On the Relationship between Mortality and Higher Education in the Kingdom of Saudi Arabia

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

BY

HALIMAH ABDULLAH ALRASHDI

DR. MUNNI BEGUM-ADVISOR

BALL STATE UNIVERSITY

MUNCIE, INDIANA

DECEMBER 2016

On the Relationship between Mortality and Higher Education in the Kingdom of Saudi Arabia

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

BY

HALIMAH ABDULLAH ALRASHDI

Committee Approval:

____________________________ ______________________________

Committee Chairperson Date

____________________________ ______________________________

Committee Member Date

____________________________ ______________________________

Committee Member Date

Departmental Approval:

____________________________ ______________________________

Departmental Chairperson Date

____________________________ ______________________________

Dean of Graduate School Date

BALL STATE UNIVERSITY

MUNCIE, INDIANA

DECEMBER 2016

Dedicated

To

My parents

&

My husband

Acknowledgements

First and foremost, I thank Allah for endowing me with the strength, patience, and knowledge to complete this research. Acknowledgement is due to Ball State University for the support given through its excellent facilities and for granting me the opportunity to pursue my postgraduate studies.

I would like to express my deep and sincere gratitude to my supervisor, Dr. Munni Begum, Professor, Department of Mathematical Sciences, Ball State University, for her openhanded guidance, instruction, continuous support, detailed and constructive comments, endurance, and advice on all matters throughout the period of this research. Her logical way of thinking has been of great value for me. Without her help and encouragement this project report could never have been completed or materialized at all. Also, I am indebted to her open door policy for students to consult with her. In fact, I do not have words to express my gratitude to her.

I would also like to express my immense appreciation for Dr. Rebecca Pierce and Dr. Xiao Yayuan for agreeing to participate in this committee and for reviewing my paper.

I would like to thank all of my friends for their constant support and encouragement in the preparation of this report and their friendship during my graduate studies. Without their encouragement and understanding it would have been impossible for me to finish this work.

Finally, I would like to express my deepest gratitude to my husband for his moral support throughout my academic career and also for his love, patience, encouragement and prayers.

Contents

List of the Tables

List of the Figures

Abstract

THESIS: On the Relationship between Mortality and Higher Education in the Kingdom of Saudi Arabia.

STUDENT: Halimah Alrashdi

DEGREE: Master of Science

COLLEGE: Sciences and Humanities

DATE: December 2016

PAGES: 47

Research on the relationship between deaths and demographic and socio-economic factors is common in the social and behavioral science. In the current study, we investigate whether the proportion of deaths can be explained by age, sex, education and geographic location for the Kingdom of Saudi Arabia (KSA). We considered a secondary data set consisting of different parts of KSA. The study shows that the proportion of deaths is affected by geographic location. The proportion of deaths is significantly different in different cities of KSA. We also found that the proportion of deaths is higher in the male population. Individuals aged more than 24 years have a higher risk of deaths which is quite natural. Surprisingly, the data do not show any statistically significant difference in mortality for different levels of education.

The two candidate models, binomial model with probit link and beta regression model, depict the relationship between proportion of deaths and some other predictors. Although the literature suggests using a beta regression model for modeling the proportion, in this study the binomial model with probit link shows better results with maximum significant predictors.

.

Chapter 1

Introduction

From many years there has been an area of study to determine the association of mortality with socioeconomic status. There is an established yet striking relationship amongst health and education [1]. Crossway over genders, races and time, more-instructed individuals appreciate preferred wellbeing over less-taught individuals. Indeed, even at lower levels of training, these connections are solid. A more elevated amount of instruction is a solid indicator of life span because of numerous elements, including higher salary and economic wellbeing, more beneficial practices, and enhanced social and mental prosperity [1]. A study by scientists at the University of Colorado Denver, New York University, and the University of North Carolina at Chapel Hill evaluates the quantity of deaths that can be connected to contrasts in instruction, and finds that variety in the danger of deaths over training levels has extended extensively. Michael Grossman's case that "years of formal tutoring finished is the most vital connect of good health" [2]. From a wellbeing point of view, the relationship recommends that training could be an intense apparatus for enhancing wellbeing, particularly given the vagueness in the profits to extra human services spending [3].

Instruction may directly affect wellbeing and wellbeing practices by means of its impact on beneficial and allocative productivity [4]. That is, training may confer direct information about wellbeing and wellbeing practices, in this manner moving the wellbeing generation capacity. What's more, instruction could change the portion of wellbeing inputs.

Chapter 2

Literature Review

Investigation of effect of wage disparity has been focal point of past examination. In numerous nations, wellbeing contrasts crossway training gatherings are striking. For example, in the United States in 1999, the deaths rate among working-age grown-ups with precisely twelve years of instruction was twice as high as among those with thirteen or more years of training [5]. Epidemiological studies have demonstrated a relationship between financial elements and wellbeing. It has not been settled whether this affiliation is entirely or halfway free of traditional danger elements. The relationship between financial components, hazard elements and mortality is very much perceived for some maladies [6-9]. These epidemiological studies are suited to evaluate the relationship between instructive level and mortality. This study was intended to study whether there is any distinction in mortality between higher levels of education, after adjusting for place of residence in terms of cities, and demographic factors such as sex and age groups.

Deaths rates for Americans ages 25 to 64 who have gone to school are not as high as the rates for the individuals who halted instruction in the wake of finishing secondary school. In 1999, the latest year for which deaths rates are available across different levels of education, there were 219 deaths for each 100,000 individuals for those with 13 or more years of training, contrasted with 474 for each 100,000 for those with 12 years and 585 for each 100,000 for those with less than 12 years of instruction. These distinctions are to some degree more prominent for men than for women. [6-9]

The mortality advantage for Americans with advanced education has been developing in late decades, as indicated by a few studies that utilized enumeration and overview information. What's more, for educated individuals there is an additional point of preference in rates of ailment and handicap, so that those with some school training appreciate additional years of sound life. Two components that somewhat clarify that institutional training leads to favorable position are way of life (educated individuals are less inclined to smoke or participate in other dangerous practices) and medical coverage (individuals with more training have a tendency to have preferable scope of coverage compared to those with less education). [9]

The motivation behind this research is to figure out if level of education causally affects wellbeing, specifically mortality. The negative relationship amongst training and mortality, the most essential measure of wellbeing, has turned out to be entrenched subsequent to the celebrated Kitagawa and Hauser [10] study, which discovered huge contrasts in deaths rates crossway over instructive classes for both genders. Later studies [11] affirm these discoveries. Elo and Preston [12] control for an assortment of other mortality elements, for example, pay, race, conjugal status, district of living arrangement, and area of birth. Rogers et al. [13] further control for access to human services, protection, smoking, activity, occupation, and other components.

Steady proof has demonstrated that financially advantaged people, whether communicated as far as training, pay, or occupation, have a tendency to have preferred wellbeing over hindered people [5]. By and large, these financial wellbeing differentials are not restricted to a little minimized gathering of society yet are communicated as a slope over the full range of social stratification [14]. The instrument’s fundamental purpose is social designing of infection and the way contributing elements are interrelated are still considered to be ineffective. Among the most widely recognized clarifications are those underlining 1) material conditions (e.g., absence of essential pleasantries and access to administrations), 2) way of life and behavioral variables, and 3) mental understandings that accentuate immediate and roundabout impacts of anxiety because of being lower in the financial progressive system [15].

