Strathmore
UNIVERSITY

Using the Generalized Linear Model (GLM) to model (specific) chronic diseases
Chronic disease in focus: Diabetes

Aleeq Salim - 101497

Submitted in partial fulfillment of the requirements for the Degree of BBS in Actuarial Science at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

February 2021

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.

Aleeq Salim [*Name of Candidate*]

 [*Signature*]

.....

February 10, 2021 [*Date*]

This Research Project has been submitted for examination with my approval as the Supervisor.

Elphas Okango [*Name of Candidate*]

 [*Signature*]

.....

February 10, 2021 [*Date*]

Strathmore Institute of Mathematical Sciences

Strathmore University

TABLE OF CONTENTS

List of Tables and Figures

List of Abbreviations

**Arranged in alphabetical order*

CDC	Centre for Disease Control and Prevention
GLM	Generalized Linear Model
IDF	International Diabetes Federation
IHME	Institute for Health Metrics and Evaluation
KDHS	Kenya Demographic and Health Survey
KNBS	Kenya National Bureau of Statistics
NCD	Non Communicable Disease
MLE	Maximum Likelihood Estimation
SS	Screenshot
Vs	Versus or against
WHO	World Health Organization

Acknowledgements

I am deeply indebted to my supervisor, Dr. Elphas Okango, for his continuous and unfailing assistance over the course of finishing this project. My colleagues Natacia Magomba and Ivy Kinyua also have a hand in the completion of this report through their support and encouragement.

I would also like to thank Deepa Ranpara for her constant support and assistance over the duration of completing this report.

This report would be incomplete without them.

Abstract

As per the World Health Organization, chronic diseases rank highly in incidence, prevalence, and mortality on a global scale with around 60% of the 56.5 million total global deaths in 2001. Among the most prevalent of the chronic diseases is diabetes. Diabetes is rapidly increasing in its rates of incidence and mortality. It was ranked as the seventh leading cause of mortality in 2016, despite the fact that it was not even featured in the top 10 list in the year 2000. Furthermore, according to the International Diabetes Federation there are around 552,400 diagnosed cases of diabetes in Kenya. It is evident from the above that diabetes is a serious and prevalent chronic disease in the country. As a result, the aim of this research project is to model and perform a regression analysis on diabetes using Kenyan data. The modeling process was done using a Generalized Linear Model. The estimators are expected to display optimal properties since GLMs use the MLE method of parameter estimation. Once the model was completed the model diagnosis displayed a significant R2, suboptimum deviances and a high AIC value. In conclusion, the most significant risk factors of diabetes in the country are high blood pressure, age, gender, and level of education.

Chapter 1: Introduction

Background Information

As per the definition of the U.S. National Center for Health Statistics, a chronic disease is one that lasts 3 months or more. Chronic diseases are inclined to become more common with age and they cannot be precluded by vaccines, cured by medication and they do not simply vanish (Shiel, 2018). Other definitions for chronic diseases claim different durations for the condition, such as the one provided by the CDC which states that chronic diseases last one year or more (NCCDPHP, 2019). Simply put, chronic diseases last for a long time, as opposed to acute diseases which are short term. This study aims to model a prevalent chronic disease in Kenya. For the modeling process, we will use a class of models referred to as GLMs, within which the response variable y_i is assumed to follow a distribution from the exponential family with mean μ_i, which is assumed to be a (often nonlinear) function of $x_i^T\beta$. We will get into more detail regarding the GLM later in the paper.

According to the WHO, chronic diseases rank highly in global incidence, prevalence, and mortality. In 2001, approximately 60% of the 56.5 million reported deaths and 46% of the global disease burden was attributed to chronic diseases. By 2020, the burden of NCDs is expected to increase to 57%. Nearly half of the deaths caused by chronic diseases are attributable to cardiovascular diseases. Moreover, diabetes and obesity are becoming a reason to worry since they already affect a large proportion of the population, and they are starting to appear at earlier stages of life (WHO, 2007). Furthermore, the situation regarding chronic diseases in Kenya is also worrisome. According to the economic survey in 2018 by KNBS, various types of cancer, HIV/AIDS and heart diseases were among the top ten causes of death in Kenya, all of which are chronic diseases. There was a total of 29, 251 deaths recorded by the KNBS from just these chronic diseases in 2017 (Njugunah, 2018).

Health damaging activities are a leading contributor to chronic diseases (Shiel, 2018). According to the IHME website, among the top ten risk factors that cause the most deaths and disability include unsafe sex, alcohol use, tobacco use and poor diets (IHME, 2019). These risk factors are all contributors to chronic diseases, and they play a crucial role in studies that model (chronic) diseases. The risk factors can potentially play the role of the explanatory variables in the modeling process. Understanding these factors would help in explaining how the burden of chronic diseases can be reduced and/or eradicated.

Among all the serious illnesses is diabetes, a non-communicable chronic disease. As of 2016, diabetes was the seventh leading cause of mortality while it never featured anywhere among the top 10 in the year 2000 (WHO, 2018). Diabetes Mellitus (scientific name for Type 2 Diabetes) is one of the four major NCDs leading to about 4 million deaths in 2017. Low income countries are expected to face a 92% increase in mortality as a result of diabetes by the year 2040. (Mohamed *et al.*, 2018). Specifically, in Kenya there are around 552, 400 cases of diabetes. (IDF, 2020) It is therefore evident that diabetes is a severe chronic disease with high prevalence and mortality not only in Kenya but globally. As a result, the primary aim of this study will be to model diabetes in Kenya using the GLM.

