Arab American University
Faculty of Graduate Studies
Department of Natural, Engineering and Technological Sciences
Master Program in Applied mathematics

The Influence of Genetic and Environmental Factors on the Occurrence of Chronic Diseases and their Prediction.

Yousef Hasan Hussain Weshahi
202120151

Supervision Committee:

Dr. Saleh Afaneh
Dr. Abdilhalim Ziqan
Dr. Mohammed Dawabshah

This Thesis Was Submitted in Partial Fulfilment of the Requirements for the Master Degree in Applied Mathematics

Palestine, September / 2025

Arab American University

Faculty of Graduate Studies

Department of Natural, Engineering and Technological Sciences

Master Program in Applied mathematics

Thesis Approval

The Influence of Genetic and Environmental Factors on the Occurrence of Chronic Diseases and their Prediction.

Yousef Hasan Hussain Weshahi

202120151

This thesis was defended successfully on 23/9/2025 and approved by:

Thesis Committee Members:

Name	Title	Signature
1. Dr. Saleh Afaneh	Main Supervisor	
2. Dr. Abdelhalim Ziqan	Member of Supervision Committee	
3. Dr. Mohammed Dawabshah	Member of Supervision Committee	

Palestine, September / 2025

Declaration

I declare that, except where explicit reference is made to the contribution of others, this thesis is substantially my own work and has not been submitted for any other degree at the Arab American University or any other institution.

Student Name: Yousef Hasan Hussain Weshahi

Student ID: 202120151

Signature: 

Date of Submitting the Final Version of the Thesis: 30/12/2025

Dedication

I dedicate this work to my dear parents, whose constant support, prayers and encouragement have been my guiding light throughout this journey. I thank my family and friends for standing by my side. Finally, I thank my supervisors whose guidance and advice have been invaluable.

Student Name: Yousef Hasan Hussain Weshahi

Acknowledgments

First and foremost, I dedicate this work to the brave prisoners whose unwavering resolve motivates us to continue the struggle for justice and freedom, as well as to our steadfast people in Gaza, and to our righteous martyrs who sacrificed their lives in defense of our homeland. You will always remain in our hearts.

I would like to extend my sincere thanks to my esteemed supervisor, Dr. Saleh Afaneh, and the members of the thesis committee, Dr. Abdel Halim Ziqan and Dr. Mohammed Dawabsheh, for their guidance and support throughout this research journey, which enabled me to complete this research.

My mother has been my source of inspiration and strength, and my dear father, my role model. I would like to express my gratitude to my brothers and sisters for their constant support and participation.

I also extend my sincere thanks to my family, colleagues, and friends for their continued support, which has been a constant source of strength.

Finally, I sincerely wish peace, prosperity, and security for our steadfast people in Gaza, the West Bank, and beyond.

Title: The Influence of Genetic and Environmental Factors on the Occurrence of Chronic Diseases and their Prediction.

Student's Name: Yousef Hasan Hussain Weshahi

Supervision Committee:

Dr. Saleh Afaneh

Dr. Abdilhalim Ziqan

Dr. Mohammed Dawabshah

Abstract

Chronic diseases are one of the biggest global challenges and the most common causes of death. They do not occur in a vacuum; rather, there are factors that may lead to their occurrence. Therefore, this study aims to examine the influence of genetic, environmental, and lifestyle factors on the prevalence of chronic diseases and their prediction using binary logistic regression. The study was conducted on a random sample of 387 individuals from the Jenin Governorate, and their data were analyzed using SPSS.

The analysis revealed the influence of seven factors at a 0.05 confidence level: age, the presence of affected siblings, a specific familial pattern of the disease, regular exercise and exposure to fresh air, psychosocial stress, alcohol or drug addiction, and regular checkups. The model's prediction accuracy reached 80.4. The study highlights the importance of these factors and their role in disease prevention and reducing their spread as much as possible through public awareness and adopting a healthy lifestyle.

Keywords: Logistic Regression, Maximum Likelihood Estimation Method (MLE), Chronic Diseases, Genetic and Environmental Factors.

Table of Contents

List of Tables

List of figures

Chapter One: Introduction

Mathematics plays a pivotal role in many fields, including medicine and human health. Statistical mathematics encompasses the methods and models used to predict the occurrence of events, including chronic diseases. Chronic diseases, such as diabetes, heart disease, and hypertension, pose a major global health challenge, burdening societies and individuals. They cause 74% of global deaths, mostly in poor communities and low-income countries, posing a significant challenge for medical research and health prevention.

These diseases do not occur in a vacuum; rather, they are influenced by a combination of factors and causes that can be classified as genetic, environmental, and lifestyle factors. Genetics is an important factor, as it refers to the biological predisposition to certain diseases resulting from genetic mutations, making some individuals more susceptible to chronic diseases than others. Various environmental factors, such as pollution and exposure to toxic substances, as well as a person's lifestyle, behavior, and diet, are also important causes of these diseases. Economic, social, and psychological factors also play a role. Based on these important factors, predicting the occurrence of chronic diseases is a major challenge, requiring numerous studies and careful analyses using various method.

The gap in this research lies in the scarcity of studies in Palestinian society on the relationship between genetic, environmental, behavioral, and lifestyle factors and the occurrence of chronic diseases. This prompted us to conduct this study, which targeted a research sample to examine the relationship between these factors and the occurrence of chronic diseases.

Palestinian society is considered a low- and middle-income society. According to Ministry of Health data, cardiovascular disease is the most common cause of death in Palestine, with the death rate from heart attacks reaching 11.7% of deaths. The percentage of individuals suffering from at least one chronic disease has increased, according to data from the Palestinian Central Bureau of Statistics, from 18% to 20% over the past decade, and the

percentage of elderly people suffering from at least one chronic disease has reached more than two-thirds. Smoking rates have also increased over the past decade from 23% to 31%, and more than half of adults in the community suffer from depression.

In addition, community-based studies on the relationship between genetic, environmental, behavioral, and lifestyle factors and the occurrence of chronic diseases are scarce. One of the main reasons for conducting this study, which targeted a sample, was to conduct a study on the relationship between these factors and the occurrence of chronic diseases.

Study Problem

The problem of the study lies in the rapid spread of chronic diseases in society. These diseases have multiple, interconnected causes, including genetic, environmental, behavioral, and lifestyle factors. Despite the existence of global research on this topic, Palestinian society lacks such studies and models that help understand the relationship between these factors and determine the extent of their influence on the occurrence of chronic diseases. This study will develop a statistical model using binary logistic regression to predict the occurrence of chronic diseases, including genetic, environmental, and lifestyle factors. This model may aid in prevention, mitigation, and early detection of diseases.

Study Objectives

- Identifying the factors influencing the occurrence of chronic diseases. This is done by analyzing a logistic regression model of sample data taken from Jenin Governorate, then accepting statistically significant factors (i.e., those that have an impact on the occurrence of the disease).
- Ranking the factors influencing the occurrence of chronic diseases, i.e., identifying the strength and extent of each factor's influence, and ranking them based on their impact on the occurrence of chronic diseases, using the results of the regression model.

- Formulating the model equation or prediction equation, through which the probability of predicting the occurrence of chronic diseases in an individual is determined, based on their data related to the statistically significant factors.
- Providing and proposing preventive medical recommendations that may contribute to reducing the prevalence of chronic diseases as much as possible in the future.

The process of identifying the influencing factors that play a role in the occurrence of chronic diseases is carried out using a logistic regression model. These factors are also ranked and classified using the model. Logistic regression also provides a model or equation for predicting the occurrence of chronic diseases in individuals. This, in turn, provides data that may help develop preventive measures to mitigate the impact of influencing environmental factors. It also highlights the importance of influencing genetic factors and raises awareness of them. Furthermore, it focuses on individual behavior and lifestyle, which enhances awareness of their health and behavior. All of this contributes to reducing the incidence of chronic diseases and curbing their spread through medical recommendations and health and preventive measures.

Study Methodology

The study relies on a quantitative research methodology and is conducted in the following stages:

1. Data Collection: A random sample of 387 individuals from both sexes was used from the Jenin Governorate.
2. Data Collection Method: An electronic questionnaire containing questions about genetic, environmental, and lifestyle factors was administered.
3. Data Analysis: Logistic regression model analysis using SPSS was used to identify statistically significant factors.

   Binary Logistic Regression Analysis Model: $\ln \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{k} \beta_i X_i$

   P: probability, β_0: intercept coefficient, β_i: regression coefficients, X_i: independent variables.

4. Regression Model Evaluation: The model's suitability, accuracy, and predictive ability were evaluated using the following tests: This is done through the following tests: classification table, goodness of fit tests, which include the chi-square test, the Hosmer-Lemeshow test, the Wald test, confidence interval estimation, coefficients of determination, multicollinearity test, and likelihood ratio test.

Research Questions

- What are the causes of chronic diseases, and what is the relationship between genetic, environmental, and lifestyle factors?
- Do genetic factors interact with environmental and lifestyle factors to lead to chronic diseases?
- Do unhealthy practices and habits, such as smoking, alcohol, and other factors, play a role in the occurrence of chronic diseases?
- Are genetic factors the primary cause of chronic diseases, or are environmental factors, lifestyle, and individual behavior the primary cause?
- Does the effect of these factors differ across different age groups?
- How accurate is the logistic regression model in predicting outcomes, and can additional factors be added or specific modifications made to improve the accuracy of the logistic regression model in predicting outcomes?

- **The study hypotheses**

H_0 : Independent variables have no effect on the dependent variable.

H_1: Independent variables have an effect on the dependent variable.

The first chapter of the thesis includes an introduction to the thesis, its importance, objectives, research questions, methodology, hypotheses, and chapters, the second chapter includes a general definition of logistic regression, its importance, uses, and previous studies that have used it, It also includes a definition of chronic diseases, their types and classifications, as well as the causes and factors affecting their occurrence, It also includes

previous studies on the topic and a literature review, the third chapter covers the study design, statistical methods and tools used, analytical techniques, outcome assessment tests, sample collection methods and size, and sample analysis results. the fourth chapter presents the sample results and their analysis using the logistic regression model, along with the analysis outputs, regression equations, and prediction models. Finally, the fifth chapter provides conclusions, discussion, recommendations, and suggestions for future research.

Chapter Two: Literature review

Chronic diseases are among the most complex and challenging issues facing the world. Chronic diseases (non-communicable diseases) are long-term medical conditions that develop slowly and often worsen over time. According to the World Health Organization's definition, a chronic disease is a disease that accompanies an individual chronically and cannot be treated, and is diagnosed by a specialist doctor and takes continuous treatment (World Health Organization, 2018).

Chronic diseases include many common and well-known diseases in the world, such as heart disease, blood pressure, diabetes, cancer, chronic respiratory diseases such as asthma, chronic mental illnesses, thalassemia, and other chronic diseases. Chronic diseases are the cause of a large number of annual deaths and a large percentage, as they cause the death of 41 million people annually around the world, which is equivalent to 74% of global deaths, and 86% of deaths from chronic diseases occur in poor, low- and middle-income countries.

Heart disease is the leading cause of death among chronic diseases, followed by cancer of various types, then chronic mental illnesses, then diabetes, as these four diseases constitute the majority of deaths at a rate of 80% of deaths (WHO). These chronic diseases pose a major challenge to global health systems and societies, due to their impact on human life and the suffering they cause, the long-term impact on individuals' health and quality of life, and the huge health costs, so they are among the biggest challenges in the modern area. Chronic diseases are classified into several groups, the most famous of which are:

- Cardiovascular diseases: These diseases include many conditions such as high blood pressure, coronary artery disease, and heart failure.
- Cancer of all types: It is a disease that causes disorder and abnormal growth of cells.
- Chronic mental illnesses: Such as neurodevelopmental disorders, schizophrenia spectrum, and other disorders.

- Diabetes: It is a disorder in the glucose metabolism process, and high blood sugar causes serious complications.
- Respiratory diseases: such as asthma, bronchitis, pulmonary fibrosis and other diseases that affect breathing.

Chronic diseases occur as a result of many causes and factors, and the most famous of these factors can be classified into 3 factors:

- Genetic factors: Genes are associated with chronic diseases as they can lead to certain diseases, and having a family history of a certain disease increases the risk of developing it through genetic transmission (Centers for Disease Control and Preventioon, 2021).
- Environmental factors: Air pollution can lead to respiratory diseases, and water pollution can lead to kidney failure, and exposure to chemicals at work or in the residential area can lead to cancer (American Lung Association, 2020).
- Behavioral factors (lifestyle): Unhealthy diet, lack of physical activity, psychological, social and economic stress, and bad habits, such as addiction to alcohol, drugs or smoking, can lead to chronic diseases. Many genes can interact with each other and with the environment, leading to increased susceptibility to disease (National Institute of Mental Helth (NIMH)(https://www.nimh.nih.gov/, 2019).

To study genetic, environmental, and lifestyle factors and their relationship to the occurrence of chronic diseases, studies have been conducted in many countries around the world. One study (Cui et al., 2023) examined the predisposition to systemic lupus erythematosus (SLE), a chronic autoimmune disease that primarily affects women. However, the causes of this disease are multiple and complex, encompassing factors, environment, and behavior. A study was conducted in the United States to predict this disease on a sample of 1,274 women. A logistic regression-based model for predicting the disease was presented, and it is considered the first model to predict this disease. This study is unique in that it was conducted on a specific chronic disease only, but a gap in the research is that the average age

of those affected was in the early 50s, which requires more studies on younger and more diverse populations.