In this study, we are interested in the proportion of deaths of Saudi people across different cities and to see whether there is any association with gender, age, or education level. We will compare the proportion of deaths for different categories of different predictors and search for the significance differences in the first part of methodology. In later part, we will establish the relationship of the response with the predictors using a binomial model with different link functions and with the beta regression.

Chapter 3

Methodology

3.1 Data

We considered data to study the impact of institutional education on mortality. The data was retrieved from the website of Ministry of Education of Saudi Arabia. The survey was conducted under the Population & Housing Census 1425 (Anno Hegirae or Hijri calendar) or equivalently 2003 in Gregorian calendar.

The data has been collected from 13 cities in Saudi Arabia. There are different sections of the survey such as, demographic characteristics (Gender, Governorate, Nationality), education characteristics (Enrolled by Governorate and Schooling Stage, Field of Specialization), economic characteristics (Activity Status, Employment Status, Main Occupation), and disability (Disability Status by Gender, Age Groups and Marital Status, Type of Disability). For our study, we have selected Governorate (Administrative Area), Gender, Age, Education Status, total number of population and deaths. For simplicity, we have re-organized age group into two categories and educational status into three categories.

3.2 Software Tools

The data have been analyzed using R software (https://www.r-project.org/), a free software environment for statistical computing and graphics. MS excel also has been used to do some initial calculations.

3.3 Statistical Analysis

Statistical Analysis

In order to analyze the data, we summarize the main characteristics of the data, which is called exploratory data analysis (EDA). Primarily EDA is conducted for examining what the data can tell us before conducting formal modeling or hypothesis testing task. EDA is usually conducted to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.

Univariate analysis provides summary measures for each variable of interest in the data. In other words, univariate examination investigates every variable in a data set, independently. It looks at the range of values, as well as the central tendency of the values [16-24]. It describes the pattern of a variable and its distribution. Univariate analysis is usually performed for facilitating more complicated analyses such as, bivariate and multivariate analysis. Univariate descriptive statistics describe individual variables. In this section, we present appropriate numerical descriptive statistics and graphical methods, which summarize the data. We present bar diagrams for graphical representations of the descriptive statistics. We present numerical descriptive statistics in a tabular form and discuss related findings.

Bivariate analysis, explores the concept of association between two variables [16-24]. Association is based on how two variables simultaneously change together; the notion of co-variation. Bivariate descriptive statistics involves simultaneously analyzing (comparing) two variables to determine if there is a pairwise relationship between the variables. The purpose of this part is to go beyond univariate statistics, in which the analysis focuses on one variable at a time. In order to carry out this analysis, we generated cross tabulations for finding association among variables. Initially, we test whether two variables are associated or not. If two variables are associated, then we find strength of this association using appropriate statistics.

3.3.1 ANOVA

Analysis of variance (ANOVA) is specific example of general linear models where the response is continuous and the predictors are categorical. ANOVA is used primarily to test if there is statistically significant difference among the group means of the response variable. In its most straightforward structure, ANOVA gives a factual test regardless of whether the method for a few gatherings are equivalent, and in this manner generalizes the t-test to more than two means. ANOVAs are valuable for looking at (testing) three or more means (groups or variables)

3.3.2 Modeling for Proportions

In statistics, regression is a methodology for demonstrating the relationship between a scalar or vector dependent variable y and one or more informative variables (or autonomous variables, predictors or independent variables) denoted by X. In other words, regression analysis is a statistical process for estimating the linear or nonlinear relationships among variables. Regression models focus on the relationship between a dependent variable and one or more independent variables (or 'predictors') [16-24].

When the response or dependent variable is categorical with two levels, regression models for binary outcomes are used. These models include logistic regression, or logit model, probit model and cloglog model. Alternatively, Beta regression models can be considered for the proportion of the response of interest. In this research, to study the relationship between deaths and demographic and socio economic factors, we considered regression models for binary outcomes and Beta regression model for proportion.

We want to describe the proportion of successes, $P_i = \frac{Y_i}{n_i}$, in each subgroup in terms of factor levels and other explanatory variables which characterize the subgroup. Since $E(Y_i) = n_i\pi_i$, $E(P_i) = \pi_i$. So, we model the probabilities π_i as

$$g(\pi_i) = x_i^T\beta$$

where x_i is a vector of explanatory variables (dummy variables for factor levels and measured values for covariates), β is a vector of parameters and g is a link function [16-24].

The simplest case is the linear model $\pi = x_i^T\beta$.

Regression models for binary outcome can be seen as instances of the generalized linear models and also can be derived using the notion of a tolerance distribution. The idea behind a tolerance distribution is from bioassay where animals are set to different levels of exposure and a response of 'yes' or 'no' is recorded. If we define *D* to be the minimum exposure required to produce a response in a subject, then *D* is a random variable. Its distribution is called the tolerance distribution, and is denoted by F_D. A response is obtained at d_i if and only if the tolerance is less than or equal to d_i, i.e., $D \leq d_i$. Thus, the probability π_i of deaths at exposure d_i is given by

$$\pi_i = P(D < d_i) = F_D(d_i)$$

Different distributional assumptions on the tolerance distribution lead to different models as discussed below.

Probit analysis

Based on the assumption of a normal tolerance distribution, i.e.

$$D \sim N(\mu_d, \sigma_d^2)$$

then

$$\pi_i = P(D < d_i) = \Phi\left(\frac{d_i - \mu_d}{\sigma_d}\right),$$

Where Φ is the $N(0,1)$ cdf. Thus,

$$\Phi^{-1}(\pi_i) = \frac{d_i - \mu_d}{\sigma_d} \equiv \beta_0 + \beta_1 d_i,$$

Where $\beta_0 = -\mu_d/\sigma_d$ and $\beta_1 = 1/\sigma_d$.

Hence, the mean and standard deviation of the tolerance distribution can be estimated by the regression parameters in this model.

Logit analysis

Assume now that the tolerance distribution is the logistic distribution, i.e.

$$f_D(d) = \frac{\exp\{(d - \mu_d)/\tau\}}{\tau[1 + \exp\{(d - \mu_d)/\tau\}]^2},$$

Where $-\infty < d < \infty, -\infty < \mu_d < \infty$ and $\tau > 0$. Then $E(D) = \mu_d$, and $Var[D] = \pi^2\tau^2/3$

(where $\pi = 3.31415 \dots$). Under this assumption,

$$\pi_i = F_D(d_i) = \frac{\exp\{(d_i - \mu_d)/\tau\}}{1 + \exp\{(d_i - \mu_d)/\tau\}}$$

$$\Rightarrow \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \frac{d_i - \mu_d}{\tau} = \beta_0 + \beta_1 d_i$$

Where $\beta_0 = -\mu_d/\tau$ and $\beta_1 = 1/\tau$.