Problem Statement

As mentioned above, chronic diseases are among the leading causes of mortality in Kenya. Information from the Kenya STEPwise Survey for Non Communicable Diseases Risk Factors 2015 report mentions that the WHO states that cardiovascular diseases, various types of cancer, diseases involving the respiratory system, and diabetes result in 82% of all NCD related deaths, with around 31.2 million deaths annually, on a global scale (Ministry of Health, 2015), all of which are chronic diseases. In terms of diabetes specifically, the WHO predicts a rise in the estimate of the prevalence of diabetes in Kenya from 3.3% to 4.5% by 2025. (Jones, 2013). Furthermore, in East Africa the average cost of care for a patient diagnosed with type 1 diabetes is $229 per annum, with insulin purchases taking up around 60-70% of that amount. Kenya does not have the funds for diabetes care and/or prevention (Jones, 2013), which leaves many Kenyan diabetics with

the risk of poorer health and poverty. From the aforementioned statistics and information and from information about diabetes that is common knowledge, we can state that diabetes is a lifelong disease that is expensive to live with. Thus, using the modeling process to better understand the risk factors of diabetes will provide invaluable information for tailor made strategies including lifestyle changes. Furthermore, the information obtained from the modeling process may also help insurance companies engineer products for these specific groups of individuals thus improving their quality of life and lifespans.

Studies regarding (specific) NCDs and chronic diseases have been carried out multiple times in Kenya with reports including but not limited to the Kenya Demographic Health Survey (KDHS) in 2014 and the Kenya STEPwise Survey for Non Communicable Diseases in 2015. Moreover, a study of chronic diseases using the GLM was carried out by a researcher in Ghana titled "Analysis of Chronic Disease Conditions in Ghana Using Logistic Regression". However, this study aims to contribute to Kenya in a similar way to this Ghanaian study by determining risk factors and prevalence of diabetes through the use of the GLM, and therein lies the gap that this study aims to fill.

This study is categorized under Disease Modeling and involves working with the GLM. The results from the modeling process can regurgitate useful information regarding the problem at hand which can only be done with a reliable execution of the modeling and analysis processes.

Research Objectives and Significance of the Research

The primary objective of this study is to employ the GLM to model diabetes, a prevalent chronic disease in Kenya. The model and the information obtained from it are required to complete more specific objectives of this study.

One of those research objectives is to determine the individual hazardousness (level of significance) of the risk factors leading to chronic illnesses, more specifically diabetes. This objective is in line with my area of study as it requires a reliable execution and implementation of a regression analysis such that reliable results are obtained regarding the risk factors. A regression analysis explains the variation in the response variable as a result of the variation in the independent variables. As mentioned earlier, the risk factors can play the role of explanatory (independent) variables in the regression model. With a suitable response variable, using the regression analysis we will be able to determine which of the risk factors cause the most variation and which ones cause the least, and hence determine the hazardousness of each. This information can be beneficial to the population as a whole, especially those with pre-existing conditions, since understanding which of the risk factors leads to the highest variation can be useful in determining lifestyle changes as a measure to avoid one from getting infected with a chronic disease or from worsening their condition if they are already diagnosed with diabetes or any other chronic disease. The information can also be useful for insurance companies in making policies targeted at specific parts of the population.

With millions of annual deaths and cases of chronic diseases increasing annually, as mentioned earlier in this chapter, the severity of these illnesses can and should not be overlooked. Hence, this justifies the importance of this kind of study.

Chapter 2: Literature Review

A number of recent studies have focused on chronic diseases establishing the importance of addressing this topic. While there have been many studies involving chronic diseases, few researchers have modeled and done mathematical analyses of diseases (chronic and acute). This section will discuss from literature what studies have said about chronic diseases; it will discuss the (statistical) methodologies that have been used to model or analyze diseases in the past, it will discuss a study that was done in Kenya regarding diabetes, and it will outline a study that has incorporated the GLM, focused on chronic diseases.

Firstly, we will briefly discuss what studies have said about chronic diseases. I will start by looking at the burden of chronic diseases faced by the African continent. The WHO predicts that Africa will experience the largest increase in mortality from cancer, respiratory diseases, cardiovascular diseases, and diabetes over the next ten years (de-Graft Aikins *et al.*, 2010). As a result of this projection, in my opinion, it is justifiable to use one of these four diseases for this study. An article called "Tackling Africa's chronic disease burden: from the local to the global", discusses the "medical, psychological, community and policy dimensions of the chronic disease burden in sub-Saharan Africa" (de-Graft Aikins *et al.*, 2010) emphasizing that chronic diseases are a burden to multiple African nations. The article concludes that there is a need for primary and secondary interventions and a need for relevant African authorities to focus on policies surrounding chronic diseases (de-Graft Aikins *et al.*, 2010). Another study titled "Barriers to managing chronic illness among urban households in coastal Kenya" aimed at describing management of chronic diseases at a household level in urban, coastal Kenya and the barriers the households faced to successful management (this study focused only on Kenya, not multiple African nations). The study involved focus group discussions and interviews with 22 households to identify barriers to chronic illness care. The report stated that the costs associated with chronic illnesses are the main barrier to its management. Other barriers included: "patient knowledge and beliefs, stigma, quality and trust in providers and long care pathways." (Porter *et al.*, 2009). Within the paper, a comparison

was also made of the results from the study of Porter *et al.* with existing literature from the same field done in other settings. From the aforementioned discussions (including points mentioned in the earlier parts of this chapter), the issue surrounding the severity of chronic diseases becomes evident and thus establishing a rational justification to focus my research project on this topic.