The researchers (Ordovas, Shen, 2008) also highlighted the common factor between genes and diet in chronic diseases. Developed countries suffer from malnutrition, which leads to the influence of factors that contribute to chronic diseases. This means that the interaction between genes and nutrition plays a significant role in the occurrence of these diseases and increases the risk of chronic diseases, highlighting the role of healthy nutrition and a balanced diet. The researchers recommended increasing the target sample size to obtain more accurate and better results, in addition to using various statistical methods to determine the complex relationships and interactions.

A study (Meng et al., 1999) in Hawaii also demonstrated the role of behavioral and lifestyle factors in the occurrence of chronic diseases. Demographic information was collected between 1975 and 1980 on a sample of 15,693 men and 16,007 women, who were followed until 1994. This study demonstrated that behavioral factors such as smoking and body mass index play a major role in the occurrence of these diseases. However, this study aimed to study behavioral or lifestyle factors and did not address environmental and genetic factors.

In Finland, (Kaprio, Koskenvuo, 2002) conducted a four-stage study involving thousands of Finnish twins who were followed. This study was based on a study of obsessive-compulsive disorder. The study demonstrated the role of genetic factors in the development of the disease. However, while this study assessed non-shared environmental factors, it did not address detailed environmental and behavioral factors such as nutrition, work environment, pollution, and physical activity.

(Kujala, 2011) also summarized the relationship between physical activity and chronic diseases, providing evidence for this. High physical activity is associated with reduced morbidity and mortality through several mechanisms. Genetic factors and genes play a role in predisposition to chronic diseases. Although this study examined the role of genes and the

interaction of genes with physical activity, it did not examine other detailed environmental factors, such as a family history of chronic diseases, the presence of first- or second-degree relatives with chronic diseases, or the presence of a specific family pattern of chronic diseases.

In a (Kelishadi, Poursafa, 2014). study examining the leading cause of death worldwide, cardiovascular disease, they examined its most important causes and the interaction between environment and genes. They pointed to the risk of early obesity and the role of environmental factors such as lack of physical activity, unhealthy nutrition, pollution, and smoking in the development of the disease. The researchers recommended early detection of disease risks and made recommendations such as improving maternal nutrition, breastfeeding, and physical activity. This study did not use statistical analysis methods to study the importance of each factor or predict future disease occurrence.

In the United States, (Wehby et al., 2017) analyzed data from 9,317 older adults (aged 65 and older) to measure their genetic predisposition to coronary artery disease, diabetes, and other diseases, and to study their daily activities, dreams, wealth, and other factors. They found that genetic risks are associated with poor functional health and socioeconomic outcomes, as genetic predispositions to chronic diseases are associated with these factors. This highlights the need for studies that focus on environmental and behavioral aspects as well, and provide predictive models for disease occurrence.

In China, (Hung et al., 2023) also demonstrated a relationship between poor health and mortality rates, and that several factors contribute to poor health, such as temperature fluctuations, pollution, and an unbalanced diet with a focus on seafood, red meat, and eggs. Increased educational resources also helped reduce poor health. Although the study highlighted the significant role of environmental factors, it did not address lifestyle and behavioral factors, such as smoking, alcohol, and physical activity. It also did not address genetics, and only used statistical methods such as linear regression, although other methods could have been used.

The logistic regression model is a nonlinear parametric regression model where the dependent variable is binary. Its importance lies in its use in predicting future events through an equation formulated using logistic regression analysis. It is an advanced analytical tool that can predict chronic diseases by analyzing genetic, environmental, and lifestyle factors. The concept of logistic regression was first developed by Berkisson (1944), and its most common uses are in the medical field.

Binary logistic regression was used to predict the likelihood of an individual contracting COVID-19 by identifying the most significant symptoms associated with the virus. Data were collected from 260 individuals at Alexandria University Hospital. The study concluded that the independent variables contributed 79.4% to explaining changes in the dependent variable, 59% to improving prediction accuracy, and 89.6% to correctly classifying individuals. A gap in this research lies in the small sample size for an area like Alexandria, which limits the generalizability of the results (Abdul khaliq, 2020).

Logistic regression was also used to identify the most important factors influencing lung cancer in Iraq. A sample of 1,500 individuals with lung cancer was found to have five variables that influence lung cancer incidence. The gap in this research lies in the number of independent factors, which is 15. Given the large sample size and the nature of the study, it would have been preferable to increase the number of independent variables to obtain more accurate and better results (Ibrahim, Taher, 2017).

Logistic regression was also used to examine the most important factors influencing the incidence of diabetes in the Iraqi city of Hillah. The study included 150 individuals and consisted of 15 independent variables. The study concluded that most variables were statistically significant and influential, with a correct classification rate of 92.7%. The gap in the research lies in the small sample size for the city of Hillah (Al_Bairmani, Ismael, 2021).

Logistic regression was also applied to the economic aspect to study and analyze the factors influencing the choice of capital change for an economic institution. The study included a sample of 100 financial observations from 20 diverse institutions. The study

included nine independent variables, five of which were found to be statistically significant and influential, with a correct classification rate of 83%. The study concluded that logistic regression is the optimal model for estimating the parameters of this type of model (Abdullah, Abdul Qader, 2021).

The logistic regression model was also used to assess the factors influencing creative thinking among employees in educational institutions. It is suitable for classifying binary qualitative variables. The logistic regression model was used on a sample of 100 male and female workers in 34 Algerian institutions. Twenty-nine independent factors were tested, and nine independent factors had a statistically significant impact on creative thinking. One of the most important results of the study is that the logistic regression model for assessing creativity has a high predictive efficiency of 84%. The sample size in this study is relatively small for educational institutions in a country like Algeria, which may limit the generalizability of the results (Al-Saeed, Zeina, 2021).

Logistic regression was also used in the financial aspect, along with the decision tree classification method, to predict financial distress in companies. Data was collected from 100 Taiwanese companies over a period of 7 years, half of which were experiencing financial distress and the other half were not. 37 factors were tested, classified into 6 groups, and it was proven that regression is more accurate in long-term prediction. However, the study included 100 Taiwanese companies, which means that the study was only local, and the results cannot be generalized globally (Chen, 2011).

The researcher (Aljammal, 2023) used logistic regression to predict the incidence of osteoporosis in patients with asthma and chronic obstructive pulmonary disease in Palestine. She took a stratified sample of 60 people distributed into 3 categories, each with 20 people. She examined the effect of 8 independent variables on the dependent variable, which is osteoporosis. She showed that 4 variables had statistical significance and an effect. However, this research studied only 8 independent variables on a small sample size, which gives inaccurate results and weak generalizability.

In a study conducted by (Malak et al, 2025) in Palestine on 1,249 infants to examine the causative factors of anemia and its prediction, the study showed that bottle feeding, low household income, and complementary foods before 6 months of age had a statistically significant impact on the occurrence of anemia. This study lacked an examination of genetic and environmental factors and was limited to factors related to the child's and mother's nutrition, as these factors may play a role in the occurrence of anemia. It did not provide an equation for predicting anemia.

In a study conducted by (Safi, Elnamrouty, 2012) to investigate the determinants of poverty in Palestine, where data was collected from 852 families, and a logistic regression model was used to identify influential and significant factors, it was shown that 9 factors influence the occurrence of poverty, including family size, residential area, number of unemployed in the family, number of children under 18, and other factors. Despite the importance and results of the study, the sample size is small for the population of Palestine, and a larger sample size would have provided better and more accurate results.

Previous studies and research have addressed some gaps and challenges that were overcome and utilized in this study. For example, gaps in some studies were represented by small sample sizes. In the study (Cui et al., 2023), the average age of the sample was in the fifties, which required increasing the sample size to include young adults. Furthermore, in the study (Abdul Khaliq, 2020), the sample size was small and insufficient for a city like Alexandria. Similarly, in the study (Al_Bairmani and Ismael, 2021), the sample size was insufficient for a city like Hillah, which in turn may not provide accurate results and lead to generalizations. Therefore, an appropriate sample size was determined for Jenin Governorate using appropriate statistical methods.

Another gap was the small number of independent variables. Ibrahim and Taher (2017) studied the impact of only 15 independent variables, while Aljammal (2023) only studied 8 independent variables, which could also lead to inaccurate results. Therefore, 19 independent variables were formulated in this study.

Another gap is that some studies that investigated the impact of genetic, environmental, and lifestyle factors only examined genetic, environmental, behavioral, or lifestyle factors, and did not combine these factors. Kujala (2011) did not examine genetic factors, Malak et al. (2025) neither genetic nor environmental factors were studied, and Meng et al. (1999) also lacked a study of behavioral and lifestyle factors, which may interact with each other and increase their impact. Therefore, the impact of all these factors was studied in this study. Some studies focused on factors with or without statistical significance and did not provide a predictive model. In this study, the influence of factors was examined, their statistical significance was ranked, and a model for predicting disease incidence was presented.

Chapter Three: Methodology

In this chapter of the study, the statistical methods and techniques used will be explained, starting with explaining regression, its characteristics and types, then logistic regression, its types, characteristics and advantages, and the reasons for using binary logistic regression, then explaining the method of analyzing the logistic regression model and its characteristics, passing through tests and methods of examining the suitability of the logistic regression model, and ending with the appropriate sample size for Jenin Governorate, and the method of collecting data represented by the questionnaire.

3.1 Regression

The first person to use the term "regression" in English was Galton (1866), Regression in statistics is a data analysis method used to determine the relationship between two or more variables. Regression analysis is used to predict the value of the dependent variable based on the values of the independent variables through a relationship between the dependent and independent variables.

Logistic regression is one of the parametric regression models, which assume that data follow a normal distribution. These models include quantitative models such as linear and nonlinear regression, and logistic regression, in which the dependent variable is binary. These models use the maximum likelihood method or other methods. These models are characterized by ease of interpretation and a simple estimation algorithm (Ravikumar et al., 2009).

There are several regression models, including linear regression, which is the simplest and most widely used form of regression, as it assumes a linear relationship between the dependent and independent variables. Linear regression analysis is not used when the dependent variable is binary such as 0,1 because it fails in the analysis. (Babtin, 2010).

Linear regression will not be used in this study for the following reasons:

1- Irrational expectations that do not fall within the range 0,1. They may be outside this range.
2- Linear regression assumes that the data is normally distributed, which may not be suitable for binary variables.
3- Linear regression assumes that the variance is constant, which may lead to inaccurate results.
4- Linear regression assumes that the relationship between the variables is linear, and therefore may give inaccurate results when the relationship is non-linear.
(Gujarati, Porter, 2009).

3.2 Logistic regression

Logistic regression model is one of the parametric that considered linear for coefficients of model, and nonlinear model for the probability, and the dependent variable is binary or more, and it is used to predict the occurrence of something in the future, and it has many important applications in many areas of life such as medicine, marketing, finance, education and social sciences (Garson, 2014).

In research and analysis, we often encounter situations in which we want to predict the occurrence of something or a result, based on a set of independent variables, and mathematics is one of the broadest and most widely used sciences in a wide range, as it enters into many applications, and among these applications are those related to the medical aspect and human health, where we can predict the occurrence of a certain disease using statistics, through logistic regression, which is a model for statistical analysis, used to find the relationship between variables, when the dependent variable takes two or more values, and thus develop models for prediction for variables, and variables that are subject to parametric conditions (Montgomery, and others, 2012).

Logistic regression is also known as the logit model and others, and is used in many scientific and commercial applications and is one of the most widely used models in the field of machine learning (Alexopoulos, 2010). As we mentioned, linear regression is used when the variables have a linear relationship, especially data that has continuous outcomes (0,1) where the outcomes are not binary, so logistic regression is the alternative for this data in which the dependent variable is binary (Komarek, 2004).

In this study, we will study the relationship between the variables: age, gender, chronic disease, whether a genetic test has been performed before, genetic counseling has been provided, whether one or both parents have a chronic disease, whether brothers or sisters have been infected, or second-degree relatives, if a specific disease has been observed to be prevalent in the family, living in the countryside or the city, a healthy and balanced diet, physical and sports activity, exposure to the open air, psychological and social pressures, exposure to the oppression of the occupation and its punitive measures, exposure to toxic substances in the workplace, living near factories or sources of pollution, smoking, drinking alcohol or drugs, exposure to passive smoking, undergoing periodic medical examinations and obtaining the necessary health care when needed.

This is done through a random sample and filling out a questionnaire that includes these variables, and using the results to estimate the value of the dependent variable (the emergence of a chronic disease) and predict it, In the binary variable, the result is P(Y=0) when failure or the event does not occur or the disease is not present, or the result is P(Y=1) when success or the event occurs or the disease is present. The logit function is used to convert the relationship into a linear one by:

$$\text{Log}(\text{p}(\text{xi})) = In(\frac{p(xi)}{1-p(xi)}) = \beta_0 + \beta_1 \text{X}_1 + \beta_2 \text{X}_2 + \ldots + \beta_i \text{X}_i \text{ , for i= 1,2,3,.....}$$

By exponent both side, the equation become:

$$\frac{p(xi)}{1-p(xi)} = e^{\beta_0 + \beta_1 \text{X}_1 + \beta_2 \text{X}_2 + \ldots + \beta_i \text{X}_i} \text{ , for i= 1,2,3,.....}$$

$$p(xi) = \frac{e^{\beta_0 + \beta_1 \text{X}_1 + \beta_2 \text{X}_2 + \ldots + \beta_i \text{X}_i}}{1 + e^{\beta_0 + \beta_1 \text{X}_1 + \beta_2 \text{X}_2 + \ldots + \beta_i \text{X}_i}} \text{ , for i= 1,2,3,....}$$

$P(X_i)$: probability of success.