Complementary log-log analysis

Similarly, the assumption of the extreme-value tolerance distribution

$$F_D(d_i) = 1 - e^{-e^{\frac{a-d_i}{b}}}$$

(where b<0) leads to the complementary log-log model. [25]

The above model for binary outcome is written with exposure variable *d_i*, which can be generalized into models including more than one variable. In the following section, we describe the general linear logistic regression model. It is to be noted that probit and cloglog models can also be written similarly.

3.3.3 General logistic regression model

The simple linear logistic model $\log[\pi_i/(1 - \pi_i)] = \beta_1 + \beta_2 x_i$ used in earlier discussion is a special case of the general logistic regression model

$$\text{logit } \pi_i = \log[\pi_i/(1 - \pi_i)] = x_i^T \beta$$

where x_i is a vector of continuous measurements corresponding to covariates and dummy variables corresponding to factor levels and β is the parameter vector .

This model is very widely used for analyzing data involving binary or binomial responses and several explanatory variables. When the response variable is considered as proportion instead of a binary variable, we can consider a Beta regression model which is discussed below.

3.3.4 Beta Regression

The linear regression model is not suitable for circumstances where the response variables is restricted to the interval (0, 1) [16-24]. A possible procedure is to transform the dependent variable and model the mean of the transformed response on the exogenous variables. This methodology, in any case, has downsides, one of them being the way that the model parameters can't be effortlessly deciphered. Another weakness is that the measures of proportions generally show asymmetry, and consequently inferring the normality assumption mislead. We are attempting to utilize a regression model that is customized for circumstances where the dependent variable (y) is measured persistently on the interval (0,1), i.e. $0 < y < 1$. The Beta distribution, is extremely adaptable for displaying extents since it can have entirely distinctive shapes relying upon the estimations of the two parameters.

Chapter 4

Analysis of Data

4.1 Introduction

In this chapter, we analyze the data from different perspectives. At first, we try to get some idea about the data using simple statistical analysis. We use some tables, graphs, and cross tabulations to see the pattern and behavior of the data. Rates, proportion and percentages also will be used to show the characteristics of the data.

4.2 Basic Statistical Analysis

4.2.1 Distribution of population according to city

Table 1: Number of people according to city

	City	Freq	Cum Freq	Percentage	Cum Percentage
1	Aseer	1056663	1056663	8.68	8.68
2	Baha	246915	1303578	2.03	10.71
3	East	1904742	3208320	15.66	26.37
4	Hail	336816	3545136	2.77	29.14
5	Jizan	736236	4281372	6.05	35.19
6	Jouf	220473	4501845	1.81	37
7	Madina	812349	5314194	6.68	43.68
8	Makkah	2699208	8013402	22.18	65.86
9	Najran	246060	8259462	2.02	67.88
10	North	172116	8431578	1.41	69.3
11	Qassem	596223	9027801	4.9	74.2
12	Riyadh	2734884	11762685	22.48	96.68
13	Tabouk	404271	12166956	3.32	100

The table above shows the population distribution among the cities. The maximum populations occur in Makkah and Riyadh. Figure 1 below shows the population structure graphically.

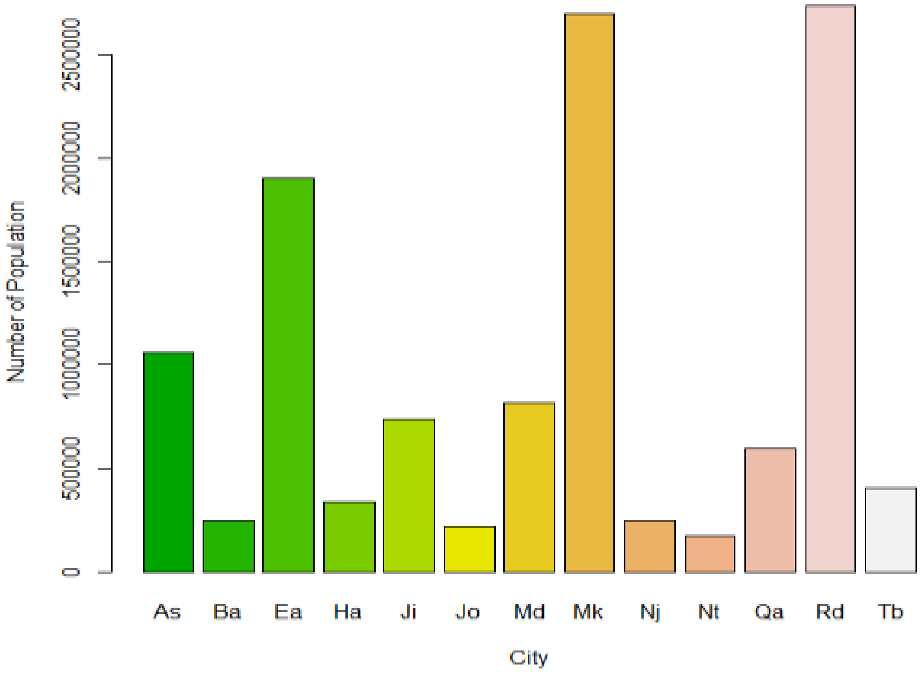

Figure 1: Population according to City

4.2.2 Distribution of number/proportion of deaths according to city

The table below shows the number of deaths in different cities. Number of deaths is highest in Makkah.

Table 2: Number of deaths according to city

	City	Freq	Cum Freq	Percentage	Cum Percentage
1	Aseer	8325	8325	16.34	16.34
2	Baha	1431	9756	2.81	19.15
3	East	6363	16119	12.49	31.64
4	Hail	1557	17676	3.06	34.7
5	Jizan	2862	20538	5.62	40.32
6	Jouf	1017	21555	2	42.31
7	Madina	2583	24138	5.07	47.39
8	Makkah	12123	36261	23.8	71.18

9	Najran	837	37098	1.64	72.83
10	North	801	37899	1.57	74.4
11	Qassem	2367	40266	4.65	79.05
12	Riyadh	9585	49851	18.82	97.86
13	Tabouk	1089	50940	2.14	100

The table below shows the distribution of proportion of deaths in different cities. We see that the proportion of deaths are almost same for all the cities. Among the cities, Makkah has the highest level and Tabouk has the lowest. The bar diagram below shows the deaths structure graphically.