Various studies have been conducted in the past to model or analyze various diseases (or disease classes) using a variety of methodologies. It is obviously impossible to be able to review all methodologies that have been used, however looking at a few can give a modest idea of how the modeling process was completed.[1] A study titled "Dietary Fat and Coronary Heart Disease: A Comparison of Approaches for Adjusting for Total Energy Intake and Modeling Repeated Dietary Measurements" involved examining associations between intakes of four major types of fat and risk of coronary heart disease over a period of 14 years (1980-1994) using alternative methods for energy adjustment. The authors compared four risk models for energy adjustment and within each model, the authors compared four different approaches for analyzing repeated dietary measurements (Hu *et al.*, 1999). The primary exposures for this study were the four main types of fat (saturated, monounsaturated, polyunsaturated, and trans fats), which were all correlated with total energy intake. As mentioned above, four methods were used to adjust for total energy, the explanations for which were taken straight from this study's report and can be found in Appendix 1. To analyze the repeated dietary measurement, the statistical model used was *the pooled logistic regression model.* This involved treating each 2 year examination interval as an independent follow-up study. "Observations over all intervals are pooled into a single sample, and a logistic regression is used to examine the association between the risk factors and development of disease during all follow-up periods" (Hu *et al.,* 1999). Further details regarding the analysis of the repeated dietary measurements can be found in Appendix 2. That basically summarizes the methodology that was utilized for this study.

Another useful article I came across, which was mentioned in the problem statement, involves both chronic diseases and the GLM, and was carried out in Ghana. The aim of

[1] It should be noted that gaining full understanding of the studies reviewed will require knowledge about the diseases/conditions that are being modeled or analyzed.

the paper is to investigate the effect of socio-demographic/socio-economic factors of Ghanaians on their chronic disease conditions. This study used data from the WHO study on AGEing and Adult Health, Wave 1, 2008-9. Each chronic disease considered within this study played the role of the dependent variable, and they included depression, oral health, injuries, and asthma. The predictor variables included age, gender, occupation, income level, ethnicity, marital status, and religion (Azumah Karim et al., 2008). After the regression analysis was performed, the prevalence of the individual chronic conditions could be determined and hence the association of the socioeconomic factors on chronic conditions in Ghana was established.

A study focused on diabetes and pre-diabetes was carried out in Kenya, titled "Prevalence and factors associated with pre-diabetes and diabetes mellitus in Kenya: results from a national survey". Within this paper, the prevalence of diabetes and pre-diabetes was determined using descriptive statistics. In terms of prevalence, the study showed a rather low prevalence of diabetes at just 2.4%, however a large proportion of people, specifically 52.8%, are not diagnosed with diabetes but face the possibility of developing issues. Additionally, the prevalence of diabetes in Kenya is relatively higher than its western neighbor, Uganda, with a 1.4% of diabetes prevalence. The study also determined a higher prevalence of pre-diabetes when compared to diabetes. The paper also arrives at a conclusion that the levels of diabetes treatment are low with just 21.3% of patients reporting that they are on treatment and just 41% among those aware of their diagnosis were on treatment (Mohamed *et al.,* 2018). These statistics indicate the relative severity of diabetes and also discuss the issues of diabetes treatment and awareness. This further gives importance to the determination of risk factors and raising awareness, not only for diabetes, but for chronic diseases as a whole.

The papers mentioned in the discussions above highlight the potential ways in which a modeling process can be done. It provides insights on the modeling process which is achieved by seeing how they did their regression analysis, which variables they used in their regression and why those variables were the appropriate ones. Since this research project also involves using the GLM (and) to carry out modeling, these studies were extremely relevant to this project's area of study and they have underlined what areas within this field have already been covered.

Knowledge Gap

As established within this report, the severity surrounding diabetes and chronic diseases cannot be overemphasized. Multiple research articles have had chronic diseases as their area of study. Moreover, there have also been studies within which chronic diseases and chronic conditions have been modeled and analyzed using statistical methodologies such the Ghanaian study that used the Logistic Regression. There have also been studies carried out in the past in Kenya whose area of interest was chronic diseases and diabetes. This study aims to further contribute to existing information within this area of study, with the research gap (or the objectives the study is trying to accomplish) of the study being the determination of the significance of potential risk factors associated with diabetes.

In order to contribute to the aforesaid research gap, this study intends to use the Generalized Linear Model to model one of the prevalent chronic diseases in Kenya. The (non-communicable), chronic disease that this study will focus on will be diabetes. Information from the WHO states that diabetes is now a growing problem in developing nations and that over 80% of the 1.5 million global deaths caused by diabetes in 2012, occurred in low/middle-income nations (WHO, 2014). Furthermore, information from the IDF states that there are now around 552,400 adults in Kenya that have been diagnosed with diabetes (IDF, 2020). Hence, diabetes is clearly a severe issue worth attention. With the aim of contributing to this research gap effectively, the data that will be used in the modeling process will specifically be Kenyan based.

Chapter 3: Research Methodology

Research Design

This research project has a correlational research design, implying that it involves the use of data to determine the relationship between a set of variables, which is ultimately the type of research design that regression analyses fall under. The aim is to identify the association between the independent variables and dependent variable and use that information to make interpretations that will be useful in achieving the research objectives of this study.

The Generalized Linear Model and the Conceptual Framework

For this analysis, we are going use a regression with the response variable being the diabetic status of the individual, and the explanatory variables will be a number of the risk factors that can cause diabetes.