$1 - p(X_i)$: probability of failure.

$\beta_0, \beta_1, \beta_2, \ldots, \beta_i$: coefficients of regression.

$X_1, X_2, \ldots, X_i$; independent variables.

The previous model is called the logistic regression model. (Al-Khazali, 2021).

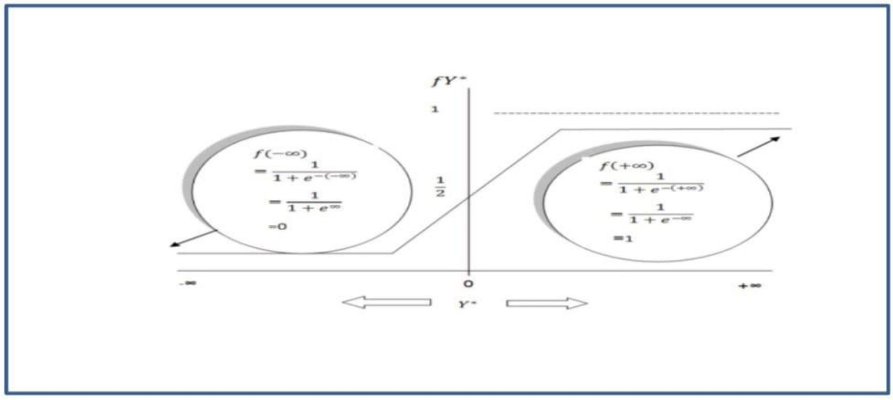

Figure 3.1: representing the logistic function.

From the previous figure, we notice that the function is bounded between 0 and 1.

The Importance of Logistic Regression:

- Dealing with binary variables with high efficiency that carry the result (success or failure, yes or no).
- Easy to interpret as the role of each independent variable on the probability of the event occurring can be clarified.
- Wide applications in many fields in medicine, finance, marketing, insurance and social sciences.
- Logistic regression is computationally efficient and can be used on very large data without the need for huge resources.
- Logistic regression can classify independent variables according to their impact on the dependent variable (Menard, 2002).

The Types of Logistic Regression:

- Binary Logistic Regression
- Multiple Logistic Regression
- Ordinal Logistic Regression

3.2.1 Binary logistic regression

It is one of the logistic regression models, and is used when the dependent variable is binary, meaning that it takes only two values such as 0,1 or (yes, no), and takes the value 1 when a specific event occurs and takes 0 when the event does not occur (Hosmer et al., 2013).

This regression helps determine the relationship between the independent variables and the binary dependent variable, which helps predict the occurrence of a specific event based on the values of the independent variables.

Conditions for binary logistic regression:

- The dependent variable must be binary, and have one of the two values 0,1 or (yes, no).
- There must be no multicollinearity between the independent variables, meaning that the independent variables must not be highly correlated.
- It must consist of more than one independent variable.
- There must be no outliers in the independent variables, and these outliers can be identified using (Mahalanobis test).
- It is preferable to have a large sample size, because this gives more accurate and stronger results. (Peduzzi, and others, 1996).

These conditions must be met in order to apply logistic regression (Sari, Daaish, 2017).

3.2.2- Multiple Logistic Regression

One of the types of logistic regression models, it is distinguished by its great importance in data analysis and has many uses in many fields., and is used when the response variable is due to nominal or ordinal variables, and consists of more than two levels, and the probability function is used to estimate the model parameters, and this regression is an extension of the binary logistic regression, and depends on the multinomial distribution (El-Habil, 2012).

It is preferable to use multiple logistic regression for several reasons, including flexibility in modeling, as it is more flexible in non-linear relationships between variables and thus increases its accuracy in predicting results (Hosmer and others, 2013).

It also deals with categorical variables, as linear (ordinary) regression deals with numerical variables, while multiple logistic regression deals with categorical variables such as classifications that consist of more than one category (Daish, Sari, 2017).

Multiple logistic regression has many methods for estimating model parameters, the most famous of which is the Maximum Likelihood Estimation (MLE) method, which is considered the most widely used (Dobson, Barnett, 2018). The derived equations are also solved using the maximum likelihood method by using the Newton-Raphson method, which uses the least squares method. (Al-Khazaali, 2021).

3.2.3- Ordinal logistic regression

It is one of the types of logistic regression, and ordinal because it is used when the dependent variable is an ordinal variable, i.e. it can be classified into categories such as (always, often, sometimes, never) and other classifications, and it works to estimate the probability of each category of the dependent variable occurring based on the independent variables, this type of logistic regression is more prevalent in medical aspects, market research and opinion polls (Dobson, 2018).

The advantages of ordinal logistic regression are:

- It is characterized by high accuracy.
- It is considered a distinctive test in determining the strength of relationships.
- It is used when the dependent variable is an ordinal variable (Warner, 2008).

Ordinal logistic regression function

$$\text{logit}[\Pr(Y = j|x)] = \log\left(\frac{\Pr(Y=j|x)}{1-\Pr(Y=j|x)}\right) = \alpha_j + \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i \ , i = 1,2,3,\ldots$$

$P(X_i)$: probability of success.

1-$P(X_i)$: probability of failure.

β_0: intercept parameter.

β_i: slope parameter.

3.3 Logistic regression transformation

When the dependent variable is binary, it carries one of the two values 1 or 0, and then we may face some problems, to solve these problems we use the odd ratio, which represents the ratio of the probability of an event occurring to the probability of its non-occurrence. if P is the probability of an event occurring, then (1-P) is the probability of its non-occurrence, and thus odds is the result of dividing the probability of occurrence by non-occurrence and is equal:

$$\text{Odd ratio (OR)} = \frac{p}{1-p}$$

It determines whether there is a link between two events through the odds ratio, and it can be considered as the probability of an event occurring when there is an exposure, compared to the probability of this event occurring when there is no exposure. (Suzmilas, 2010), For example, Odds ratio is the ratio of the probability of something occurring to its non-occurrence. If the probability of something occurring is 0.6, then odds equals 3/2 or 1.5, While probability is the ratio of the occurrence of something to everything that can

happen, another example, if the probability of an event (P) is 3/4, then the odds ratio will be 3, meaning 3 times it will happen for every time it won't happen.

As a simple explanation, when the odds ratio is less than 1, there are more failures than successes, and when the odds ratio is equal to 1, the chances of success and failure are equal, this means that there is no relationship between X and Y, and when the odds ratio is greater than 1, there are more successes. (Abu saada, 2013).

$$P = \alpha + \beta x$$

The probability value may not be within the range of 0.1 because the value of x changes, so we will solve this problem by converting the probability using the logarithm

$$\text{Logit(OR)} = \text{In}(\frac{p}{1-p}) = \alpha + \beta x$$

P**:** probability of something happening.

α, β: Regression parameters.

X **:** independent variable.

With this equation, the probability is modeled as a linear function of the independent variable x, By exponentiation both sides of equation:

$$\frac{p}{1-p} = e^{\alpha + \beta x}$$

After doing some simple steps the final look will be:

$$P=P(y = \text{certain value}, X=x) = \frac{1}{1 + e^{-(\alpha + \beta x)}}$$

This logistic regression equation represents the prediction equation, which transforms the linear regression to obtain a positive value. Which indicates the probability of the event occurring This equation, when represented, is given in the form of the letter S with the points 0 and 1. The extreme points but in the case of multiple logistic regression (in which the independent variable x is multiple), the equation changes to another form as follows:

Logit(P(Y=1)) = $\beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta ixi,$ where k is number of predictors for y.

(El-habil, 2012)

$$P = \frac{e^{\ \alpha + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ + \ldots + \beta_i X_i}}}{1 + e^{\alpha + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ + \ldots + \beta_i X_i}}}$$

when y is specific value and $X = X_1, X_2, \ldots, X_i$

$$P = \frac{1}{1 + e^{-(\alpha + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ + \ldots + \beta_i X_i})}} \qquad \text{for i: 1,2,3,.....}$$

(mnahi, kamr,2018)

3.4 Maximum likelihood method

In logistic regression, the maximum likelihood (MLE) method is used to estimate the parameters of this model, and it is more appropriate than the least squares method, as seen by (Eliason, 1993), due to its compatibility with the logistic regression model and its distinctive statistical properties, especially when the sample size is large.

The maximum likelihood method is one of the most commonly used statistical methods for estimating the parameters of statistical models, and this method depends on repetition in the estimation process, as it depends on repeating the arithmetic operations repeatedly, i.e. many times, until the appropriate optimal estimates are obtained, which enables us to interpret the data.

In general, the maximum likelihood method is suitable for linear and nonlinear models, while the least squares method is suitable for linearity only, and the maximum likelihood method does not require any conditions for the independent variables, which means that it is suitable regardless of whether the independent variable is nominal, ordinal or categorical.

It is used in logistic regression to calculate the logit coefficients, and the aim of using it is to know the extent to which the outcome of the dependent variable can be predicted, based on the independent variables, and this method depends on repetition, as it starts with a specific value and then determines the amount of change in the logit coefficients, which increases the logarithm of the probability (Walker, 1996).

The likelihood function

$$L(\beta,Y) = \prod_{i=1}^{N} P(X_i)^{Yi} (1 - P(X_i))^{Ni-Yi} \quad (1)$$

After simplification the equation will become:

$$L(\beta,Y) = \prod_{i=1}^{N} \left(\frac{P(Xi)}{(1-P(Xi)}\right)^{Yi} (1 - P(X_i))^{Ni} \quad (2)$$

Form the previous chapter we know:

$$\text{In } \left(\frac{p(Xi)}{1-P(Xi)}\right) = \beta_0 + \beta_i X_i \quad (3)$$

By exponent the both side of equation:

$$\frac{p(Xi)}{1-P(Xi)} = e^{\beta 0+\beta iXi} \quad (4)$$

So by compensation in (2) equation:

$$L(\beta,Y) = \prod_{i=1}^{N} \left(e^{\beta 0+\beta iXi}\right)^{Yi} \left(1 - \frac{e^{\beta 0+\beta iXi}}{1+ e^{\beta 0+\beta iXi}}\right)^{Ni} \quad (5)$$

By algorithm both side of equation:

$$\ln L (\beta, Y) = \sum_{i=1}^{N} [Y_i (\beta_0 + \beta_i X_i) - N_i \text{In}(1 + e^{\beta 0+\beta iXi})] \quad (6)$$

L (β, Y): the likelihood function.

P: the probability.

N: the number, (Nasrawi,2017).

Then we differentiate the possibility function with respect to β, and then we set the derivative equal to zero, and then we solve this by certain methods such as the Newton-Raphson method, as this method gives suitable estimates of the parameters. (Al-Khazaali, 2021).

Logistic model transactions

We will interpret and explain logistic regression coefficients from several aspects Logistic regression is interpreted in terms of the logit coefficient which is used to estimate the logit coefficient, so the dependent variable is equal to 1 for each unit change that occurs in the independent variable, and logistic regression calculates the value of the change in the logarithm of the weighting coefficient of the dependent variable and not the change in the dependent variable like linear regression (Garson, 2006).

It can be said that the change in the coefficients in terms of the logit gives an interpretation like the interpretation in linear regression, but the difference is that the units of the dependent variable represent the logarithms of the weighting coefficients (Babtin, 2009), and we take the absolute values of the regression coefficients, to determine the effect of each variable and then we can rank the variables based on these absolute values.

Also for the odds ratio (OR) where it expresses the relative change for each increase in the independent variable, in addition, in terms of probabilities, it can be said that an increase in the independent variable by one unit will lead to an increase in the logit or logarithm of the weighting coefficient. (Garson, 2006), He also stated these interpretations and explanations (Hosmer et al., 2013), (Menard, 2002).

3.5 Model suitability evaluation

There are many methods and means used in logistics to examine a small part of this model, so we will learn about these methods and means.

3.5.1- table Classification

A classification table is a table that shows the number of (positive) cases that have a particular trait predicted by the model, and the number of (negative) cases that do not have

that trait predicted by the model compared to the actual observed cases, whether positive or negative. (Peng et al, 2002).

Table 3.1: The Classification table.

Observed	Positive Expected (PE)	negative Expected (NE)	Total
Positive (P)	True positive (TP)	false negative (FN)	TP + FN
Negative (N)	False positive (FP)	True negative (TN)	FP + TN
Total	TP + FP	FN + TN	TP + FP + FN + TN

- Sensitivity (S): It is calculated through the table, it is also known as the positivity rate and it is considered the probability value that the classification is positive for the case that is considered positive, and it is calculated as follows:

$$S = \frac{TP}{TP+FN}$$

- Specificity (SP): It is calculated through the table, it is also known as the negative rate, and it is considered the probability value that the classification is negative, for the case that is negative, and it is calculated as follows:

$$S_P = \frac{TN}{FP+TN}$$

- Hit ratio (accuracy): It is calculated through the table, and is also known by another name, which is the efficiency ratio, and is considered a value for the probability of correct classification, and is calculated as follows:

$$E_T = \frac{TP+TN}{\mathrm{TP + FP + FN + TN}}$$

- Odds :We can say that it is considered the probability of success relation to the fail, and it is calculated as follows:

$$\mathrm{Odd} = \frac{\mathrm{TP + FN}}{\mathrm{FP + TN}}$$

- Odd ratio: It is defined as the probability of a certain thing occurring, divided by the probability of the thing not occurring, and the Odd ratio is the ratio of the odd coefficient of a variable to the odd coefficient of another variable, and it is calculated as follows:

$$\text{OR} = \frac{\text{TP} \,.\, \text{TN}}{\text{FN} \,.\, \text{FP}}$$

3.5.2 - Goodness-of-fit test

Goodness-of-fit in binary logistic regression refers to how closely the model's predicted probabilities match the observed outcomes, and includes two tests:

3.5.2.1 - chi-square test (x^2)

It is a test proposed by Pearson, and is known by the following mathematical formula**:**

$$x^2 = \sum_{i=1}^{N} \frac{(X_i - Y_i)^2}{Y_i}$$

X_i: viewed values.