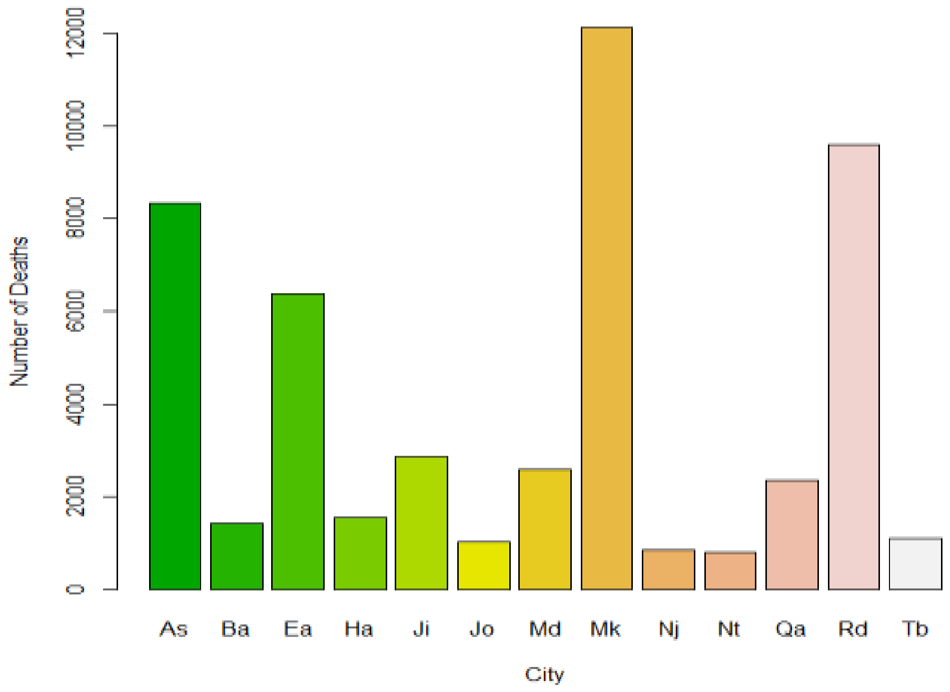

Figure 2: Deaths according to City

Table 3: Proportion of deaths according to city

	City	Population	Deaths	Proportion of deaths
1	Aseer	1056663	8325	0.007879
2	Baha	246915	1431	0.005796
3	East	1904742	6363	0.003341
4	Hail	336816	1557	0.004623
5	Jizan	736236	2862	0.003887
6	Jouf	220473	1017	0.004613
7	Madina	812349	2583	0.00318
8	Makkah	2699208	12123	0.004491
9	Najran	246060	837	0.003402
10	North	172116	801	0.004654
11	Qassem	596223	2367	0.00397
12	Riyadh	2734884	9585	0.003505
13	Tabouk	404271	1089	0.002694
		12166956	50940	

4.2.3 Distribution of proportion deaths according to Age

Table 4: Population according to Age

	Age	Freq	Cum Freq	Percentage	Cum Percentage
1	24-	5671944	5671944	46.62	46.62
2	24+	6495012	12166956	53.38	100

Around fifty-three percentage of the population are of age 24 or more and the rest are under 24 years. The number of deaths in group 24 or more is 41895 which is 0.6% of the population of that age group whereas the proportion of deaths is around 0.2% in the age group under 24 years.

Table 5: Proportion of deaths by Age Group

	Total	Deaths	Prop_Deaths

24-	5671944	9045	0.001595
24+	6495012	41895	0.00645
	12166956	50940	

From the bar graph (figure-3, we clearly see the huge difference between these two groups in terms of proportion of deaths.

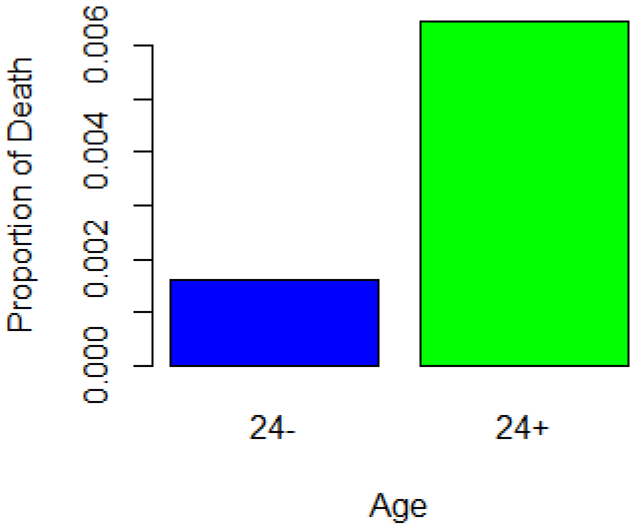

Figure 3: Proportion of deaths by age

4.2.4 Distribution of proportion deaths according to Education Level

Table 6: Population by Education Level Group

	Education	Freq	Percentage
1	Not educated	3573107	29.37%
2	Some Level of Education	7411363	60.91%
3	University Graduate	1182486	9.72%
		12166956	100%

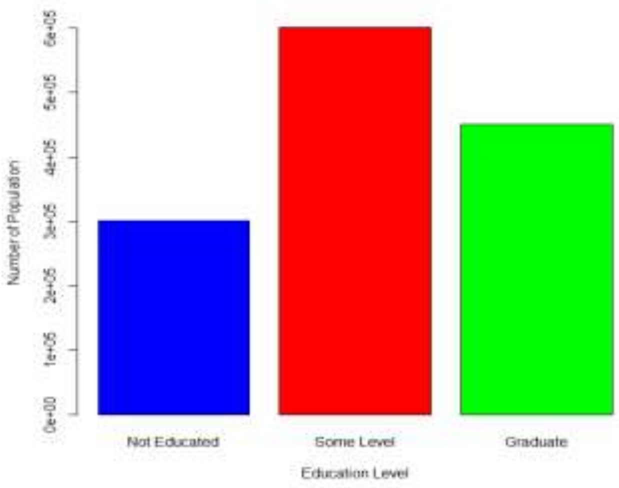

Figure 4: Population according to education level

Most of the people in the study have some level of education (about 61%), around 10% have university degree and around 29% are uneducated (Table 6 & Figure 4).

The table below shows the number of deaths are also high in the group of people having some level of education which is more than 44%.

Table 7: Deaths by Education Level

	Education	Freq	Cum Freq	Percentage	Cum Percentage
1	Not educated	11320	11320	22.22	22.22
2	Some Level of Education	22640	33960	44.44	66.67
3	University Graduate	16980	50940	33.33	100

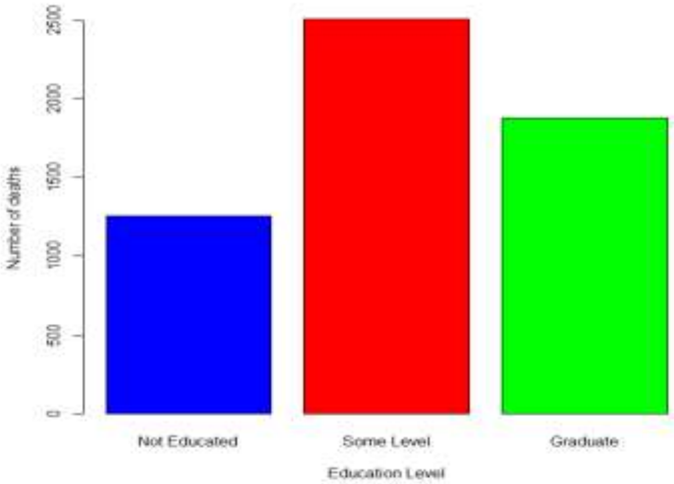

Figure 5: Number of Deaths by Education

4.2.5 Distribution of proportion deaths according to Sex

Table 8: Proportion of Deaths by Sex

	Total	Deaths	Prop_Deaths
Female	6072885	18234	0.0030
Male	6094071	32706	0.0054

The proportion of deaths is higher for male. It is almost double.