While defining and explaining GLMs below, I will assume an understanding of simple linear regression models. [2]There are 3 components to any GLM. Firstly, we have the Random Component which denotes the distribution of the response variable Y_i. In GLMs, we assume the response variable follows any distribution from the exponential family, as opposed to standard linear regression models which assume a normal distribution. Secondly, we have the Systematic Component that refers to the linear predictor. In GLMs we have a variety of explanatory variables $(X_i, i = 1, \dots, k)$ that are linear in parameters and they vary simultaneously. These are called the covariates; the function of the covariates is referred to as the linear predictor (for example $\beta_0 + \beta_1 x_1 + \beta_2 x_2$). Lastly, we

[2] Information regarding the Generalized Linear Model is sourced from the Statistics site of PennState University, with the following citation:
PennState University. (2018). *Introduction to Generalized Linear Models*. Retrieved from STAT 504 : https://online.stat.psu.edu/stat504/node/216/

have the link function which states the relationship between the linear predictor and the distribution's mean. In simple linear models, the mean is equivalent to the linear predictor. Estimating the parameters allows us to calculate the mean of the distribution and hence make predictions regarding the response variable. To estimate the parameters, one can use the Maximum Likelihood Estimation (MLE) method. Other methods can also be used.

The main reason why the GLM is more suitable for this study, as compared to the simple linear model, is because the response variable in a simple linear model is assumed to always be normally distributed which might not necessarily be the case in our study. Moreover, the models are fitted through the MLE method and as a result of this, the estimators will display optimum properties.

Information from the Mayo Clinic[3] states that there are two kinds of chronic diabetes, there is type 1 diabetes and type 2 diabetes. Additionally, there are two kinds of potentially reversible diabetes, pre-diabetes, and gestational diabetes. Pre-diabetes is a state when one's blood sugar levels are high, but not as high as a diabetic and gestational diabetes occurs during pregnancy but can resolve after delivery (MayoClinic, 2018). "Type 2 diabetes is far more common, accounting for about 90% of all diabetes worldwide" (WHO, 2014). The risk factors for each type of diabetes differ. For an understanding of which risk factors lead to which type of diabetes, see the following table:

[3] The information in this section regarding the types of diabetes and the risk factors of diabetes are from the Mayo Clinic website, with the following citation:
MayoClinic. (2018, August 8). *Diabetes*. Retrieved from Mayo Clinic:
https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

Type of Diabetes	Risk Factors
Type 1 Diabetes	• Family history • Environmental factors • Presence of damaging immune system cells • Region/location
Type 2 and Prediabetes	• Weight • Lack of exercise or activity • Family History • Race – it is unclear why • Age • History of gestational diabetes • Polycystic ovary syndrome • High blood pressure • Abnormal cholesterol and triglyceride levels
Gestational Diabetes	• Age • Family or personal history • Weight • Race

Table 1: Diabetes Risk Factors

Coming back to the model, the response variable, i.e., the diabetes status, is binary since it can have only 2 possible outcomes. For that reason, we will use a Binary Logistic Regression since it models how a binary response variable is explained by a set of explanatory variables.[4] As mentioned earlier, all GLM's have 3 key components (Random Component, Systematic Component and Link Function). For this regression, the distribution of the response variable (i.e., the random component) will be the Binomial Distribution,

$$Y \sim Binomial(n, p), \quad where\ p\ is\ the\ probability\ of\ success$$

[4] PennState University. (2018). *Introduction to Generalized Linear Models*. Retrieved from STAT 504 : https://online.stat.psu.edu/stat504/node/216/

The systematic component refers to the linear predictor, which will be a function of the k explanatory variables. Lastly, we have the link function, which under a Binary Logistic Regression is the Logit link, which is:

$$logit(p) = log\left(\frac{p}{1-p}\right)$$

The coefficients produced from the GLM will then be logits and can be converted to odds ratios by being exponentiated.

The x's are the explanatory variables which can be continuous, discrete, or even both. They are assumed to be linear in parameters.

This fundamentally summarizes the key components that will be involved in our regression analysis.

Population, Sampling and Data Collection

The data that was used for this research project is from the 2014 KDHS report. The details regarding the sampling procedure carried out when the study for the 2014 KDHS was being done can be found in Appendix 3.

For this research project, the data was simply taken from the KDHS report as mentioned above. However, the primary source of data for the 2014 KDHS report itself was questionnaires. When the questionnaires were completed they were sent to the KNBS data processing center. It needed to be ensured that the data was consistent with the sample list and then that the data entered into their system was accurate and complete. A detailed explanation of the data collection for the 2014 KDHS can be found in Appendix 4.

The modeling process is going to be performed using the R programming language. The data collected from the 2014 KDHS report is raw data and is stored in an Excel file in a csv format such that it can be read into R with ease.