Y_i: expected values.

N: number of views.

If the test value is greater than the significance level that was determined, i.e. it is not significant, then the observed values and expected values are equal, and this confirms that the model is appropriate and suitable for the data.

It is a non-parametric test used to test independence, i.e. to test whether there is a relationship between categorical variables, whether they are independent or related. (Ramadan, 2023).

The null hypothesis is accepted if the chi-square value is equal to the tabular chi-square value, which means that there are no differences between the observed and expected frequencies. The null hypothesis is rejected if the chi-square value is not equal to the tabular

chi-square value, which means that there are differences between the observed and expected frequencies.

3.5.2.2 Hosmer-lemeshow test

It is a statistical test developed by Hosmer and Lemeshow in 1980. This test is used to evaluate the suitability of the logistic regression model, as it tests the hypothesis of differences between the observed data and the expected data based on the model, which means in other words that it tests the model's ability to accurately predict events.

This test is of great importance in the analytical aspect of statistics and is widely used for its simplicity of use and ease of interpretation of the results, it is widely used in medicine, for example, (Nashef et al., 1999), made a model to predict early deaths among heart surgery patients in Europe, where they evaluated the model using the Hosmer and Lemeshow test.

The test divides the data into groups and the expectations are calculated for each group. This test is considered a chi-square test because it uses the chi-square in the final comparison, where the n records in the data set are sorted according to the estimated probability of success, and then the group is divided into other smaller groups of equal size. (Hosmer, Lemeshow, 1980).

In a study conducted by (Paul et al., 2013), when examining the suitability of the regression model using the Hosmer and Lemshow test. They recommended that the upper limit of the target sample should not exceed 25 thousand, and if it exceeds it, the strategies they recommended should be used.

Hosmer-Lemeshow C statistic evaluation:

$$\hat{C}_g = \sum_{i=1}^{g} \left(\frac{(O_{s.i} - E_{s.i})^2}{E_{s.i}} + \frac{(O_{f.i} - E_{f.i})^2}{E_{f.i}} \right)$$

$O_{s.i}$ *:* Observed number of success.

$E_{s.i}$ *:* Expected number of success.

$O_{f.i}$: Observed number of failure.

$E_{f.i}$: Expected number of failure.

the p-value for the Hosmer–Lemeshow:

$$\mathrm{P} = \int_{\hat{C}_g}^{\infty} x_{g-2}^2\,(x)dx.$$

x_{g-2}^2 : It symbolizes to probability density function of x^2 distributions with g-2 degrees of freedom evaluated at x . (Paul et al., 2013).

3.5.3 Wald test

It is one of the tests used to evaluate logistic regression, and it tests whether the model coefficients differ from zero, in other words the test is used to know the effect of each independent variable in the regression, and this test follows a distribution, and the value of Wald is directly proportional to the effect and contribution of the independent variable. The test is done by assuming the logit coefficient according to the null theory that it is equal to zero, then calculating the value of Wald (Kleinbaum & Klein, 2010)

The process of comparing the Wald statistic with one degree of freedom with the critical value from the Chi-square distribution is done, and the significance of the parameters must be less than 0.05 in order to reject the null hypothesis and accept the alternative.
The null hypothesis states as follows:

H_0 : β = 0, There is no effect of the independent variable on the dependent variable.

H_1 : β ≠ 0, There is an effect of the independent variable on the dependent variable.

β: The regression coefficient of the independent variable.

The null hypothesis assumes that the B is equal to zero, meaning that the independent variable has no effect on predicting the value of the dependent variable, and the alternative hypothesis assumes that the B is not equal to zero, meaning that the independent variable has a contribution and role in predicting the value of the dependent variable (Menard, 2002).

The Wald statistic with one degree of freedom is compared to the critical value of the chi-square distribution, and the significance of the parameters must be less than 0.05 in order to reject the null hypothesis and accept the alternative.
The Wald statistic is calculated based on the following formula:

If w^2 tracking x^2distribution , the formula is , $W = (\frac{\beta}{S_{EB}})^2$

If w^2 tracking z distribution , the formula is, $W = \frac{\beta}{S_{EB}}$

W: wlad statistics.
β: Regression coefficient of the independent variable.
S_{EB}: Standard deviation of the regression coefficient of the independent variable.

It is widely used in medical and social studies to determine and evaluate the importance of independent variables in regression, in addition, it was found that the power of the test may be weakened under some circumstances, especially if there is a large amount of variation in the data. (Basu, Ghosh, 2016).

3.5.4 Confidence interval estimation

A statistical tool used to test the accuracy of the logistic regression, where the possible range of the true coefficient is estimated based on a specific sample and a certain level of confidence, usually (95%). In order to evaluate the extent of the influence of the independent variables on the dependent variable.

Estimating a confidence interval is essential in logistic regression because it assesses the confidence in the estimates that have been determined. (Hosmer et al., 2013).

The confidence interval for the parameter is as follows:

$$\mathrm{CI} = \hat{B}_i \pm Z_{(1-\frac{\alpha}{2})} . \mathrm{SE}(\hat{B}_i)$$

CI: confidence interval.

$\hat{B}_i$: parameter.

$Z_{(1-\frac{\alpha}{2})}$: critical value.

SE($\hat{B}_i$) : standard deviation

Where the standard deviation is the square root of the diagonal elements in the covariance matrix, And the confidence interval for the odd ratio is:

$$\mathrm{CI} = e^{\hat{B}_i \pm Z_{(1-\alpha/2)}.\mathrm{SE}(\hat{B}_i)}$$

The confidence interval for statistically significant factors should not include the number 0 for the regression model coefficients, and the confidence interval for the likelihood ratio should not include the number 1.

3.5.5 Coefficient of determination

It is a relative variance in the dependent variable that is to be predicted from the independent variable, and its value ranges between 0 and 1, and expresses the strength of the model in prediction, and determining the proportion and role of the independent variables and their effect on the dependent variable.

This is done through the following Coefficients:

3.5.5.1 Cox and Snell coefficient ($R_{CS}{}^2$)

$$R_{CS}{}^2 = 1 - e^{\left[\frac{2}{n}(Ln(New) - Ln(Baseline))\right]}$$

$$R_{CS}{}^2 = 1 - [e^{\frac{1}{2}(New - Baseline)}]^2$$

Ln (New): Log likelihood of the new model after including the variables

Ln (Baseline): Log likelihood of the original model after including only the constant term

N: Total number، (Babtin, 2009).

But this coefficient does not include the number 1, so a coefficient was developed that includes the integer 1 in a simple way by dividing by the maximum value of the previous coefficient ،this is done through :

3.5.5.2 Negelkerke coefficient (R_N^2)

The advantage of this test is that it includes reaching the maximum value and number (1).

$$R_N{}^2 = \frac{R_{CS}{}^2}{1 - e^{\frac{[2\ln(Baseline)]}{n}}}$$

$$R_N{}^2 = \frac{R_{CS}{}^2}{1 - [e^{-\frac{Baseline}{2}}]^{\frac{2}{n}}}$$

3.5.6 multicolinearity analysis test

This test is used in logistic regression to check if the independent variables are correlated with each other, which negatively affects the model. This is done in SPSS by checking the variance inflation factor (VIF), which expresses the size of the inflation in the variance of the regression coefficient due to the correlation between the independent variables.

VIF values:

- If its value is between 1 and 5, there is no major problem in the correlation between the independent variables.
- If its value is between 5 and 10, there is a moderate problem in the correlation between the independent variables.
- If its value is greater than 10, there is a major problem in the correlation between the independent variables and it must be dealt with. (O'Brien, 2007)

3.5.7 likelihood-ratio test

This test is one of the most important tests in evaluating and diagnosing the suitability of the logistic regression model, where two different models are compared, one of which contains all the independent variables, and the other does not contain all the independent variables. This is done in order to determine whether the independent variables improve the performance of the logistic regression model.

It is calculated using the following rule:

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

H_0 : $\beta = 0$

H_1 : $\beta \neq 0$

Then we compare the value of G with the value of x^2. If G is greater, we reject the null hypothesis.

3.6 Sample size

The sample size plays an important role in the accuracy and credibility of the study, and there are many factors that affect it, such as the size of the target community and the required and possible degree of confidence. The sample size in the study was determined based on scientific statistical methods, through the use of a simple random sample, which aims to randomize the selection and reduce bias as much as possible and objectivity.
The target sample size can be determined through several methods, including: -

3.6.1 Krejcie and Morgan method.

This method is used to determine the appropriate sample size based on the population size, the required confidence level, and the acceptable error rate, and the mathematical formula for this method is:

$$S = \frac{X^2 .\ N\ .\ P(1-P)}{D^2 .\ (N-1) + X^2 .\ P(1-P)}$$

S: sample size.

X: Standard value of confidence level.

N: population size.

P: The proportion of the trait in the population is usually 0.5 if it is not known.

D: acceptable error rate.

The sample to be taken is from Jenin Governorate, whose population, according to the Palestinian Central Bureau of Statistics, is 359,934 people for the year 2024. According to the 95% confidence level, and 0.05 acceptable error rate. so The sample size is

$$S = \frac{(1.96)^2 .\ (359934)\ .\ (0.5)(1-0.5)}{(0.05)^2 .359933 + (1.96)^2 .(0.5)(1-0.5)} = \frac{(3.8416).(89983.5)}{(0.0025).(359933)+(3.8416).(0.25)} = \frac{345680.6136}{900.7929} = 383.75$$

$$\approx 384$$

3.6.2 Thompsons method

This method is also used to determine the sample size and its mathematical formula is:

$$n = \frac{N.Z^2 .\ P.(1-P)}{E^2 .(N-1) + Z^2 .\ P.(1-P)}$$

n: sample size.

N: population size.

Z: Standard value of confidence level.

P: The proportion of the trait in the population is usually 0.5 if it is not known.

E: acceptable error rate.

After calculating the sample size through this equation, the same result will be obtained for the sample, which is approximately 384 people.

3.7 Questionnaire

After determining the target sample size, data was collected using a questionnaire. The questionnaire was designed to include questions related to genetic factors, environmental, behavioral, and lifestyle factors, based on data from the World Health Organization. All questionnaire questions were binary. The following table shows all the independent factors, along with their codes, as the questionnaire questions addressed these factors, the factors studied are 19, and are arranged as follows:

Table 3.2: the factors and their symbols.

Factors	The symbol
Age group	X1
Genetic counseling or testing	X2
Living in an urban or rural area	X3
Genetic counseling or testing	X4
One or both parents having chronic diseases	X5
Brothers or sisters having chronic diseases	X6
specific pattern of disease in A family or extended family	X7
Following a healthy and balanced diet	X8
Committing to exercise and exposure to fresh air	X9
Smoking	X10
Addiction to alcohol or drugs	X11
Being in areas where smoking occurs(passive smoking)	X12
Exposure to psychological and social stress	X13
Suffering from financial and Economic stress	X14
Exposure to oppression by the occupation or its punitive measures	X15
Exposure to chemicals at work	X16
The presence of factories or sources of pollution nearby	X17
Regular medical checkups help identify and detect potential diseases	X18
Receiving healthcare when needed	X19

Chapter Four: Results

This part of the study analyzes the collected data and presents a comprehensive and detailed picture of the results obtained using appropriate statistical methods and hypothesis testing, and this includes several sections, beginning with presenting the preliminary results of the sample, then converting them to percentages and analyzing them using a binary logistic regression model in SPSS. The final analysis results are then obtained, the model equation is formulated, and a model is presented to predict the likelihood of developing chronic diseases.

The binary logistic regression model will be used as a tool to analyze the sample data, with the goal of developing a model to predict the likelihood of developing a chronic disease using SPSS. The effect of the independent variables on the binary dependent variable, the likelihood of developing a chronic disease, will be determined, and the extent of the independent variables' influence on the dependent variable will be clarified and ranked accordingly.