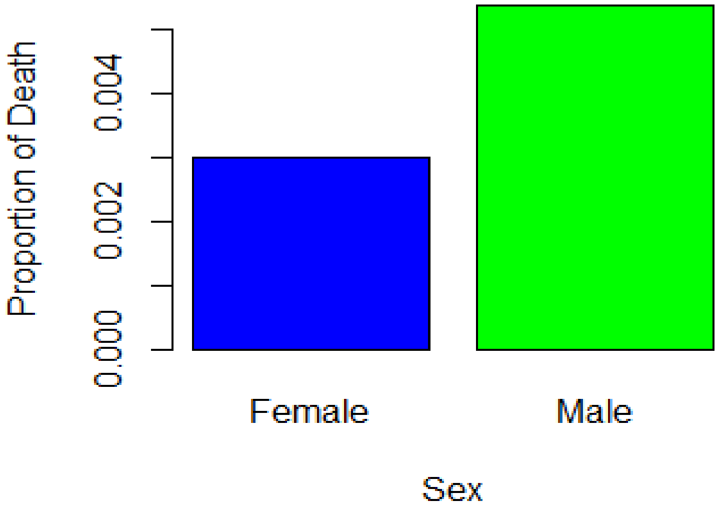

Figure 6: Proportion of Deaths by Sex

The bar plot above shows the clear significant difference of proportion of deaths between male and female.

4.2.6 Distribution of proportion deaths according to City and Education Level

Table 9: Population by City and Education Level

	Total		
	Not educated	Some Level of Education	University Graduate
Aseer	234818	469636	352227
Baha	54870	109740	82305
East	423272	846545	634909
Hail	74847	149694	112270

Jizan	163602	327204	245403
Jouf	48995	97991	73493
Madina	180525	361050	270788
Makkah	599827	1199654	899740
Najran	54680	109361	82021
North	38246	76491	57368
Qassem	132498	264996	198747
Riyadh	607750	1215500	911625
Tabouk	89836	179672	134754

Table 10: Deaths by City and Education Level

	Deaths		
	Note ducated1	Some Leve of Education1	University Graduate1
Aseer	1853	3705	2779
Baha	314	627	470
East	1410	2819	2114
Hail	345	691	518
Jizan	636	1272	954
Jouf	215	429	322
Madina	573	1145	859
Makkah	2694	5387	4040
Najran	181	362	272
North	171	341	256
Qassem	525	1051	788
Riyadh	2128	4255	3191
Tabouk	241	481	361

4.2.7 Proportion of deaths according to Sex and age group

Table 11: Proportion of deaths according to Sex and age group

Age Group	Total Population		Deaths		Proportion of Deaths	
	Sex					
	Female	Male	Female	Male	Female	Male
24-	2881215	2790729	1377	7668	0.000478	0.002748
24+	3191670	3303342	16857	25038	0.005282	0.00758

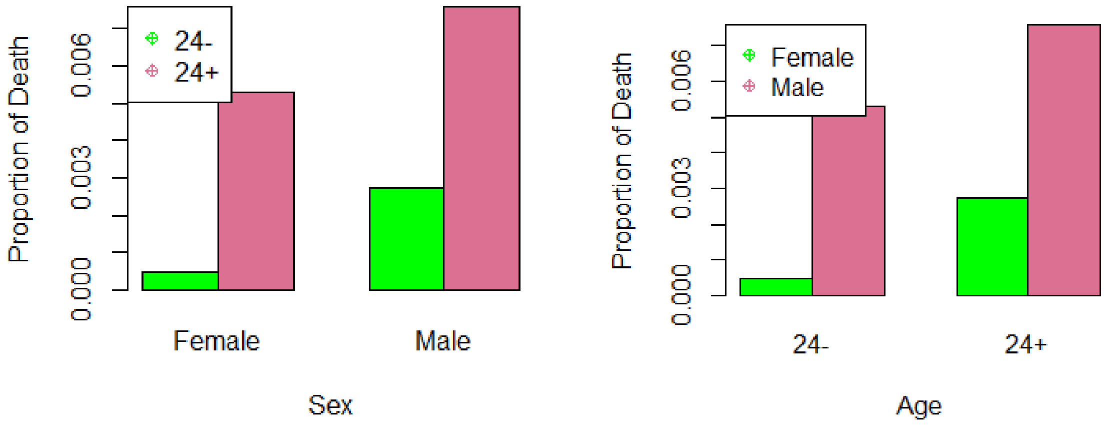

Figure 7: Proportion of deaths according to Sex and age group

The proportion of deaths in the female population for age group of 24- is about 0.00045 and for age group 24+ is 0.005, proportions are higher in the male population for age group 24- and 24+ respectively which are 0.003 and 0.007.

4.3 Multivariate Statistical Analysis

4.3.1 ANOVA Analysis

In this chapter we will discuss some multivariate aspects. Using some advanced statistical tools, we will try to investigate the relationship of proportion of deaths and some predictors. Initially using the ANOVA technique, we will study whether there is any association between proportion of deaths and some pre-defined predictors' one at a time. To check the statistical significance, we set up the cut off level of significance as $\alpha = 0.05$. The corresponding p-value which is less than $\alpha = 0.05$ will be considered as a significant variable. After analyzing all pre-defined predictors one by one we will try to build a relationship model. As our dependent variable is proportion of deaths which is rate/proportion, we will try some generalized linear models, like Poisson model, binomial model and beta model.

4.3.1.1 ANOVA Analysis for proportion of deaths with Sex

Table 12: ANOVA Analysis for proportion of deaths with Sex

	Df	Sum Sq	Mean Sq	F Value	P-value
Sex	1	0.0066	0.006613	35.54	2.82e-09***
Residuals	2806	0.5222	0.000186		

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

The ANOVA table above shows that the proportion of deaths is significant in male and female.

4.3.1.2 ANOVA Analysis for proportion of deaths with Age

Table 13: ANOVA Analysis for proportion of deaths with Age

	Df	Sum Sq	Mean Sq	F Value	P-value
Age	1	0.0273	0.027326	152.9	<2e-16***
Residuals	2806	0.5014	0.000179		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The ANOVA table above states that the proportion of deaths is significant in different levels of age group

4.3.1.3 ANOVA Analysis for proportion of deaths with City

Table 14: ANOVA Analysis for proportion of deaths with City

	Df	Sum Sq	Mean Sq	F Value	P-value
City	12	0.0042	0.00035	1.865	0.034 *
Residuals	2795	0.5246	0.0001877		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The ANOVA table above states that the proportion of deaths is significantly different among different cities.

4.3.2 Logit, Probit and cloglog models for proportion

In this section we present results from three regression models for proportion, namely logit model, probit model and cloglog models. Akaike Information Criterion (AIC) is used to measure the goodness-of-fit for each model. Table 15 presents the estimates of regression coefficients, their standard errors and corresponding z-values, p-values and odds ratios obtained from the logit model.

