	We need to replace units we use	1	
		The patient just gets the blood for no cost	2	
		We pay for the blood	3	
		We pay for the tests	4	
		Do not know	5	
		Other	6	
Q4.13	*If you answered "We pay for the blood or test" to Q 4.12 then answer this question, else proceed to Q4.14* What do you pay for? And how much	Tests	Q4131	Amount
		Service	Q4132	Amount
		The use of blood	Q4133	Amount
		We pay the donor	Q4134	Amount
		Total	Q4135	Amount
		Do not know	Q4136	
Q4.14	What tests do you know the blood bank does?	HIV	Q4141	
		AIDS	Q4142	
		Malaria	Q4143	
		Hepatitis B	Q4144	
		Hepatitis C	Q4145	
		Blood group	Q4145	
		Syphilis (gonorrhoea)	Q4146	
		Do not know	Q4147	
Q4.15	What do you think of the system in place in Mali?	I think it is.....	Q4151	
		Nothing in particular	Q4152	
		I do not know	Q4153	
Q4.16	What would you change?	I would change.....	Q4161	
		Nothing in particular	Q4162	
		I do not know	Q4163	

SECTION 5: ATTITUDES AND PERCEPTIONS TO DONATION

Q5.1	How should blood donors be treated? *(Tick all that are applicable)*	As patients	1
		As valued customers	2
		As servants of the blood service	3
		As clients to be paid	4
		Do not know	5
		Other (please explain)	6
Q5.2	What information should be provided to donors before donating? (*Tick all that are applicable*)	Risks of donation	Q521
		Mechanism of donation	Q522
		Health exclusion criteria	Q523
		Uses made of their blood	Q524
		Importance of donors	Q525
		Value of donation	Q526
		Do not know	Q527
		Other	Q528

Q5.3	What health checks should be conducted before donation? (*Tick all that are applicable*)	Completion of registration form	**Q531**
		Health questionnaire	**Q532**
		Individual confidential counselling	**Q533**
		Basic health check	**Q534**
		Has the donor taken any medication before donation?	**Q535**
		Blood pressure check	**Q536**
		Do not know	**Q537**
		Other	**Q538**
Q5.4	What lifestyle check should be conducted before donation? *(Tick all that are applicable*)	Monogamous	**Q541**
		Polygamy	**Q542**
		Non-smoker	**Q543**
		Non-alcoholism	**Q544**
		Sportive	**Q545**
		Heterosexual or homosexual	**Q546**
		Do not know	**Q547**
		Other	**Q548**
Q5.5	What do you think about blood donation? (Tick all that are applicable)	It is a good practice	**Q551**
		It is a dangerous process	**Q552**
		I have no strong feelings	**Q553**
		It is important and everyone should donate	**Q554**
		I have not given it any thought	**Q555**
		Other	**Q556**
Q5.6	Why do certain people donate blood while others do not? (Tick all that are applicable)	They may be too old or too young	**Q561**
		They are the wrong gender	**Q562**
		They carry a disease that can be passed on	**Q563**
		They do not know about donating blood	**Q564**
		General fear of blood and the donation process	**Q565**
		Fear of the blood screening	**Q566**
		Their cultural system does not allow it	**Q567**
		It is against their religion	**Q568**
		Do not know	**Q569**
		Other	**Q5610**
Q5.7	In your opinion what is an appropriate way to give blood? (Tick all that are applicable)	Voluntary non remunerated (unpaid) donation	**Q571**
		At the request of relatives	**Q572**
		Be paid to give blood	**Q573**
		Only when it can be used for me	**Q574**
		People should not donate blood	**Q575**
		Do not know	**Q576**
		Other	**Q577**
Q5.8	Do people who donate blood receive something in exchange?	Yes	**1**
		Yes in some cases	**2**
		No	**3**
		Do not know	**4**

Q5.9	What rewards do they receive as compensation? **(please indicate all that you are aware of)**	Money	**Q591**
		Food and/or drinks	**Q592**
		Gifts	**Q593**
		Moral satisfaction	**Q594**
		Transportation to town	**Q595**
		Do not know	**Q596**