4.1 Descriptive analysis of survey data

The target sample in Jenin Governorate, which totaled 387 individuals, was randomly selected. Of these, 253 were free of chronic diseases and 134 were suffering from chronic diseases. The results of the 19 variables were converted to percentages. The independent variables were classified into five groups, each represented by a set of figures

The first group represents demographic factors (sex, age group, urban or rural area). Regarding the first factor, gender, it was observed that the proportion of healthy males was higher than that of females, and that the proportion of affected males was lower than that of females. This may indicate a potential influence of gender on disease onset, possibly due to certain hormonal factors. (figure 4.1)

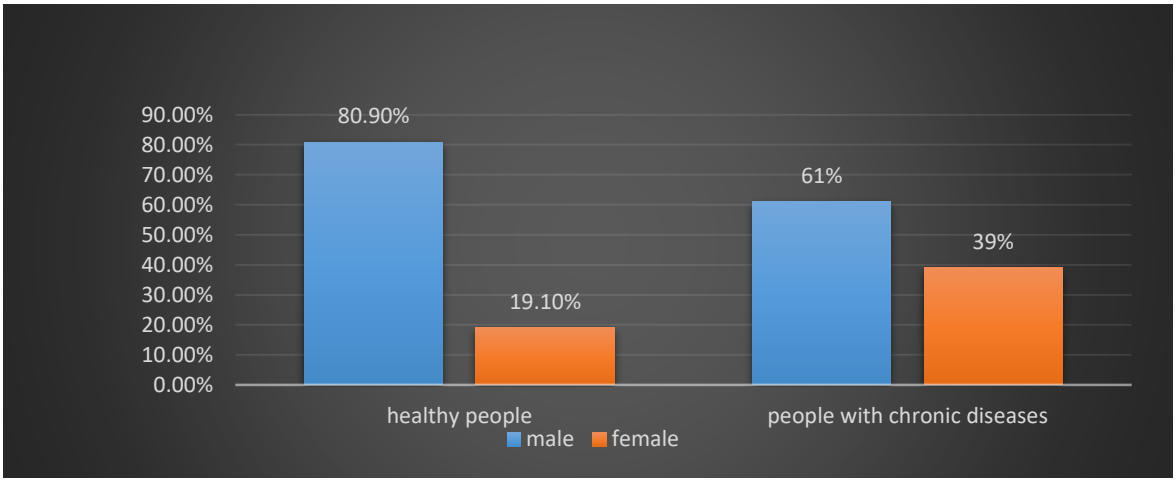

Figure 4. 1: Distribution of infected and healthy people by gender.

The second demographic factor is age. While the proportion of healthy and infected children did not change significantly, the proportion of young people constituted the largest group among both healthy and infected individuals, despite their lower proportion among infected individuals, as most study participants were young. Conversely, the proportion of elderly people among infected individuals increased significantly compared to healthy individuals, suggesting that aging may influence the incidence of chronic diseases. (figure 2)

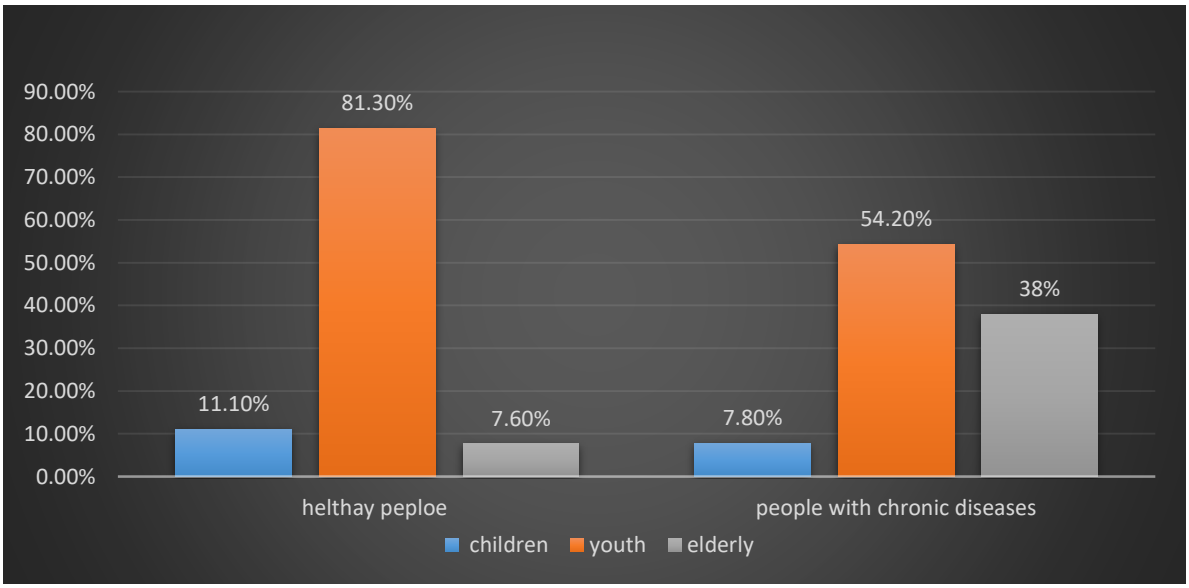

Figure 4. 2: Distribution of individuals by age group.

The third demographic variable is the place of living in the city or countryside, From the figure, it was noted that there was no significant difference between healthy and infected people between rural and urban areas, with a slight increase in infected people in rural areas, which may indicate the possibility of an impact on the occurrence of diseases, even if it is small, and this may be due to agricultural activities or the presence of pollution sources. (figure 4.3)

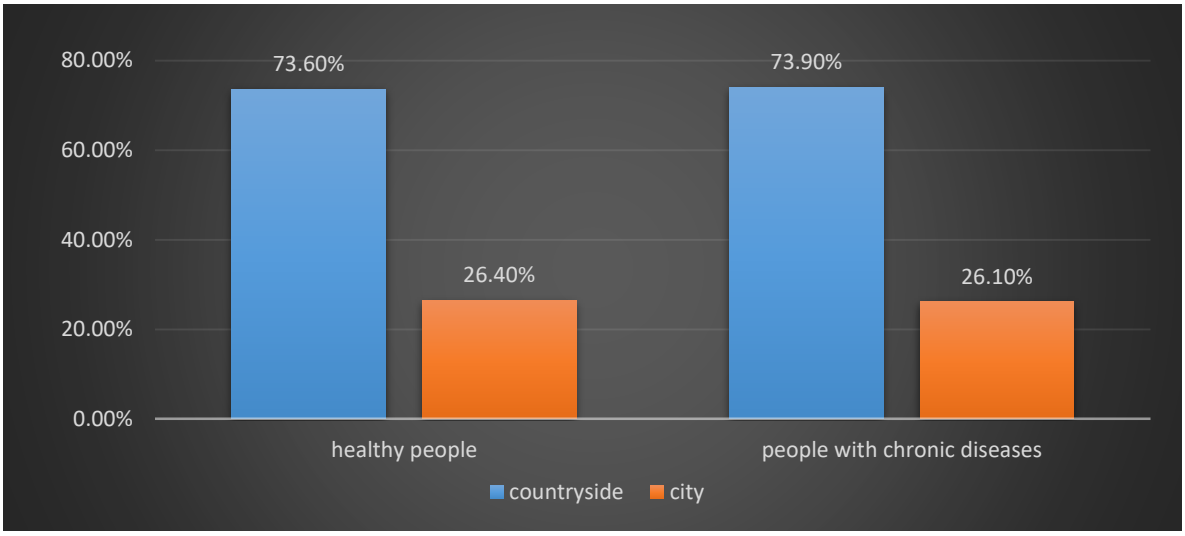

Figure 4. 3: Distribution of individuals according to place of residence.

The second group of variables represents genetic factors and family history of disease. It includes five factors: genetic testing or counseling, one or both parents with chronic diseases, siblings with chronic diseases, second-degree relatives with chronic diseases, and the presence of a specific pattern of diseases in the extended family, (Figure 4.4) shows the distribution of individuals with and without chronic diseases who responded positively to these factors.

It is noted from the figure that the proportion of individuals who underwent genetic testing or counseling among those with chronic diseases was higher than among those without. This may indicate a potential impact on the onset of chronic diseases, highlighting the importance of early screening.

It is also noted from the figure that the proportion of individuals with chronic diseases whose one or both parents were affected was higher than the proportion of those without chronic diseases whose one or both parents were affected. This indicates a potential impact on the occurrence of chronic diseases based on the presence of one or both parents.

The figure shows that the proportion of affected individuals with affected siblings is higher than the proportion of healthy individuals with affected siblings, indicating the potential impact of having siblings with chronic diseases on an individual's chronic disease incidence rate.

The figure also shows that the proportion of affected individuals whose extended family members suffer from a specific type of chronic disease is higher than the proportion of healthy individuals whose extended family members suffer from a specific type of disease, indicating the potential impact of having a specific type of disease in the family on an individual's chronic disease incidence rate. (Figure 4.4)

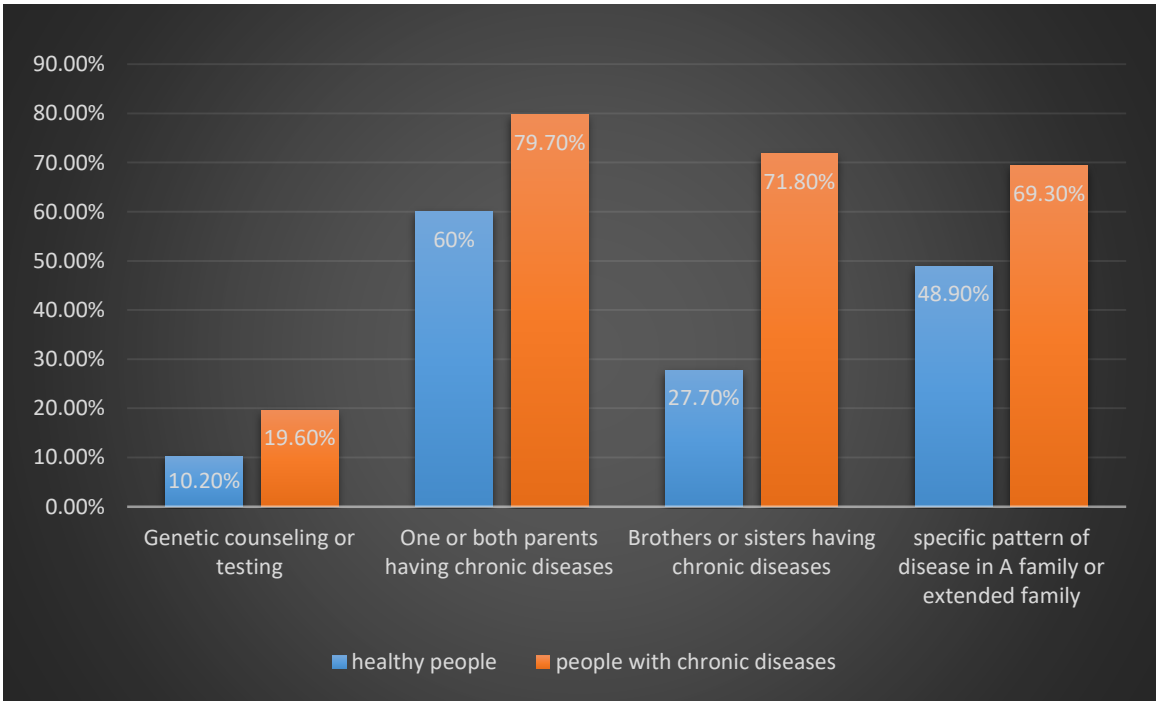

Figure 4. 4: Genetic factors and family history of diseases

The third group of variables is behavioral and lifestyle factors, which include five factors: adherence to a healthy and balanced diet, regular exercise and exposure to fresh air, smoking, alcohol or drug addiction, and exposure to secondhand smoke, (Figure 4.5) shows the distribution of infected and healthy individuals who responded positively to these factors.

It is evident from the figure that the proportion of healthy and affected individuals who adhere to a healthy and balanced diet is very similar, with a slight increase in the proportion of those who adhere to this diet compared to the healthy individuals. This indicates that it does not have a significant impact, but may have a long-term effect.

The figure also shows that the proportion of people who exercise and are exposed to fresh air is higher among healthy individuals than among those with chronic diseases, reflecting a potential impact on the incidence of chronic diseases.

The figure indicates that the proportion of smokers with the disease is lower than the proportion of healthy smokers. This is because most smokers are young and less susceptible to chronic diseases, and because smoking plays a significant role in the occurrence of chronic diseases, its effects are long-term. The figure shows that the percentage of alcohol, cannabis, and drug addicts is very small and low, and that the percentage of addicts suffering from chronic diseases is greater than the percentage of healthy addicts, indicating the potential impact of this on the incidence of chronic diseases.

The figure also shows a significant similarity in the numbers exposed to secondhand smoke, both those infected and those not infected, given the widespread prevalence of smoking in Palestinian society. (figure 4.5)

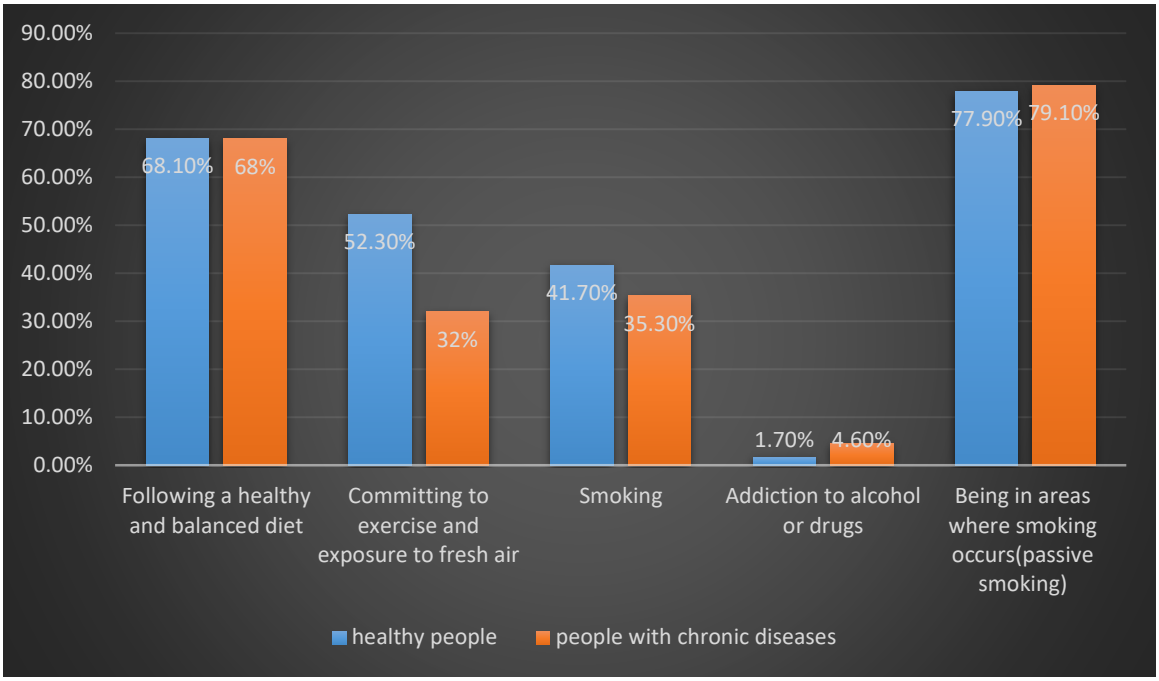

Figure 4.5: Behavioral and lifestyle factors.