Table 15: Coefficients of Binomial Model with logit link

	Estimate	Std..Error	z.value	Pr...z..	Odds.Ratio
(Intercept)	-6.10473	0.017679	-345.318	0	0.002232
CityBaha	-0.31311	0.028745	-10.8928	1.25E-27	0.73117
CityEast	-0.88852	0.016718	-53.1488	0	0.411265
CityHail	-0.5448	z	-19.6578	4.96E-86	0.579956
CityJizan	-0.69301	0.021747	-31.8675	7.5E-223	0.50007
CityJouf	-0.51002	0.033341	-15.297	8E-53	0.600484
CityMadina	-0.93233	0.022593	-41.2664	0	0.393637
CityMakkah	-0.62089	0.014304	-43.4062	0	0.537469
CityNajran	-0.83722	0.036363	-23.0242	2.7E-117	0.432913
CityNorth	-0.50037	0.03713	-13.476	2.17E-41	0.606309
CityQassem	-0.68952	0.023376	-29.4969	3.2E-191	0.501817
CityRiyadh	-0.85228	0.015048	-56.6376	0	0.42644
CityTabouk	-1.10986	0.032303	-34.3576	1.1E-258	0.329606
Age24+	1.396219	0.011617	120.1905	0	4.039895
SexM	0.576495	0.009272	62.17293	0	1.77979
EducationSome Level of Education	2.34E-16	0.011549	2.02E-14	1	1
EducationUniversity Graduate	1.51E-16	0.012174	1.24E-14	1	1

```
AIC: 125844
```

Using the binomial model with logit link function, we see proportion of deaths is significantly different in different cities, between male and female, and in different age groups. The data do not show any statistical difference in proportion of deaths for different education settings.

The proportion of deaths is approximately 1.77 times higher for males than females. The proportion of deaths in the age group of 24+ is 4.039895 times higher than age group of 24 or less. To compare the proportion of deaths in different cities, we first set Aseer as the reference city and compare other cities with it. The results show that the proportion of deaths is less in all the cities than in Aseer. The odds ratio (OR) for city Bahah is .73 with a p-value < 0.05 which reveals that the odds of proportion of deaths for city Baha is 27% less than city Aseer.

Table 16: Coefficients of Binomial Model with cloglog link

	Estimate	Std..Error	z.value	Pr...z..	Odds.Ratio
(Intercept)	-6.10757	0.017631	-346.417	0	0.002226
CityBaha	-0.31002	0.028616	-10.8336	2.39E-27	0.733435
CityEast	-0.88385	0.016657	-53.0623	0	0.413189
CityHail	-0.54099	0.027611	-19.5928	1.78E-85	0.582174
CityJizan	-0.68875	0.02167	-31.783	1.1E-221	0.502203
CityJouf	-0.50644	0.03322	-15.2453	1.77E-52	0.602634
CityMadina	-0.92748	0.022525	-41.1761	0	0.395551
CityMakkah	-0.6169	0.014242	-43.3171	0	0.539614
CityNajran	-0.83261	0.036261	-22.9617	1.1E-116	0.434913
CityNorth	-0.49682	0.036995	-13.4291	4.08E-41	0.608465
CityQassem	-0.68536	0.023296	-29.4196	3.1E-190	0.50391
CityRiyadh	-0.8478	0.014989	-56.5617	0	0.428354
CityTabouk	-1.10493	0.032231	-34.282	1.5E-257	0.331233
Age24+	1.392751	0.011598	120.0817	0	4.02591
SexM	0.574433	0.009246	62.12514	0	1.776123
EducationSome Level of Education	-2.4E-16	0.011512	-2.1E-14	1	1
EducationUniversity Graduate	-2.4E-16	0.012135	-2E-14	1	1

```
AIC: 125859
```

Using the binomial model with cloglog link function, we find that the proportion of deaths is significantly different in different cities, between males and females, and in different age groups. The data do not show any statistical difference in proportion of deaths for different education settings.

Table 16 is similar to table 15, the proportion of deaths is approximately 1.77 times higher for males than females. The proportion of deaths in the age group of 24+ is 4.02 times higher than age group of 24 or less. To compare the proportion of deaths in different cities, we first set Aseer as the reference city and compare other cities with it. The results show that the proportion of deaths is less in all the cities than in Aseer. The odds ratio (OR) T for city Bahah is .0.73 with a p-value < 0.05 which reveals that the odds of proportion of deaths for city Baha is 27% less than city Aseer.

Table 17: Coefficients of Binomial Model with probit link

	Estimate	Std..Error	z.value	Pr...z..	Odds.Ratio
(Intercept)	-2.82836	0.006017	-470.071	0	0.05911
CityBaha	-0.12983	0.010366	-12.5243	5.5E-36	0.878248
CityEast	-0.32643	0.005867	-55.6372	0	0.721497
CityHail	-0.21049	0.009794	-21.4915	1.9E-102	0.810187
CityJizan	-0.2616	0.007623	-34.3198	4E-258	0.769816
CityJouf	-0.19606	0.011739	-16.7006	1.3E-62	0.821966
CityMadina	-0.34273	0.007788	-44.0062	0	0.70983
CityMakkah	-0.23585	0.005139	-45.8945	0	0.7899
CityNajran	-0.30959	0.012453	-24.8612	2E-136	0.733747
CityNorth	-0.19253	0.013069	-14.7322	4.01E-49	0.824871
CityQassem	-0.25888	0.008175	-31.6664	4.5E-220	0.771916
CityRiyadh	-0.31232	0.005328	-58.6193	0	0.731748
CityTabouk	-0.39358	0.010777	-36.5217	5E-292	0.674639
Age24+	0.475815	0.003749	126.9202	0	1.609326
SexM	0.203595	0.003176	64.09756	0	1.225802
EducationSome Level of Education	2.41E-16	0.004009	6.02E-14	1	1
EducationUniversity Graduate	-9.3E-16	0.004226	-2.2E-13	1	1

```
AIC: 125223
```

Using binomial model with probit link, like the logit link, the coefficient table above shows that except for education all other variables are significant.

Since there are similar significant variables with all three link functions, we need to decide on a model. There are many model selection criterions which might be used. Here we use AIC criterion for selecting the best model. Accordingly, we choose the lowest AIC for the binomial model with probit link function.

Using the binomial model with probit link function, we see the proportion of deaths is significantly different in different cities, between males and females, and in different age groups. The data do not show any statistical difference in proportion of deaths for different education settings.

The proportion of deaths is approximately 1.23 times higher for males than females. The proportion of deaths in the age group of 24+ is 1.61 times higher than age group of 24 or less. To compare the proportion of deaths in different cities, we first set Aseer as reference city and compare other cities with it. All cities have a proportion of deaths smaller than Aseer.