		Nothing	**Q597**
		Other	**Q598**
Q5.10	In your opinion what could be done to encourage more people to donate blood? **(please indicate all that you are aware of)**	Additional informational material	**1**
		Mass media sensitization campaigns	**2**
		Donors should be paid	**3**
		Improved donor appreciation	**4**
		Gifts to donors	**5**
		Round table seminars	**6**
		Do not know	**7**
		Other	**8**
Q5.11	Can something bad happen to a person who donates blood?	Yes	**1**
		No	**2**
		Do not know	**3**
Q5.12	*If you answered "Yes" to Q 5.11 then answer this question, else proceed to Q5.13* What can happen?	Contract disease	**Q5121**
		I could possibly die	**Q5122**
		Feel weak temporarily	**Q5123**
		Loss of health	**Q5124**
		Lose my spiritual power	**Q5125**
		My heart will be affected	**Q5126**
		Transfer of characteristics and traits	**Q5127**
		Usage for evil purposes	**Q5128**
		Do not know	**Q5129**
		Other	**Q51210**
Q5.13	What could be done to maintain a donor's health?	The blood service must use new collection materials each time	**Q5131**
		Test donors for diseases before donation	**Q5132**
		Medical exam before donation	**Q5133**
		Offer food and liquids after donation	**Q5134**
		Refrain from strenuous activities	**Q5135**
		Do not know	**Q5136**
		Refer donors to treatment if they test positive for an infection	**Q5137**
		Other	**Q5138**
Q5.14	If you had to convince a person to donate blood what would you say to him/her?	Brief idea:	
		Do not know	
Q5.15	What comes to your mind when I say the word Blood?	Symbolizes life	**Q5151**
		Symbolizes family	**Q5152**
		It is a gift of God	**Q5153**
		It is private	**Q5154**
		It is a communal resource	**Q5155**
		I do not know	**Q5156**
		Other	**Q5157**
Q5.16	Do you think you have enough blood to donate?	Yes	**1**
		No	**2**
		If not, why not	**3**
		Do not know	**4**
Q5.17	How much blood do you think you have?	Less than one litre	**1**
		1-2 litre	**2**
		3-5 litre	**3**
		5-7 litre	**4**
		8-10 litre	**5**
		More than 10 litre	**6**
		I don't know	**7**
Q5.18	Are there any factors that may influence the volume of blood in the body?	My health	**Q5181**
		The proportions of my body	**Q5182**
		My gender	**Q5183**
		Menstruation	**Q5184**
		My age	**Q5185**

		My nutrition	**Q5186**
		I don't know	**Q5187**
		Other	**Q5188**
Q5.19	What can you do to replace the blood given through donation?	Food and drinks	**Q5191**
		Rest	**Q5192**
		No strenuous activities	**Q5193**
		You can never get it back	**Q5194**
		I didn't know it was possible to get it back	**Q5195**
		I don't know	**Q5196**
		Other	**Q5197**
Q5.20	*If you answered "Food and drinks" to Q 5.19 then answer this question, else proceed to Q5.21* Specify which food and drinks?	Please explain	**Q5201**
Q5.21	Are there other practices that may replace blood that you know of?	No /I am not aware of it	**Q5211**
		Assisted by a traditional healer	**Q5212**
		Alternative remedies, such as Plants & herbs	**Q5214**
		Prayer	**Q5215**
		Other	
Q5.22	When will you consider agreeing to receive a blood transfusion?	When it is advised by a doctor	**Q5221**
		After I've explored other options	**Q5222**
		As a last resort	**Q5223**
		When my life is in danger	**Q5224**
		Other	**Q5225**
Q5.23	If you needed a transfusion, in what order would you want to receive blood from whom and why?	Family – friend- stranger	**1**
		Friend-family-stranger	**2**
		Stranger-family-friend	**3**
		Stranger-friend-family	**4**
		From the blood bank	**5**
		I have no preference	**6**
		Explain:	**7**
Q5.24	Under what circumstances would you consider requesting reimbursement for your blood?	Poverty	**Q5241**
		Hunger	**Q5242**
		For compensation of time and travel	**Q5243**
		I would never	**Q5244**
		I do not know	**Q5245**
		Other	**Q5246**
Q5.25	Do you consider yourself a Voluntary non-remunerated blood donor?	Yes Why?	**1**
		No Why?	**2**
		I don't know	**3**

SECTION 6: COMMUNICATION CHANNELS AND BEHAVIOUR

Q6.1	Have you ever seen or heard messages about blood donation?	Yes	**1**
		No	**2**
		I cannot remember	**3**
Q6.2	*If you answered "Yes" to Q 6.1 then answer this question, else proceed to Q6.4.*	Basic message:	