The fourth group of variables is environmental and psychological factors, which includes four factors: exposure to chemicals in the workplace, the presence of factories or pollution sources close to the place of residence, exposure to psychological and social stress, and exposure to oppression and punitive measures by the occupation, (Figure 4.6) shows the distribution of infected and healthy individuals who responded positively based on these factors.

The figure shows that the incidence rate of the disease is very similar between those exposed to chemicals and those not exposed to them in the workplace, with a slight bias in favor of those exposed due to their long-term effects, indicating their weaker impact on the occurrence of chronic diseases.

The figure also indicates that the incidence rate of the disease is similar between those living near pollution sources or factories and those living farther away, with a slight bias in favor of those living near pollution sources due to their long-term effects, reflecting their weaker impact on the occurrence of chronic diseases.

The figure shows that the percentage of those with this condition who are exposed to psychological and social stress is higher than the percentage of healthy individuals who are exposed to it. This is because psychological and social stress can develop and cause depression, anxiety, or overthinking. This increases the risk of chronic diseases, indicating that these stressors have a high potential for causing chronic diseases.

As noted in the figure, the percentage of infected and healthy individuals subject to occupation sanctions is close to the percentage of those not subject to them. This indicates that exposure to occupation sanctions and oppression may have a weak impact on the incidence of chronic diseases, due to several reasons, including the absence of settlements or military bases in the Jenin area, there is no direct contact between residents and settlers, and most of those who participated in the survey had not been imprisoned before. (figure 4.6)

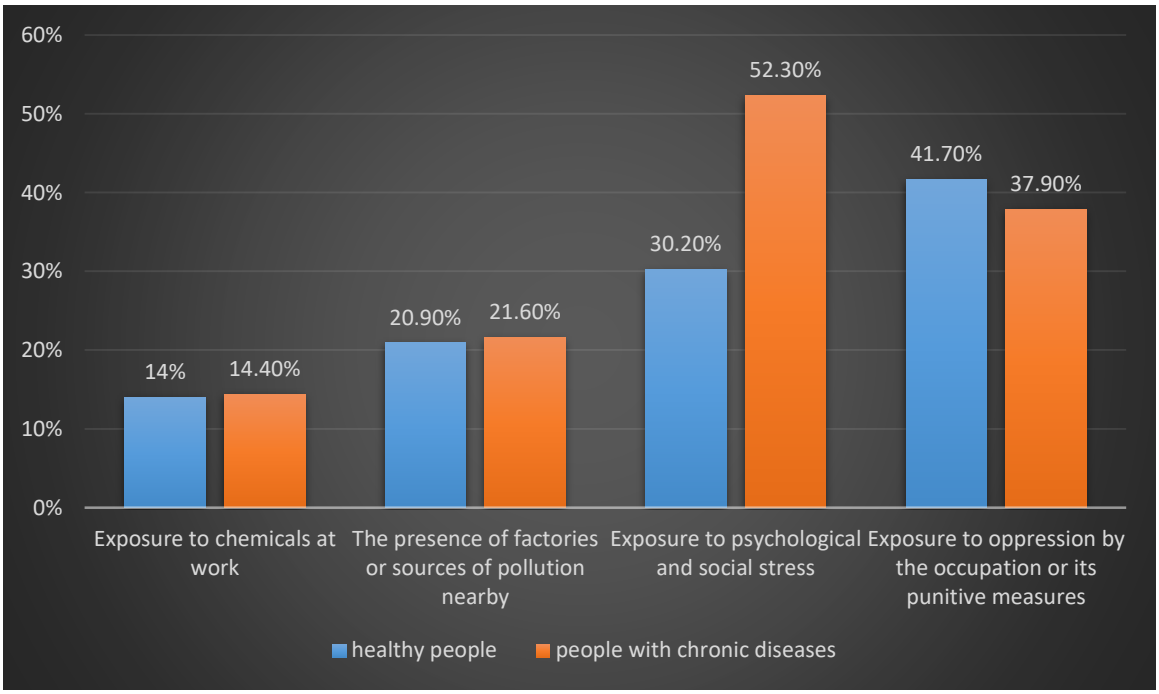

Figure 4.6: Environmental and psychological factors.

The fifth group of variables is health and economic factors which include: suffering from financial and economic stress, regular medical checkups, and receiving adequate

healthcare when needed. (Figure 4.7) shows the distribution of those who answered positively according to these factors.

The figure shows that the percentage of infected individuals who suffer from financial and economic stress is higher than the percentage of healthy individuals. This is due to the psychological and health effects resulting from malnutrition, which indicates its potential impact on the development of chronic diseases.

The figure also shows that the percentage of those who undergo regular testing among infected individuals is higher than the percentage of healthy individuals. This is due to a lack of interest in testing except when symptoms appear, and a lack of awareness and health prevention.

The figure shows that the percentage of those who receive adequate healthcare when needed is very high in Palestinian society, regardless of whether they are infected or not. However, the percentage of infected individuals is slightly higher, which may indicate the weak impact of this factor on the incidence of chronic diseases. (figure 4.7)

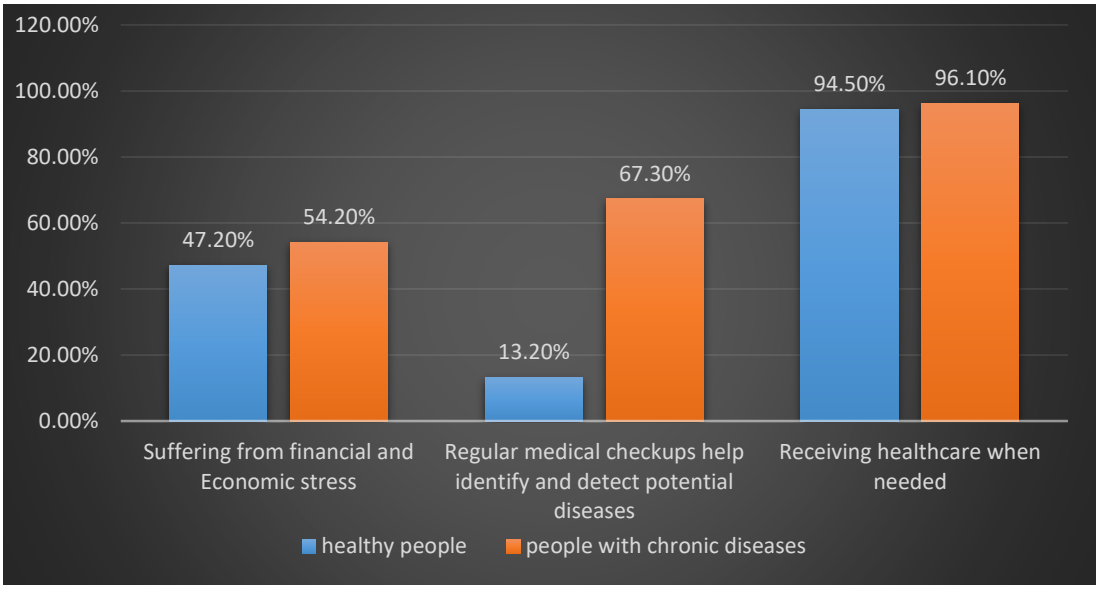

Figure 4.7: health and economic factors.

4.2 Logistic regression model analysis

In this section, the results of the logistic regression analysis of the effect of independent variables represented by genetic and environmental factors on chronic diseases are explained, as the incidence of chronic disease is considered a binary dependent variable, so the least squares method is not suitable for its application , and The model parameters are estimated using the maximum likelihood method, and this regression ranks the independent variables according to their effect on the occurrence of the dependent variable, The binary dependent variable takes the value (1) when the event occurs with probability (p), and takes the value (0) when the event does not occur with probability (1-p).

In this study, the logistic regression function will be estimated, as it consists of a binary variable that represents the incidence of a chronic disease and takes the value (1), or the absence of a chronic disease and takes the value (0). The table (4.1) shows the independent variables and the values they take:

Table 4.1: Variables and their values.

Factors	Value (1)	Value (0)
Gender	M	F
Age group	Elderly	others
Living in an urban or rural area	Rural	Urban
Genetic counseling or testing	Yes	No
One or both parents having chronic diseases	Yes	No
Brothers or sisters having chronic diseases	Yes	No
specific pattern of disease in A family or extended family	Yes	No
Following a healthy and balanced diet	Yes	No
Committing to exercise and exposure to fresh air	Yes	No
Smoking	Yes	No
Addiction to alcohol or drugs	Yes	No
Being in areas where smoking occurs(passive smoking)	Yes	No
Exposure to psychological and social stress	Yes	No
Suffering from financial and Economic stress	Yes	No
Exposure to oppression by the occupation or its punitive measures	Yes	No
Exposure to chemicals at work	Yes	No
The presence of factories or sources of pollution nearby	Yes	No
Regular medical checkups help identify and detect potential diseases	Yes	No
Receiving healthcare when needed	yes	No

4.3 Logistic regression model analysis result

The attached table below shows the results of a binary logistic regression analysis to estimate the independent variables affecting the incidence of chronic diseases in Palestine using SPSS. The factors were coded from X1 to X19, as previously assumed. Some

independent variables are statistically significant at a confidence level of 0.05, including: age, the presence of siblings with chronic diseases, a specific pattern of chronic diseases in the family, regular exercise and exposure to outdoor air, suffering from psychological and social stress, alcohol or drug addiction, and undergoing periodic medical checkups. The remaining independent variables are not statistically significant at a confidence level of 0.05

Table 4.2: Data analysis results for binary logistic regression

Factors	Odd ratio	P-value	Lower confidence interval	Upper confidence interval
X1	0.591	0.154	1.396	5.471
X2	2.763	0.004	1.287	3.217
X3	2.288	0.063	0,955	5.480
X4	1.918	0.079	0.927	3.967
X5	3.732	< 0.001	1.979	7.037
X6	2.22	0.018	1.149	4.307
X7	0.707	0.357	0.337	1.480
X8	0.958	0.844	0.623	1.472
X9	0.482	0.020	0.261	0.892
X10	1.457	0.045	1.009	2.105
X11	1.633	0.158	0.827	3.226
X12	1.054	0.913	0.408	2.721
X13	1.477	0.342	0.661	3.302
X14	0.606	0.147	0.308	1.193
X15	11.718	0.007	1.946	70.554
X16	1.065	0.877	0.482	2.355
X17	1.364	0.347	0.714	2.605
X18	12.935	< 0.001	6.8333	24.483
X19	3.660	0.97	0.791	16.9222

The seven factors of influence and significance are as follows, along with the reasons for their importance:

Age group(X2): The results of the model indicate that age group plays a prominent role in the incidence of chronic disease, as the odds ratio reached 2.763, which means that the occurrence of a chronic disease at one age group is 2.763 times more than at another (with other variables held constant). This is explained by the fact that the age group was divided into three categories, and the percentage of those with chronic diseases in the elderly group was greater than the youth group, which is also greater than the children group in terms of the percentage of those with chronic diseases, meaning that with age, the chances of developing chronic diseases increase, The same result was obtained in a study on the predictive value of sociodemographic factors in chronic diseases, conducted in the United States, where data was collected from 372,050 people. Age was found to be the most predictive and influential factor in health conditions, meaning that the relationship between age and disease incidence is directly proportional (Kunnath *et al*, 2024).

Having brothers or sisters with chronic diseases(X5): The results of the model indicate that having brothers or sisters with chronic diseases increases the chances of developing chronic diseases to a greater extent, as the odds ratio reached 3.732, meaning that people who have brothers or sisters with chronic diseases are 3.732 times more likely to develop chronic diseases than those who do not have brothers or sisters who are not infected (with other variables held constant). A US study that also examined the effect of siblings having cardiovascular disease on the risk of developing found that the odds ratio for people with affected siblings was 1.45 after adjusting for risk factors .The study also showed that this ratio had a greater impact than the parental risk of developing these diseases (Morabito *et al.*, 2005).

The presence of a certain pattern of chronic diseases in the family(X6): The results of the model indicate that the presence of a certain pattern of diseases in a person's family increases the chance of developing diseases, as the odds ratio reached 2.225, meaning that people who have a certain pattern of diseases in their family are 2.225 times more likely to

develop chronic diseases than people who do not have a certain pattern of chronic diseases in their family, (with other variables held constant). a study on the risk of developing hypertension and its transmission across generations in the United States also found that the presence of the disease in the extended family had an effect, with affected grandparents having a 1.33 odds ratio for their grandchildren to develop hypertension (Niiranen *et al,* 2017).