The dataset has ten variables, each represented by a certain factor index; the index is explained in the following table:

VARIABLE	No. OF LEVELS	FACTORS
Age	NA	Continuous
Region	8	1 = "Coast" 2 = "North Eastern" 3 = "Eastern" 4 = "Central" 5 = "Rift Valley" 7 = "Western" 8 = "Nyanza" 9 = "Nairobi"
POResidence	2	1 = "Urban" 2 = "Rural"
Education Level	4	0 = "No Education" 1 = "Primary" 2 = "Secondary" 3 = "Higher"
Wealth_Index	5	1 = "Poorest" 2 = "Poorer" 3 = "Middle" 4 = "Richer" 5 = "Richest"
Marital Status	6	0 = "Never in union" 1 = "Married" 2 = "Living with partner" 3 = "Widowed" 4 = "Divorced" 5 = "No longer living together/Separated"

High_Blood_Pressure	2	0 = No 1 = Yes
Diabetes	2	0 = No 1 = Yes
Gender	2	0 = "Female" 1 = "Male"
Exercise	2	0 = No 1 = Yes

Table 2: Raw Data Index

Chapter 4: Data Analysis and Interpretation

Variable Selection

As mentioned multiple times in this report, the explanatory variables in the regression are best suited to being the risk factors causing diabetes. In Table 1 we have information from literature, specifically from the Mayo Clinic, stating what leads to the different types of diabetes. The variables used in the regression, listed in Table 2, include the risk factors highlighted in Table 1 to provide a more comprehensive regression output. Variables such as Age, Region, Exercise, High Blood Pressure and POResidence are directly mentioned in Table 1. Other variables such as Gender, Wealth and Education level are also possible risk factors, however they are not mentioned in Table 1.

Parameter Estimation

The GLM uses the MLE method for the estimation of parameters.[5] The model we will produce will therefore display optimal properties of the estimators, seeing that it is an inherent advantage of using the GLM as a regression tool. This in turn increases the reliability of the output of the model.

[5]PennState University. (2018). *Introduction to Generalized Linear Models*. Retrieved from STAT 504: https://online.stat.psu.edu/stat504/node/216/

Output of the GLM

The GLM used was aimed at predicting the effect of all other variables on the 'Diabetes' variable. This GLM was stored into a string variable labeled 'logistic', the summary output of the model is displayed below:

```
> summary(logistic)

Call:
glm(formula = Diabetes ~ ., family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0698  -0.1015  -0.0849  -0.0730   3.6264

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -5.969212   0.372427 -16.028  < 2e-16 ***
Age                  0.026337   0.007712   3.415 0.000638 ***
Region2              0.845070   0.299611   2.821 0.004794 **
Region3              0.119629   0.233584   0.512 0.608550
Region4              0.024657   0.252583   0.098 0.922235
Region5             -0.278040   0.223748  -1.243 0.213998
Region7              0.098673   0.280005   0.352 0.724539
Region8              0.088197   0.252421   0.349 0.726786
Region9             -0.029881   0.357703  -0.084 0.933426
POResidence2        -0.231680   0.148929  -1.556 0.119793
Education.Level1    -0.510317   0.241506  -2.113 0.034595 *
Education.Level2    -0.253458   0.261275  -0.970 0.332007
Education.Level3    -0.288994   0.303141  -0.953 0.340422
Wealth_Index2        0.089606   0.228549   0.392 0.695011
Wealth_Index3       -0.056354   0.234173  -0.241 0.809825
Wealth_Index4       -0.110609   0.237690  -0.465 0.641680
Wealth_Index5        0.058621   0.255179   0.230 0.818306
Marital.Status1     -0.050076   0.197440  -0.254 0.799782
Marital.Status2     -0.496877   0.458841  -1.083 0.278855
Marital.Status3     -0.327092   0.427328  -0.765 0.444012
Marital.Status4     -0.270898   0.441918  -0.613 0.539873
Marital.Status5      0.230915   0.307737   0.750 0.453035
High_Blood_Pressure1 3.326369   0.139737  23.805  < 2e-16 ***
Gender1              0.301394   0.141346   2.132 0.032981 *
Exercise1           -0.002619   0.128171  -0.020 0.983698
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3099.2  on 27534  degrees of freedom
Residual deviance: 2409.2  on 27510  degrees of freedom
AIC: 2459.2

Number of Fisher Scoring iterations: 8
```

Figure 1: GLM Summary Output (SS)

Model Diagnosis and Interpretation

Model's R^2

To test for the model's significance, we can calculate the R^2 of the model and test for its significance using a Chi-square test and obtaining its *p* value. These steps can be carried out in R using the following lines of code:

```
> #We can then calculate the R^2 of the model...
> ll.null <- logistic$null.deviance/-2
> ll.logistic <- logistic$deviance/-2
> R.sqd <- (ll.null - ll.logistic)/ll.null
> R.sqd
[1] 0.2226279
> #...the p-value for that R^2 using a Chi-squared distribution
> 1-pchisq(2*(ll.logistic-ll.null), df=(length(logistic$coefficients)-1))
[1] 0
> #p-value is less than 0.05 so the R^2 value is significant
```

Figure 2: Model's R-squared (SS)

Note that individual calculations were stored into unique string variables in order to make the overall calculations easier to execute.

Deviances and the Akaike Information Criterion (AIC)

A low Null Deviance indicates that the Null Model appropriately models and explains the data. A low Residual Deviance indicates that the model used is appropriate. The Null Deviance of the model is 3099.2, which is quite high. This is a positive sign since it indicates that the Null model does not explain the data well and that additional estimators should be used to model the data. The Residual Deviance is 2409.2, which is lower than the Null Deviance (a good sign) but it is still relatively high. We can therefore conclude that the model used explains the dataset better than the Null Model, however the Residual Deviance is still relatively high possibly implying that the log likelihood of the model used is not close to the log likelihood of the Saturated Model. Furthermore, it can also be stated

that model is not 'well-fitted' since the value of the Residual Deviance is not near the value of its degrees of freedom.

The Akaike Information Criterion (AIC) basically refers to the quality of the model and how well it fits the data that was used to produce the model. The AIC for the model is 2459.2, which is high, indicating a suboptimal fit.