4.3.3 Beta Regression model for proportion

Table 18: Coefficients of Beta Regression

	mean.Estimate	mean.Std..Error	mean.z.value	mean.Pr...z..	OR
(Intercept)	-5.530135662	0.080162529	-68.98654221	0	0.003965
CityBaha	0.067151648	0.085261608	0.787595364	0.430933427	1.069458
CityEast	-0.039203458	0.08598567	-0.455930135	0.648440219	0.961555
CityHail	0.073604622	0.085216191	0.863739872	0.387730805	1.076381
CityJizan	-0.077509593	0.086234738	-0.898821002	0.368748011	0.925418
CityJouf	0.231277144	0.084057978	2.751400275	0.005934108	1.260208
CityMadina	-0.010788551	0.085796838	-0.125745323	0.899933524	0.989269
CityMakkah	0.142955366	0.084717921	1.6874277	0.091521141	1.153678
CityNajran	0.081414112	0.085161007	0.95600222	0.339071113	1.08482
CityNorth	0.270384216	0.083757439	3.228181517	0.001245799	1.310468
CityQassem	0.046328283	0.085407034	0.542441075	0.58751469	1.047418
CityRiyadh	0.146357879	0.084693008	1.72809872	0.083970527	1.15761
CityTabouk	0.101580081	0.085017407	1.194815089	0.232159284	1.106919
Age24+	0.478762247	0.040005251	11.96748499	5.25989E-33	1.614075
SexM	0.240835949	0.033262286	7.240511063	4.46997E-13	1.272312
EducationSome Level of Education	3.90558E-16	0.043067336	9.06855E-15	1	1
EducationUniversity Graduate	-3.71026E-16	0.045396958	-8.17293E-15	1	1

Using the above table, unlike the previous model, the proportion of deaths is not significantly different in different cities; only two cities (Makkah and North) are different. However, like the previous model, the proportion of deaths is significantly different between males and females and in different age groups. The data do not show any statistical difference in proportion of deaths for different education settings.

The proportion of deaths is approximately 1.27 times higher for males than females. The proportion of deaths in the age group of 24+ is 1.61 times higher than age group of 24 or less. To compare the proportion of deaths in different cities, we first set Aseer as reference city and compare other cities with it.

The OR for city Makkah is 1.15 with a p-value < 0.05 which reveals that the odds of proportion of deaths for city Makkah is 15% more than city Aseer. The OR for city North is 1.13 with a p-value < 0.05 which reveals that the odds of proportion of deaths for city North is 13% more than city Aseer.

4.4 Conclusion

The two candidates models, binomial model with probit link and beta model shown above, depicts the relationship between proportion of deaths and some other predictors. The literature suggests to use beta regression model for modeling the proportion but in this study the binomial model with probit link shows the better results with maximum significant predictors.

Chapter 5

Conclusion and Discussion

This study is an attempt to see the probable connection of proportion of deaths with gender, age, education and geographic location in the Kingdom of Saudi Arabia. The results are very similar to previous studies. To study the effect of the available factors on deaths, we considered two different approaches: the binomial models with different link functions and the beta regression model. After analyzing all aspects, we selected the binomial model with probit link function to model the relationship between proportion of deaths and the selected predictors. Using this model, we found geographic location has a significant impact on proportion of deaths. Gender is another important feature to discriminate the deaths rate. Very likely, higher age group shows the greater risk associated with the proportion of deaths. Surprisingly, we do not find any significant difference in proportion of deaths in case of different levels of education.

Our study provides comparable information on the proportion of deaths and the available demographic and socio economic factors for the Saudi population. This opens a new window of research in Kingdom of Saudi Arabia to study the proportion of deaths with regard to different aspects. Since there were time and money restrictions, we could not study many related areas. However, lifestyle of the Saudi people, especially food habits and exercise habits, will be two important features to consider for future research.

References

1. Clark D, Royer H. The effect of education on adult health and mortality: evidence from Britain. Working paper. 2010. http://www.nber.org/papers/w16013.
2. Grossman M. Education and non-market outcomes. Handbook of the Economics of Education. 2006; 1:578-633.
3. Oreopoulos P, Salvanes KG. How large are returns to schooling? hint: money isn't everything. *The National Bureau of Economic Research*. 2009.
4. Weinstein MC, Skinner JA. Comparative effectiveness and health care spending - implications for reform. *New England Journal of Medicine*. 2010; 362(5): 460-465.
5. Hoyert DL, Arias E, Smith BL, et al. Deaths: final data for 1999. *National Vital Statistics Reports*. 2001; 49(8):1-116.
6. Feldman JJ, Makuc DM, Kleinman JC, Cornoni-Huntley J. National trends in educational differentials in mortality. *Am J Epidemiol.* 1989; 129(5): 919-933.
7. Liu K, Cedres LB, Stamler J, et al. Relationship of education to major risk factors and deaths from coronary heart disease, cardiovascular diseases and all causes: findings of three Chicago epidemiologic studies. *American Heart Association*. 1982; 66(6): 1308-1314.
8. Buchner HC, Ragland DR. Socioeconomic indicators and mortality from coronary heart disease and cancer: a 22-year follow-up of middle-aged men. *Am J Public Health*. 1995; 85(9): 1231-1236.
9. Dennis BH, Zhukovsky GS, Shestov DB, et al. The association of education with coronary heart disease mortality in the USSR Lipid Research Clinics Study. *Int J Epidemiol*. 1993; 22(3): 420-427.
10. Schwab JJ. Health and behavior: The interplay of biological, behavioral, and societal influences. American Journal of Psychiatry. 2003; 160(3): 603-605.
11. Kitagawa EM, Hausern PM. Differential Mortality in the United States: a Study in Socioeconomic Epidemiology. Cambridge: Mass, Harvard University Press; 1973.
12. Deaton A, Paxson C. Mortality, education, income and inequality among American cohorts. Princeton University. 1999.
13. Elo IT, Preston HS. Educational differentials in mortality: United States, 1979-1985. *Social Science & Medicine.*1996; 42(1): 47-57.

14. Black D, Morris J, Smith C, et al. Inequalities in Health: The Black Report: The Health Divide. London, United Kingdom: Penguin; 1988.

15. Adler NE, Ostrove JM. Socioeconomic status and health: what we know and what we don't. *Ann NY Acad Sci*. 1999; 896(1): 3–15.

16. Kerns GJ. Introduction to Probability and Statistics Using R. 1st Ed. GNU Free Documentation License. 2011.

17. Dobson AJ, Barnet A. An Introduction to Generalized Linear models, Second edition, Chapman & Hall/CRC. 2008.

18. Agresti A, Finlay B. Statistical Methods for the Social Sciences. 3th Ed. Prentice Hall, 1997.

19. Anderson TW, Sclove SL. Introductory Statistical Analysis. Houghton Mifflin Company. 1974.

20. Clarke GM, Cooke D. A Basic Course in Statistics. 5th Ed. Arnold, 1998.

21. StatSoft Inc. Big data, data mining, productive analysis, statistics, statSoft Electronic Statistics Textbook. 2016. http://www.statsoftinc.com/textbook/stathome.html.