Regular exercise and exposure to fresh air(X9): The results of the model indicate that regular exercise and exposure to fresh air play a prominent role in reducing the occurrence of chronic disease in a person, as the odds ratio reached 0.482, meaning that people who exercise and are exposed to fresh air are 0.518 times less likely to develop chronic diseases than people who do not exercise and are not exposed to fresh air, (with other variables held constant), A European study on the relationship between physical activity and chronic diseases, which included 30,826 people, also showed that those who were physically active were less likely to develop heart and respiratory diseases, diabetes, and obesity (Marques *et al,* 2017).

Suffering from psychological and social stress(X10): The results of the model indicate that suffering from psychological and social stress increases the chances of developing a chronic disease, as the odds ratio reached 1.457, which means that people who suffer from these stresses are 1.457 times more likely to develop chronic diseases than other people who do not suffer from psychological and social stress, (with other variables held constant), In a study on the relationship between psychological and social factors and the incidence of chronic diseases, a sample of 11,637 people in Australia was surveyed. Psychological stress was found to have a significant impact on the incidence of chronic diseases, with the odds ratio for heart disease being 2.3 for those experiencing psychological stress. Social support also slightly reduced the risk of chronic diseases (Sahle *et al,* 2020).

Alcohol or drug addiction(X15): The results of the model indicate that people addicted to drugs or alcohol are more likely to develop chronic diseases, as the odds ratio reached 11.718, which means that people addicted to drugs or alcohol are 11.718 times more likely

to develop chronic diseases than people who are not addicted to drugs or alcohol (with other variables held constant). a study conducted in Brazil on 1,781 people showed that alcohol and drug use influence the incidence of chronic diseases ،with the incidence rate increasing with the rate of individual consumption .The World Health Organization also indicated that alcohol causes 2.6 million deaths annually ،while drugs cause 0.6 million deaths annually (Bird ،2025).

Periodic medical examinations(X18): The results of the model indicate that regular medical examinations play a prominent role in the emergence of chronic diseases, as the odds ratio reached 12.935, meaning that people who undergo medical examinations are 12.935 times more likely to develop chronic diseases than people who do not undergo medical examinations (with other variables held constant), A study on the effect of regular screening for cardiovascular risk factors in Korea showed that people who undergo regular checkups have a lower risk of developing the disease than those who do not (Park *et al*., 2020).

This means that those who undergo medical screenings are more likely to detect diseases earlier, or that those with chronic diseases are more likely to undergo more frequent screenings to monitor their health and detect any other conditions that may be complications of the primary chronic disease. Furthermore, there is a noticeable decline in the uptake of medical screenings in the community, especially among the elderly and young, as most people only undergo screening when symptoms of a chronic disease appear.

Based on the previous results for significant and influential factors, the odds ratio values were as shown in (figure 4.8)

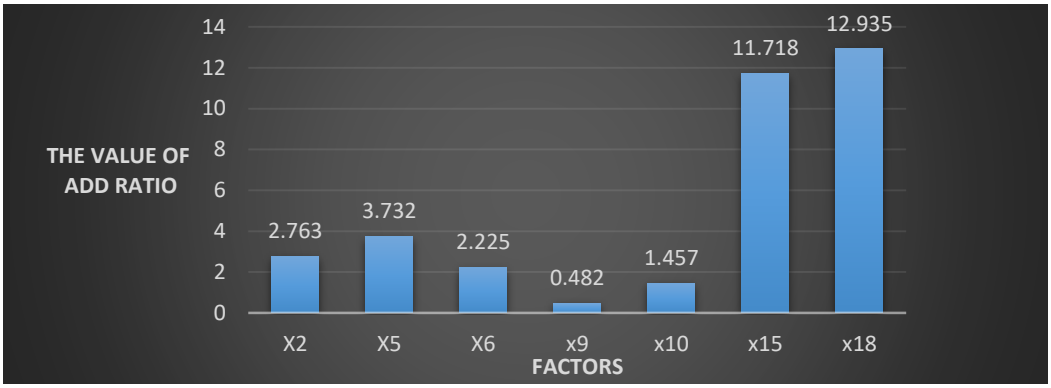

Figure 4.1: the statistically significant factors.

4.4 prediction equation

It is one of the outputs of binary logistic regression. This equation predicts the occurrence of an event with a binary answer, this equation is based on the logistic regression coefficients and the values of the independent variables or factors. The outcome of the equation is the answer to the binary variable, which is the occurrence of the disease.

After analyzing the binary logistic regression model on the program Spss, the statistically significant factors were identified, and the values of the regression coefficients were determined, which allows writing the prediction equation as follows:

$$\text{Log}\left(\frac{p}{1-p}\right) = -5.284 + 1.016\,X_2 + 1.317\,X_5 + 0.800\,X_6 - 0.729\,X_9 + 0.377\,X_{10} + 2.461X_{15} + 2.560\,X_{18}$$

The equation for predicting the incidence of chronic diseases can be written in the following common form after neglecting the independent variables that are not statistically significant as follows:

$$\text{P(Y)} = \frac{1}{1+e^{-\left(-5.284 + 1.016\,X_2 + 1.317\,X_5 + 0.800\,X_6 - 0.729\,X_9 + 0.377\,X_{10} + 2.461\,X_{15} + 2.560\,X_{18}\right)}}$$

According to the previous prediction equation of the model, if an elderly person has brothers or sisters with chronic diseases, does not have a family pattern of the disease, does not practice physical activity regularly, does not have psychological and social stress, does not drink alcohol, and undergoes regular checkups, then the probability of having a chronic disease is:

$$P(Y) = \frac{1}{1+e^{(-5.284 + 1.016(1) + 1.317(1))}} = 0.9503$$

According to the previous prediction equation of the model, if an elderly person has a family pattern of the disease, does not have brothers or sisters with chronic diseases, does not practice physical activity regularly, does not have psychological and social stress, does not drink alcohol, and undergoes regular checkups, then the probability of having a chronic disease is:

$$P(Y) = \frac{1}{1+e^{(-5.284 + 1.016(1) + 0.8(1))}} = 0.9697$$

According to the previous prediction equation of the model, if an elderly person has a family pattern of the disease, does not have a family pattern of the disease, does not have psychological and social stress, does not drink alcohol, and undergoes regular checkups, then the probability of having a chronic disease is:

$$P(Y) = \frac{1}{1+e^{(-5.284 + 1.016(1) + 0.729(1))}} = 0.9718$$

According to the previous prediction equation of the model, if an elderly person Psychological and social pressures, has no brothers or sisters with chronic diseases, has no family pattern of the disease, does not exercise regularly, does not drink alcohol, and undergoes regular checkups, the probability of having a chronic disease is:

$$P(Y) = \frac{1}{1+e^{(-5.284 + 1.016(1) + 0.377(1))}} = 0.9799$$

According to the previous prediction equation for the model, if an elderly person drinks alcohol, has no brothers or sisters with chronic diseases, has no family pattern of the disease, does not exercise regularly and does not have psychological and social pressures, and undergoes regular checkups, the probability of having a chronic disease is:

$$P(Y) = \frac{1}{1+e^{(-5.284 + 1.016(1) + 2.461(1))}} = 0.8591$$

According to the previous prediction equation for the model, if an elderly person, undergoes regular checkups, has no brothers or sisters with chronic diseases, has no family pattern of the disease, does not exercise regularly and does not have psychological and social pressures, and does not drink alcohol, the probability of having a chronic disease is:

$$P(Y) = \frac{1}{1+e^{(-5.284 + 1.016(1) + 2.56(1))}} = 0.8465$$

Chapter Five: Discussion

This section of the thesis includes a discussion of the results of evaluation tests to verify the model's suitability, and finally, a conclusion and recommendations for future work.

There are several tests to diagnose the suitability of the logistic regression model in estimating the factors influencing the onset of chronic diseases. The tests mentioned in Chapter Three of the thesis, which examine the model to determine its suitability, were conducted as follows:

5.1 Goodness-of-fit test chi-square and Hosmer-lemeshow test

Through chi-square and Hosmer-lemeshow test on the SPSS program, the following results were obtained:

Table 5.1: Hosmer-lemeshow test.

Sample size	chi-square	Sig. of Hosmer-lemeshow test
387	14.094	0.079

The chi-square value was 14.094, and the sig. value in the Hosmer-Lemnshaw test was 0.079. Since it is greater than 0.05, this means that the model is a good fit to the data, and there is no significant difference between the actual and predicted values, which reflects its suitability.

5.2 Coefficient of determination (Cox and snell (R_{CS}^2) , Negelkerke (R_N^2))

By analyzing the regression model in SPSS, the following results were obtained for the values of the coefficients of determination:

Table 5.2: moodle summary Which contains Coefficient of determination.

-2 loglihood	Cox and snell R_{CS}^{2}	Negelkerke R_{N}^{2}
301.355	0.431	0.583

The value of Cox and Snell, which is equal to 0.431, means that the model explains about 43.1% of the dependent variable, and the value of Nagelkerke, which is equal to 0.583, means that the model explains about 58.3% of the dependent variable, which means that the model is good and appropriate.

5.3 Table Classification Efficiency Test

The classification table test is one of the methods of diagnosis and examination of the efficiency and suitability of the model, as the overall evaluation percentage reached 80.4, which indicates the accuracy of the model in classifying the data correctly, as the model can classify 80.4% of the cases correctly into the target categories, and the overall evaluation percentage can be calculated by dividing the number of correct predictions by the total number of the study sample.

Table 5.3: table of classification.

		predicted	
Observed	Healthy people	People with chronic diseases	Percentage correct
Healthy people	203	31	0.868
People with chronic diseases	45	108	0.706
Overall percentage			0.804

5.4 wald test

The independent variables that were statistically significant at a confidence level of 0.05 were: age group, with a Wald value of 8.507; having brothers or sisters with chronic diseases, with a Wald value of 16.554; having a pattern of chronic diseases in the family, with a Wald value of 5.632; regular exercise and exposure to fresh air, with a Wald value of 5.409; suffering from psychological and social stress, with a Wald value of 4.025; addiction to alcohol or drugs, with a Wald value of 7.220; and regular medical checkups, with a Wald value of 61.827. The null hypothesis is assumed for these variables and the alternative hypothesis is accepted, which means that these variables have a significant effect on the dependent variable. The higher the Wald value, the greater the influence and role of the independent variable on the dependent variable. The remaining independent variables are not statistically significant at a confidence level of 0.05, so we accept the null hypothesis and reject the alternative. This means that the remaining variables have no significant influence on the dependent variable. The variables can be ranked according to the strength of influence as follows:

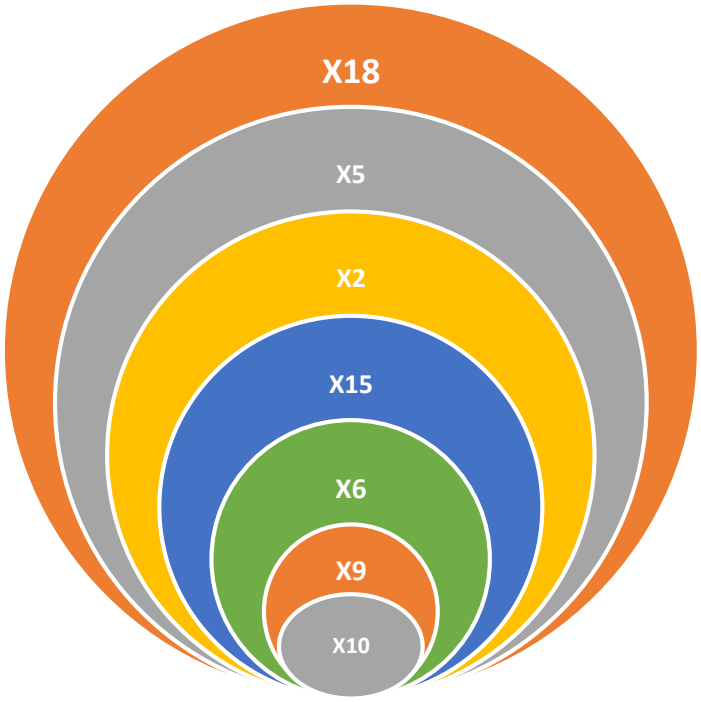

Figure 5 1: Variables ranked according to their strength and influence.

5.5 multicolinearity analysis test

Through this test and its application on the SPSS program, all values of the variance inflation factor (VIF) were between 1 and 2, where the lowest value for one of the variables was 1.050 and the highest value for another variable was 1.457, which means that there is no problem with autocorrelation, which means that there is no autocorrelation or that it is very small between the independent variables and that there is no significant overlap between them, so we can say that the model is stable because the estimation of the coefficients of the independent variables is not distorted by the overlap between the variables.

5.6 likelihood-ratio test

Table 5.4: table of likelihood-ratio with constant limit.

-2 log likelihood	Coefficients of constants
519.420	- 0.419
519.416	- 0.425
519.416	- 0.425

Table 5.5: table of likelihood-ratio after adding the variables.

	Chi-square	df	Sig
Step	218.062	20	< 0.001
Block	218.062	20	< 0.001
Modle	218.062	20	< 0.001

In this model that contains only the fixed term, in the seventh table, the lowest value of the maximum likelihood function was reached and the value of (-2likelihood) was equal to (519.416) at the third iteration when the change value was less than 0.001, From the analysis of the model, the value of (x^2) was equal to (171.019) through the overall statistics box.