Interpretation of Coefficients and Statistical Significance

From a quick look at the output of the GLM, we note that the only statistically significant coefficients are those associated with High_Blood_Pressure1, Age, Region2, Gender1, and Education.Level1 (in order of statistical significance). The statistical significance is determined through the z-value and the p-value.[6]

The interpretation for each factor variable will be in terms of its reference category. The odds for the reference category for each factor variable are assumed to be 1. Exponentiating the coefficients for each category for each of the factor variables produces its effect relative to the reference category, on the response variable, i.e., Diabetes Status. Exponentiating the coefficients returns the odds ratio of the explanatory variable.

The explanatory variable 'High_Blood_Pressure' displays the largest effect. Exponentiating the coefficient for High_Blood_Pressure1 i.e., `exp(3.326369)` returns an odds ratio of `27.83708` indicating that someone with high blood pressure is 27.84 times more likely to report diabetes than someone who does not have high blood pressure.

In terms of the explanatory variable, 'Region' we can see that individuals from Region2 are most likely, specifically 2.33 more likely, to report diabetes compared to individuals from Region1 (the reference category). Region5, on the other hand, has the lowest odds ratio of 0.78 indicating that individuals from Region5 are least likely to report diabetes compared to individuals from Region1.

[6] Not statistically significant if the z value is less than 2 standard deviations from 0 and if the p value is greater than 0.05.

'Gender' is also a significant explanatory variable. From the model output, it can be said that a male is 1.35 times more likely to report diabetes than a female.

The coefficient for Education.Level1 produced the last statistically significant coefficient. Firstly, for this explanatory variable, I would like to point that all the coefficients are negative implying that anyone with some form of formal education reduces his/her likelihood of reporting positive for diabetes. The odds ratio for Education.Level1 is 0.60 displaying the least likelihood of reporting diabetes compared to the reference category of Education.Level0 (No Education).

The remaining coefficients did not display statistical significance. Using their logits and odds ratios to make interpretations and conjectures regarding an individual's likelihood of reporting diabetes could result in unreliable information.

Chapter 5: Conclusions and Recommendations for future study

Conclusion

From the output of the GLM and the interpretation we can conclude that the risk factors (limited to the risk factors used within this study) that have the largest effect on an individual's Diabetes status in Kenya are high blood pressure status, age, gender, and level of education. We restrict the interpretation of the output model to those variables that displayed statistical significance in order to avoid unreliable interpretations and conjectures. We thus conclude that individuals with high blood pressure, males, those with no formal education, and those living in the North Eastern region are most likely to contract diabetes.

In my personal opinion, I am surprised that the 'Exercise' variable did not display the property of statistical significance. Moreover, if we were to ignore the importance of that property and still interpreted the output, the likelihood of an individual reporting diabetes is almost the same whether they exercise or not. This surprised me since lack of activity and exercise is a genuine risk factor that can lead to type 2 diabetes and pre-diabetes, and additionally is it generally accepted that keeping fit and exercising can avoid chronic illnesses in an individual's future lifetime.

One of the objectives of this study was to use the results to inform potential insurance policies. Insurance companies can use the information to engineer specific policies targeted at certain members of the population for example, target those living in a certain region.

Recommendations for future study

Despite a lot of studies having already been done by multiple researchers, there is always room for further contribution to an area of study. Studies in the future can align their focus towards other diseases, whether chronic or acute.

Regarding diabetes, studies have been done using descriptive statistics to determine the prevalence of diabetes, using multiple regression models to estimate additive costs as a result of diabetes, and studies have been done using regression models to determine the effects of various factors on diabetes. However, these results are likely to become outdated in the coming years as a result of changes in lifestyles, attitudes, health care systems etc. That opens doors to researchers to redo similar studies to determine the changes in the respective outputs/results arising from changes in lifestyles, attitudes, health care systems etc. Additionally, the older studies can guide new researchers and their research methods to ensure reliable outputs and results. This argument can also be applied to all other diseases and also different fields of research.

Appendices

Appendix 1 – Methods to adjust for total energy (Study done by Hu *et al.*)[7]

1. Standard multivariate model, in which the independent exposures are absolute intake of fats (grams per day) and total energy intakes. The relation can be expressed as the following:

*Disease risk = s1 * saturated fat + s2 * mono + s3 * poly + s4 * trans + s5 * protein + s6 * total energy intake,*

where s1-s6 are regression coefficients

2. Nutrient residual model, in which the fat residuals obtained by regressing nutrient intake on total energy intake are included as independent variables. Total energy intake is also included as a covariate. The relation can be expressed as the following:

*Disease risk = n1 * saturated fat residual + n2 * monounsaturated fat residual + n3 * polyunsaturated fat residual + n4 * trans-fat residual + n5 * protein residual + n6* total energy intake.*

   The coefficients for the nutrient residuals have the same isocaloric substitution interpretation as does the standard multivariate model, and they are quantitatively the same. The coefficient for energy, however, retains the biologic meaning of total energy intake

3. Energy-partition model, in which total energy is partitioned into that contributed by different types of fats and that contributed by other sources (protein and

[7] All information in Appendix 1 is from a study titled "Dietary Fat and Coronary Heart Disease: A Comparison of Approaches for Adjusting for Total Energy Intake and Modeling Repeated Dietary Measurements"

carbohydrates). Because units as either calories or grams for the nutrients do not in any way affect the analyses, we model absolute intakes of all energy-bearing nutrients simultaneously:

$$\textit{Disease risk} = e1 * \textit{saturated fat} + e2 * \textit{monounsaturated fat} + e3 * \textit{polyunsaturated fat} + e4 * \textit{trans-fat} + e5 * \textit{protein} + e6 * \textit{carbohydrates}$$

In this model, the coefficient for one type of fat represents the effect of increasing absolute intake of this type of fat while holding constant other fats and other energy sources (i.e., protein and carbohydrates). Therefore, it represents the effect of "adding fat," which includes both its energy and nonenergy effects.