	What was the message about?		
Q6.3	Where did you see or hear those messages? ***(please indicate all that you are aware of)***	At the transfusion centre	**Q631**
		At the doctor or dispensary	**Q632**
		At the hospital	**Q633**
		Written media	**Q634**
		School/college	**Q635**
		In church	**Q636**
		At the mosque	**Q637**
		Road adverts (bill boards)	**Q638**
		Television	**Q639**
		Radio	**Q6310**
		I do not remember	**Q6311**
		Other	**Q6312**
Q6.4	What would be the best way to spread messages about blood donation? ***(please indicate all that you are aware of)***	Radio	**Q641**
		Television	**Q642**
		Written media	**Q643**
		Other printed media	**Q644**
		By word of mouth	**Q645**
		Banners	**Q646**
		Do not know	**Q647**
		In Church	**Q648**
		In schools/colleges/university	**Q649**
		Hospitals/clinics	**Q6410**
		Telephone or SMS	**Q6411**
		Other	**Q6412**
Q6.5	Have you ever been involved in any volunteer activity?	Yes	**1**
		No	**2**
Q6.6	*If you answered "Yes" to Q 6.5 then answer this question, else proceed to Q6.8.* What was the main purpose of such activity? ***(please indicate all that you are aware of)***	Helped the elderly	**Q661**
		Helping at the local clinic or hospital	**Q662**
		Working for the church	**Q663**
		Supporting the local school	**Q664**
		Raising funds for HIV/AIDS orphans	**Q665**
		Raising funds for medical services	**Q666**
		Helping a feeding scheme e.g. soup kitchen	**Q667**
		Raising funds for the sport club	**Q668**
		Raising money to build a school	**Q669**
		Others (please explain)	**Q6610**
Q6.7	If you are involved in any voluntary activities please rate their success. If not, please move onto Q6.8.	All were very successful	**1**
		Some were successful	**2**
		None were successful	**3**
		I gave up before the end of the project	**4**
		Do not know	**5**
Q6.8	Do you know any volunteer groups in your area? If yes chose from the list.	No	**Q681**
		Church youth group	**Q682**
		Youth group at the mosque	**Q683**
		Boy or Girl scouts	**Q684**
		Youth brigade	**Q685**
		Blood donor club or association	**Q686**
		Farmers association	**Q687**
		Support for vulnerable people	**Q689**
		Red Cross	**Q6810**
		Others	**Q6811**

Q6.9	Are you aware of any HIV/AIDS related programmes in your district?	Yes	**1**
		No	**2**

SECTION 7: SOCIO DEMOGRAPHICS

Q7.1	What is your highest education level?	Never went to school	**1**
		Primary school (Grades 1 to 6)	**2**
		Second Fundamental Cycle (Grades 7 to 12) or equivalent	**3**
		Technical/Practical qualification	**4**
		Lyceum	**5**
		Post Graduate level	**6**
		First bachelor's degree	**7**
Q7.2	Civil status?	Single	**1**
		Married	**2**
		Live in long standing partnership	**3**
		Polygamous relationship or marriage	**4**
		Separated	**5**
		Divorced	**6**
		Widowed	**7**
Q7.3	How many people live in your household?	Number	
Q7.4	What is your faith or religion?		
Q7.5	Are you employed at present?	Yes	**1**
		No	**2**
Q7.6	What is your main occupation? ***(please indicate all that apply e.g. civil servant and manager)***	Public servant/government employee	**Q751**
		Manager	**Q752**
		Professional (teacher, physician, etc)	**Q753**
		Salesman	**Q754**
		Service area worker (unskilled, cleaner etc)	**Q755**
		Tradesman (builder, mechanic etc)	**Q756**
		Self employed	**Q757**
		Student	**Q758**
		Servant of God	**Q759**
		Medically boarded	**Q7510**
		Armed forces/police	**Q7511**
		Retired	**Q7512**
		Unemployed	**Q7513**
		I am in multiple employment (list)	**Q7514**
		Other (list)	
Q7.7	Average monthly income ***(please note "Variable" means an inconsistent income from month to month)*** in CFA?	zero	**1**
		≤13500	**2**
		13690 to 22500	**3**
		22950 to 45000	**4**
		45450 to 225000	**5**
		225450 to 900000	**6**
		900450 to 2250000	**7**
		>2250000	**8**
		Variable	**9**

Q7.8		Running water	**Q771**

	Which of the following objects do you have in your household?	Water borne (flush) sanitation	**Q772**
		TV	**Q773**
		Mains electricity	**Q774**
		Battery or solar power	**Q775**
		Refrigerator	**Q776**
		Fixed telephone	**Q777**
		Mobile phone	**Q778**
		Washing machine	**Q779**
		Video/DVD	**Q7710**
		Radio	**Q7711**
		Bicycle	**Q7712**
		Satellite TV	**Q7713**
		Computer	**Q7714**
		Internet	**Q7715**
		Motor cycle	**Q7716**
		Motor car	**Q7717**
Q7.9	Main material of dwelling walls?	Cement	**1**
		Stone	**2**
		Brick	**3**
		Clay	**4**
		Wood	**5**
		Tin/Iron	**6**
Q7.10	Main material of dwelling roof?	Tiles	**1**
		Concrete	**2**
		Slates	**3**
		Thatch	**4**
		Fibre glass	**5**
		Corrugated iron	**6**
Q7.11	Do you live in:?	Urban zone	**1**
		Rural zone	**2**
Q7.12	Area/Banlieue/Commune?		
Q7.13	City/town?		
Q7.14	District?		

Ask interviewee if there is anything else they would like to add

Thank interviewee and finish the interview.

SECTION 8: Additional points raised by the INTERVIEWEE:

_____
_____
_____
_____
_____
_____
_____
_____