22. Freund JE, Perles BM. Modern Elementary Statistics. Prentice-Hall, 2001.

23. Johnson RA, Bhattacharyya GK. Statistics: Principles and Methods, 2nd Ed. Wiley, 1992.

24. Moore D, McCabe G. Introduction to the Practice of Statistics. 3re Ed. Freeman, 1998.

25. Lecture notes from course MATH 627: Generalized linear models taken at Ball State University, Spring 2016.

Appendix

R code

```
setwd("C:/Users/Moamer/Documents/Desktop/fall 2016 graduate")

library(xlsx)

datah <- read.xlsx("C:/Users/Moamer/Documents/Desktop/fall 2016 graduate/Copy of data_all-1
today.xlsx",1,header=T)

head(datah)

library(car)

datah$Age=recode(datah$Age,"c('10 - 14','15 - 19','20 - 24') ='24-'; else='24+'")

datah$Education=recode(datah$Education,"c('PhD','Master','University') ='University Graduate';

                    c('RW', 'Illiterate') ='Not educated';else='Some Level of Education'")

head(datah)

#### Table with datah ####

attach(datah)

tc=round(xtabs(Total ~ City));tc

t1=round(cbind(Frequency=tc,Cum_Freq=cumsum(tc),Perc=prop.table(tc)*100),2);t1

barplot(tc,xlab="City",ylab="Number of
Population",names.arg=c("As","Ba","Ea","Ha","Ji","Jo","Md","Mk","Nj","Nt","Qa","Rd","Tb"),
main="Population according to City", col=terrain.colors(13))

td=round(xtabs(Death ~ City));td

barplot(td,xlab="City",ylab="Number of
Death",names.arg=c("As","Ba","Ea","Ha","Ji","Jo","Md","Mk","Nj","Nt","Qa","Rd","Tb"),main
="Death according to City", col=terrain.colors(13))

t2=round(cbind(Frequency=td,Cum_Freq=cumsum(td),Perc=prop.table(td)*100,Cum_Perc=cum
sum(td)/sum(Death)*100),2);t2

t3=cbind("Total"=tc,"Death"=td,"Prop_Death"=td/tc);t3 # Prop of Death by City=pdc

ta=round(xtabs(Total ~ Age));ta #Over all population size in each age group
```

```r
t4=round(cbind(Frequency=ta,Cum_Freq=cumsum(ta),perc=prop.table(ta)*100,Cum_Perc=cumsum(ta)/sum(Total)*100),2);t4

da=round(xtabs(Death ~ Age));da #Over all Death Status in each age group

#t=round(cbind(Frequency=da,Cum_Freq=cumsum(da),perc=prop.table(da)*100,Cum_Perc=cumsum(da)/sum(Death)*100),2);t

t5=cbind("Total"=ta,"Death"=da,"Prop_Death"=da/ta);t5 # Prop of Death by Age=pda

barplot(da/ta,xlab="Age",ylab="Proportion of Death",names.arg=c("24-","24+"),main="Proportion of Death by Age", col=c("blue","green"))

te=round(xtabs(Ed_No ~ Education));te #Over all education level

t6=round(cbind(Frequency=te,Cum_Freq=cumsum(te),perc=prop.table(te)*100,Cum_Perc=cumsum(te)/sum(Ed_No)*100),2);t6

de=round(xtabs(Death ~ Education)) #Death as education level

t7=round(cbind(Frequency=de,Cum_Freq=cumsum(de),perc=prop.table(de)*100,Cum_Perc=cumsum(de)/sum(Death)*100),2);t7

t8=cbind("Total"=te,"Death"=de,"Prop_Death"=de/te);t8 # Prop of Death by Age=pda

barplot(de/te,xlab="Education",ylab="Proportion of Death",names.arg=c("Not Educated","Some Level","Graduate"),main="Proportion of Death by Education", col=c("blue","red","green"))

ts=round(xtabs(Total ~ Sex))

ds=round(xtabs(Death ~ Sex))

t9=cbind("Total"=ts,"Death"=ds,"Prop_Death"=ds/ts);t9 # Prop of Death by sex=pds

barplot(ds/ts,xlab="Sex",ylab="Proportion of Death",names.arg=c("Female","Male"),main="Proportion of Death by Sex", col=c("blue","green"))

t10=round(xtabs(Total ~ City+Education));t10

t11=round(xtabs(Death ~ City+Education));t11

t12=("Prop"=t11/t10);t12

tea=round(xtabs(Total ~ Education+Age));tea

dea=round(xtabs(Death ~ Education+Age));dea

t13=cbind(tea,dea,dea/tea);t13

barplot(dea/tea,xlab="Education",ylab="Proportion of Death",

    names.arg=c("24-","24+"),
    col=c("green","palevioletred","tan"),beside=T,main="Proportion of death in different education level by age")
legend("topleft", pch=10,horiz=F,cex=1, legend=c("Not Educated","Some Level","Graduate"), col=c("green","palevioletred","tan"))
tes=round(xtabs(Total ~ Education+Sex))
des=round(xtabs(Death ~ Education+Sex))
t14=cbind(tes,des,des/tes);t14
barplot(des/tes,xlab="Education",ylab="Proportion of Death",
    names.arg=c("Female","Male"),
    col=c("green","palevioletred","tan"),beside=T,main="Proportion of death in different education level by Sex")
legend("topleft", pch=10,horiz=F,cex=1, legend=c("Not Educated","Some Level","Graduate"), col=c("green","palevioletred","tan"))
tas=round(xtabs(Total ~ Age+Sex))
das=round(xtabs(Death ~ Age+Sex))
t15=cbind(tas,das,das/tas);t15
barplot(das/tas,xlab="Sex",ylab="Proportion of Death",
    names.arg=c("Female","Male"),
    col=c("green","palevioletred"),beside=T,main="Proportion of death in different Age level by Sex")
legend("topleft", pch=10,horiz=F,cex=1, legend=c("24-","24+"), col=c("green","palevioletred"))
############## ANOVA and Measures of Association ##########################
m1 <- aov(Death/Total ~ Sex, data=datah);summary(m1)
m2 <- aov(Death/Total ~ Age, data=datah);summary(m2)
m3 <- aov(Death/Total ~ City, data=datah);summary(m3)
m4 <- aov(Death/Total ~ Education, data=datah);summary(m4)
```

```
#################### Binomial Model ######################

#############################################################

nd <- Total-Death

p<- c(Death/Total)

bd <- cbind(Death,nd)

plot(p~Age,xlab = "Age", ylab ="Proportion of Death", main = "Age-Death plot")

plot(p~City,xlab = "City", ylab ="Proportion of Death", main = "City-Death plot")

plot(p~Sex,xlab = "Sex", ylab ="Proportion of Death", main = "Sex-Death plot")

plot(p~Education,xlab = "Education", ylab ="Proportion of Death", main = "Education-Death plot")


# Fit the simple logistic regression model

logit <- glm(bd ~City+Age+Sex+Education,family =  binomial(link = "logit"));summary(logit)

probit <- glm(bd ~City+Age+Sex+Education, family = binomial(link = "probit"));summary(probit)

cloglog <- glm(bd ~City+Age+Sex+Education, family = binomial(link = "cloglog"));summary(cloglog)

logit$aic;probit$aic;cloglog$aic

c1=coef(summary(probit));c1

o1=exp(coef(probit));o1

cf1 <- cbind(c1, "Odds Ratio"=o1);cf1

# Fitted y=Number of Death

py=fitted.values(probit)*Total

#Chi-Square value for each model to test the goodness of fit

ch_p=sum(((Death-py)^2/py)+(nd-(Total-py))^2/(Total-py));ch_p

plot(probit)
```

```
############# Distribution of Death Proportion ################
prd=Death/Total ## Proportion of death
y=sort(Death/Total)
plot(y,ylab="Proportion of Death", main="Graph of Proportion of Death")
d1 <- data.frame(City,Age,Sex,Education,prd)
library(betareg)
bg=betareg(prd~City+Age+Sex+Education, data=d1)
summary(bg)
plot(bg)
cb=coef(summary(bg));cb
ob=exp(coef(bg));ob
cfb <- cbind(cb, "Odds Ratio"=ob);cfb
```