4. Multivariate nutrient density model, in which the nutrient densities (percentages of energy) from different types of fat are included as exposure variables and total energy is included as a covariate:

$$\textit{Disease risk} = m1 * \textit{percent energy from saturated fat} + m1 * \textit{percent energy from monounsaturated fat} + m3 * \textit{percent energy from polyunsaturated fat} + m4 * \textit{percent energy from trans-fat} + m5 * \textit{percent energy from protein} + m6 * \textit{total energy}.$$

The coefficients from this model also have an isocaloric interpretation, representing substitution for an equal energy from carbohydrate, but are in units of the percentage of energy.

Appendix 2 - Analysis of the repeated dietary measurements (Study done by Hu *et al.*)[8]

D'Agostino *et al.* showed the asymptotic equivalence of this approach (mentioned in the literature review) and the Cox regression model with time-dependent covariates. The necessary conditions for this equivalence include relatively short time intervals and small probability of the outcome in the intervals, both of which are satisfied here. Following D'Agostino et al., a general pooled logistic model can be written:

$$Logit\ p_i\big(X(t_{i-1})\big) = \alpha_i + \beta_1 x_1(t_{i-1}) + \cdots + \beta_p x_p(t_{i-1})$$

where $p_i\big(X(t_{i-1})\big)$ is the conditional probability of observing an event by time t_i given that the individual is event-free at time t_{i-1} and $X(t_{i-1})$ is the vector of independent variables measured at time t_{i-1}.

The above model assumes that only the current risk profile is needed to predict disease outcome. For dietary exposures, previous diet or long-term diet may be more important. Therefore, we compared the following approaches for handling repeated dietary measurements in the pooled logistic regression:

1. Baseline diet only, in which CHD incidence from 1980 to 1994 was related to 1980 fat intake.
2. Most recently measured diet, in which CHD incidence during the 1980-1984 time period was related to fat intake from the 1980 questionnaire, CHD incidence during the 1984-1986 time period was related to fat intake from the 1984 questionnaire, CHD incidence during the 1986-1990 time period was related to fat intake from the 1986 questionnaire, and CHD incidence during the 1990-1994 time period was related to fat intake from the 1990 questionnaire.
3. Cumulative average diet, in which CHD incidence between each 2-year questionnaire cycle was related to the cumulative average of fat intake calculated

[8] All information in Appendix 2 is from a study titled "Dietary Fat and Coronary Heart Disease: A Comparison of Approaches for Adjusting for Total Energy Intake and Modeling Repeated Dietary Measurements" with reference to a study titled "Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart Study" that was done in 1990 by D'Agostino *et al.*.

from all of the preceding dietary measures. Therefore, CHD incidence during the 1980-1984 time period was related to the fat intake from the 1980 questionnaire; CHD incidence during the 1984-1986 time period was related to the average intake from the 1980 and 1984 questionnaires; CHD incidence during the 1986-1990 time period was related to the average intake from the 1980, 1984, and 1986 questionnaires; and CHD incidence during the 1990-1994 time period was related to the average intake from all four dietary questionnaires.

4. As an alternative, to give more weight to the most recent diet, we calculated the average of the most recently measured diet and previous average diet. For 1980-1984 and 1984-1986 time periods, the exposures were calculated in the same way as described above. For CHD incidence during 1986-1990 time period, the exposures were calculated as the following:

[1986 diet + (1980 diet + 1984 diet)/2]/2 = 1986 diet/2 + 1980 diet/4 + 1984 diet/4.

Similarly, for CHD incidence that occurred during the 1990-1994 time period, the exposures were calculated as follows:

[1990 diet + (1986 diet/2 + 1980 diet/4 + 1984 diet/4)] 12= 1990 diet/2 + 1986 diet/4 + 1980 diet/8 + 1984 diet /8.

In both cumulative methods, we used the averages of fat intakes instead of a single measurement to reduce within-subject variation and best represent long-term diet. One method provides more weight to the baseline diet, while the other assumes that recent diet is more important.

Appendix 3 – Sampling procedure (2014 KDHS)

The sample for the 2014 KDHS was drawn from a master sampling frame, the Fifth National Sample Survey and Evaluation Program (NASSEP V). This is a frame that the KNBS currently operates to conduct household-based surveys throughout Kenya. Development of the frame began in 2012, and it contains a total of 5,360 clusters split into four equal subsamples. These clusters were drawn with a stratified probability proportional to size sampling methodology from 96,251 enumeration areas (EAs) in the 2009 Kenya Population and Housing Census. The 2014 KDHS used two subsamples of the NASSEP V frame that were developed in 2013. Approximately half of the clusters in these two subsamples were updated between November 2013 and September 2014. Kenya is divided into 47 counties that serve as devolved units of administration, created in the new constitution of 2010. During the development of the NASSEP V, each of the 47 counties was stratified into urban and rural strata; since Nairobi county and Mombasa county have only urban areas, the resulting total was 92 sampling strata. The 2014 KDHS was designed to produce representative estimates for most of the survey indicators at the national level, for urban and rural areas separately, at the regional level, and for selected indicators at the county level. In order to meet these objectives, the sample was designed to have 40,300 households from 1,612 clusters spread across the country, with 995 clusters in rural areas and 617 in urban areas. Samples were selected independently in each sampling stratum, using a two-stage sample design. In the first stage, the 1,612 EAs were selected with equal probability from the NASSEP V frame. The households from listing operations served as the sampling frame for the second stage of selection, in which 25 households were selected from each cluster." (KNBS, 2014, p. 5).

Appendix 4 – Data Collection: Questionnaires (2014 KDHS)[9]

The 2014 KDHS used a household questionnaire, a questionnaire for women aged between 15 and 49, and a questionnaire for men aged between 15 and 54. For the 2014 KDHS, data was collected from 40,300 households. To address concerns related to data quality and overall management for a survey of this size, to reduce the length of fieldwork, and to limit interviewer and respondent fatigue, a decision was made to not implement the full questionnaire in every household and, in so doing, to collect only priority indicators at the county level. Thus, a total of five questionnaires were used in the 2014 KDHS: (1) a full Household Questionnaire, (2) a short Household Questionnaire, (3) a full Woman's Questionnaire, (4) a short Woman's Questionnaire, and (5) a Man's Questionnaire. The 2014 KDHS sample was divided into halves. In one half, households were administered the full Household Questionnaire, the full Woman's Questionnaire, and the Man's Questionnaire. In the other half, households were administered the short Household Questionnaire and the short Woman's Questionnaire. Selection of these subsamples was done at the household level—within a cluster, one in every two households was selected for the full questionnaires, and the remaining households were selected for the short questionnaires. (KNBS, 2014)

[9] All the information in Appendix 4 is a paraphrased version of the explanation of how the questionnaires were used for the 2014 KDHS; all this information is from the report of the survey itself.

References

Azumah, Karim & Saeed, Bashiru & Darkwah, Kwaku & Musah, A. (2008). Analysis of Chronic Disease Conditions in Ghana Using Logistic Regression. International Journal of Statistics and Applications 2015, 5(4): 133-140. doi: 10.5923/j.statistics.20150504.01.

de-Graft Aikins, A., Unwin, N., Agyemang, C., Allotey, P., Campbell, C., & Arhinful, D. (2010). Tackling Africa's chronic disease burden: from the local to the global. Globalization and Health, 6(1), 5. doi:10.1186/1744-8603-6-5

Hu, F. B., Stampfer, M. J., Rimm, E., Ascherio, A., Rosner, B. A., Spiegelman, D., & Willett, W. C. (1999). Dietary Fat and Coronary Heart Disease: A Comparison of Approaches for Adjusting for Total Energy Intake and Modeling Repeated Dietary Measurements. American Journal of Epidemiology, 149(6), 531–540. doi:10.1093/oxfordjournals.aje.a009849

IDF. (2020, April 23). *IDF Africa Members*. Retrieved from International Diabetes Federation: https://idf.org/our-network/regions-members/africa/members/13-kenya.html

IHME. (2019). *Kenya*. Retrieved from IHME Measuring what matters: http://www.healthdata.org/kenya

Jones, T. L. (2013). Diabetes Mellitus: the increasing burden of. *South Sudan Medical Journal*, 60-64.

Kenya National Bureau of Statistics. (2014). *Kenya Demographic and Health Survey.* Nairobi.

KNBS. (2014). *Kenya Demographic and Health Survey.* Nairobi.

MayoClinic. (2018, August 8). *Diabetes*. Retrieved from Mayo Clinic: https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

Ministry of Health. (2015). *Kenya STEPwise Survey for Non Communicable Diseases Risk Factors.* Nairobi.

Mohamed, S. F., Mwangi, M., Mutua, M. K., Kibachio, J., Hussein, A., Ndegwa, Z., Owondo, S., Asiki, G., Kyobutungi, C. (2018). Prevalence and factors associated with pre-diabetes and diabetes mellitus in Kenya: results from a national survey. BMC Public Health, 18(S3). doi:10.1186/s12889-018-6053-x

National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP). (2019, October 23). *About Chronic Diseases*. Retrieved from Centers for Disease Control and Prevention (CDC): https://www.cdc.gov/chronicdisease/about/index.htm

Njugunah, M. (2018, April 28). *Pneumonia, Malaria And Cancer Are Kenya's Leading Causes Of Death*. Retrieved from Capital News: https://www.capitalfm.co.ke/news/2018/04/pneumonia-malaria-and-cancer-are-kenyas-leading-causes-of-death/

PennState University. (2018). *Introduction to Generalized Linear Models*. Retrieved from STAT 504: https://online.stat.psu.edu/stat504/node/216/

Porter, T., Chuma, J., & Molyneux, C. (2009). Barriers to managing chronic illness among urban households in coastal Kenya. Journal of International Development, 21(2), 271–290. doi:10.1002/jid.1552

Shiel, W. C. (2018, December 21). *Medical Definition of Chronic disease*. Retrieved from MedicineNet: https://www.medicinenet.com/script/main/art.asp?articlekey=33490

WHO. (2007, June 22). *2. Background*. Retrieved from World Health Organization: https://www.who.int/nutrition/topics/2_background/en/

WHO. (2014, November). *Kenya faces rising burden of diabetes*. Retrieved from World Health Organization: https://www.who.int/features/2014/kenya-rising-diabetes/en/