After adding the independent variables to the initial model with the fixed term, this led to an improvement in the model as the value of (-2likelihood) became, as shown in the 7 table, equal to (301.355) as it was (519.416) and the difference between them is equal to (218.061), which represents the value of (x^2) as shown in the 2 table, and this means that there has become a significant improvement and statistical significance at a confidence level of 0.05, which leads to the model that contains the independent variables being much better than the one that contains only the fixed term.

The model improved the classification as the percentage of correct classification according to the table (80.4%) became Through these tests that were conducted for the logistic regression model, it was found that it is appropriate and suitable for predicting the probability value of the dependent variable, which is the occurrence of chronic disease, through the independent variables, which are genetic and environmental factors and the individual's lifestyle.

5.7 Confidence interval

From Table (4.2), it is clear that the confidence intervals for the likelihood ratio do not include the number 1 for the significant factors, which means that they are significant according to the confidence intervals. Also, from the results of the analysis on the SPSS program, it is also clear that the confidence intervals for the regression coefficients do not include zero, which means that these seven factors are significant and influential.

5.8 Conclusions

In this study, the main objective was to verify the relationship and influence between genetic, environmental, and lifestyle factors on the occurrence of human chronic diseases, and to study the possibility of predicting their occurrence using binary logistic regression. The study demonstrated the influence and contribution of some factors to the occurrence of diseases at a confidence level of 0.05, namely the age group, where the older a person is, the greater the likelihood of infection. It also demonstrated the role of genetic factors, as it was

shown that the presence of brothers or sisters with chronic diseases, and the presence of a specific pattern of diseases in the family, increases the likelihood of infection. In addition, it was shown that environmental factors play a role, as exposure to psychological and social stress increases the likelihood of infection. A person's lifestyle also plays a role, as regular exercise, exposure to fresh air, periodic medical examinations, and addiction to alcohol or drugs all lead to an increased likelihood of developing chronic diseases. As for the remaining factors, it was found that they were not statistically significant at a confidence level of 0.05. The statistically significant factors were ranked according to their strength and influence on the occurrence of chronic diseases, and then a prediction equation based on these factors was formulated after analyzing the data. Using Spss program, which in turn analyzed the binary logistic regression model.

The data were verified to be free of multicollinearity, and the suitability, accuracy, and predictive ability of the logistic regression model were also verified using several tests and methods. The logistic regression model predicted the probability of developing chronic diseases at 80.4%, demonstrating the importance of integrating various factors into predictive models to prevent diseases and limit their spread as much as possible.

The study achieved its objectives, but there are some issues to consider, including the sample size, the number of different factors, and a specialized study of the disease itself. Increasing the sample size and the number of factors used to study a specific disease will provide more data, leading to more accurate and reliable results. In addition, studying a specific chronic disease itself will provide detailed information about it, thus obtaining more accurate, useful and beneficial results than studying chronic diseases in general. Therefore, the study recommends studying a specific disease itself, increasing the number of independent variables and increasing the sample size. It also recommends using logistic regression in future studies and on a wider scale in fields other than medicine, due to its predictive ability. It is also recommended to conduct studies using other statistical methods and compare them with the results of logistic regression to obtain the best results. It is hoped that the study will contribute to understanding the relationship between genetic,

environmental and lifestyle factors in chronic diseases, which may contribute to the development of effective measures for prevention and treatment.

5.9 Future study

This study recommends that future researchers increase the size of the targeted sample to obtain the best possible results, and distribute the sample to include all Palestinian regions, including cities and countryside, to achieve the best results that can be generalized to the community. It also recommends studying something specific to itself and not something general, such as studying only a specific disease, in addition to increasing the number of independent variables that are likely to have an impact on the dependent variable, to achieve accurate and better results. In addition, it recommends studying according to long-term designs, i.e., following sample individuals for a period of time to identify risk factors associated with the occurrence of diseases, which enhances causal inference rather than a random association. Finally, it also recommends using various modeling methods, other predictive models in addition to logistic regression, and machine learning methods, and comparing all of these methods together, to obtain accurate results.

References

Abdel Khaleq, A. (2020). Using the logistic regression model to determine symptoms of COVID-19 infection. *Trade and Finance, 41*(2), 123–145.

Abdullah, A., & Abdel Qader, Q. (2021). Application of logistic regression in studying and analyzing factors affecting the choice of enterprise capital. *Dirasat Iqtisadiyah [Economic Studies], 15*(3), 272–287. https://asjp.cerist.dz/en/article/165877

Al-Khazaali, T. A. M. (2021). *Estimation of a logistic regression model for a multilevel dependent variable with practical application* [Master's thesis, University of Karbala]. University of Karbala, College of Administration and Economics, Department of Statistics. https://drive.google.com/file/d/1g9DZWORh9NTl1DpqC-KZZBqbio0YDIj1/view?usp=sharing

Al-Nasrawi, N. (2017). *Using the bootstrap method in analyzing parametric and semi-parametric models and comparing between them* [Master's thesis, University of Karbala]. https://uokerbala.edu.iq/wp-content/uploads/2020/05/Rp-Use-The-Bootstrap-in-Parametric-%E2%80%8EModels-and-Semi-Parametric-%E2%80%8E-Analysis-and-Comparison-Between-%E2%80%8EThem.pdf

Abdi, M. S., & Zenadi, Z. (2021). Logistic regression model of the factors affecting the estimation of creative thinking among workers of a sample of local institutions. *Journal of Creativity*, 11(2), 285–304. https://search.emarefa.net/detail/BIM-1298258

Alexopoulos, E. C. (2010). Introduction to multivariate regression analysis. *Hippokratia, 14*(1), 23–28.

American Lung Association. (2020). *State of the air 2020*. American Lung Association. https://www.lung.org/getmedia/13cf12b8-060b-4a2c-8dd1-dd8af5c16b2f/State-of-the-Air-2020.pdf. American Lung Association

Babtain, A. A. H. (2009). Logistic regression and how to use it in building prediction models for data with two-valued dependent variables (Unpublished doctoral dissertation). College of Education, Umm Al-Qura University, Saudi Arabia.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, *39*(227), 357–365. https://doi.org/10.1080/01621459.1944.10500699.

Booth, F. W., & Lees, S. J. (2007). Fundamental questions about genes, inactivity, and chronic diseases. *Physiological Genomics, 28*(2), 146–157.

Centers for Disease Control and Prevention. (2021). Genetics and chronic disease: How family history can influence disease risk.

Chen, M.-Y. (2011). Predicting corporate financial distress based on integration of decision tree classification and logistic regression. *Expert Systems with Applications, 38*(9), 11261–11272. https://doi.org/10.1016/j.eswa.2011.02.173

Cui, J., Malspeis, S., Choi, M. Y., Lu, B., Sparks, J. A., Yoshida, K., & Costenbader, K. H. (2023). Risk prediction models for incident systemic lupus erythematosus among women in the Nurses' Health Study cohorts using genetics, family history, and lifestyle and environmental factors. *Seminars in Arthritis and Rheumatism, 58*, 152143. https://doi.org/10.1016/j.semarthrit.2022.152143

Dobson, A., & Barnett, A. (2018). An introduction to generalized linear models (4th ed.). Chapman & Hall/CRC.

Eliason, S. R. (1993). Maximum likelihood estimation: Logic and practice. Sage.

El-Habil, A. M. (2012). An application on multinomial logistic regression model. *Pakistan Journal of Statistics and Operation Research, 8*(2), 271–291.

Franzago, M., Santurbano, D., Vitacolonna, E., & Stuppia, L. (2020). Genes and diet in the prevention of chronic diseases in future generations. *International Journal of Molecular Sciences, 21*(7), 2633. https://doi.org/10.3390/ijms21072633

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland, 15*, 246–263.

Garson, G. D. (2006). Logistic regression. North Carolina State University, PA 765: Quantitative Research in Public Administration.

Garson, G. D. (2014). Logistic regression: Binary and multinomial. Statistical Associates Publishing.

Ghosh, A., & Basu, A. (2016). Robust estimation in generalized linear models: The density power divergence approach. *Test, 25*, 269–290.

Griffin, T. (2025, April 10). Just 8 drinks a week increase your chance of brain lesions by 133% — and up your odds of a key sign of Alzheimer's. *New York Post*. https://nypost.com/2025/04/10/health/eight-drinks-a-week-increase-chance-of-brain-lesions-by-133/

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill Education.

Hosmer, D. W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics — Theory and Methods, 9*(10), 1043–1069.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Huang, Z., Cao, C., Xu, M., & Yang, X. (2023). Impact of environmental exposure on chronic diseases in China and assessment of population health vulnerability. *ISPRS International Journal of Geo-Information, 12*(4), Article 155.

Ibrahim, W., & Taher, H. (2019). Using the logistic regression model to identify the most important factors affecting lung cancer in Iraq for the year 2017. *Journal of Administration and Economics, 1*(121), 283–309.

Iftikhar Mnahi, K., & Saifaldin, H. (2018). Financial indicators affecting the performance of stocks in Iraqi market of securities for 2016 using the binary logistic regression model. *Al-Dananeer Journal, 1*(1), 167–189.

Kaprio, J., & Koskenvuo, M. (2002). Genetic and environmental factors in complex diseases: The Older Finnish Twin Cohort. *Twin Research and Human Genetics, 5*(5), 358–365.

Kelishadi, R., & Poursafa, P. (2014). A review on the genetic, environmental, and lifestyle aspects of the early-life origins of cardiovascular disease. *Current Problems in Pediatric and Adolescent Health Care, 44*(3), 54–72.

Khera, A., McGuire, D. K., & Reilly, M. P. (2005). Sibling cardiovascular disease as a risk factor for cardiovascular events in middle-aged adults. *JAMA, 293*(3), 358–363. https://doi.org/10.1001/jama.293.3.358

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression: A self-learning text* (3rd ed.). Springer.

Komarek, P. (2004). *Logistic regression for data mining and high-dimensional classification* (Tech. Rep. CMU-RI-TR-04-34). Robotics Institute, Carnegie Mellon University. https://www.ri.cmu.edu/publications/logistic-regression-for-data-mining-and-high-dimensional-classification/. Robotics Institute CMU

Kunnath, A. J., Sack, D. E., & Wilkins, C. H. (2024). Relative predictive value of sociodemographic factors for chronic diseases among All of Us participants: A descriptive analysis. *BMC Public Health, 24*(1), Article 405. https://doi.org/10.1186/s12889-024-17834-1

Liu, K., Cao, H., Guo, C., Pan, L., Cui, Z., Sun, J., Zhao, W., Han, X., Zhang, H., Wang, Z., Niu, K., Tang, N., Shan, G., & Zhang, L. (2021). Environmental and genetic determinants of major chronic disease in Beijing–Tianjin–Hebei region: Protocol for a

community-based cohort study. *Frontiers in Public Health, 9*, Article 659701. https://doi.org/10.3389/fpubh.2021.659701

Loktionov, A. (2003). Common gene polymorphisms and nutrition: Emerging links with pathogenesis of multifactorial chronic diseases (Review). *The Journal of Nutritional Biochemistry, 14*(8), 426–451.

Marques, A., Santos, T., Martins, J., de Matos, M. G., & González Valeiro, M. (2018). The association between physical activity and chronic diseases in European adults. *European Journal of Sport Science, 18*(1), 140–149. https://doi.org/10.1080/17461391.2017.1400109

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.) [Quantitative Applications in the Social Sciences]. Sage.

Meng, L., Maskarinec, G., Lee, J., & Kolonel, L. (1999). Lifestyle factors and chronic diseases: Application of a composite risk index. *Preventive Medicine, 29*(4), 296–304.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Wiley.

Nashef, S. A. M., Roques, F., Michel, P., Gauducheau, E., Lemeshow, S., & Salamon, R. (1999). European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardio-Thoracic Surgery, 16*(1), 9–13.

National Institute of Mental Health. (2019). *Chronic illness and mental health* (Understanding the link between chronic disease and mental health). https://www.nimh.nih.gov/health/publications/chronic-illness-mental-health. National Institute of Mental Health

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673–690.

Ordovas, J. M., & Shen, J. (2008). Gene–environment interactions and susceptibility to metabolic syndrome and other chronic diseases. *Journal of Periodontology, 79*(8), 1508–1513.

Park, B. H., Lee, B. K., Ahn, J., Kim, N. S., Park, J., & Kim, Y. (2021). Association of participation in health check-ups with risk factors for cardiovascular diseases. *Journal of Korean Medical Science, 36*(3), e19. https://doi.org/10.3346/jkms.2021.36.e19

Paul, P., Pennell, M. L., & Lemeshow, S. (2013). Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. *Statistics in Medicine, 32*(1), 67–80.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis.

Journal of Clinical Epidemiology, 49(12), 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Ravikumar, P., Martin, J., & Lafferty, J. (2010). High-dimensional Ising model selection using ℓ1-regularized logistic regression. *The Annals of Statistics, 38*(3), 1287–1319. https://doi.org/10.1214/09-AOS691

Ramadan, A. N. (2023). Chi-square test for independence and subsequent pairwise comparisons for analyzing double tables using the Statistical Package for the Social Sciences (SPSS). *Journal of Special Education and Inclusive Education Research, 1*(2), 1–33.

Sari, M., & Daish, M. A. (2017). Logistic regression model: Concept, characteristics, applications. With an example of a binary logistic regression in SPSS. *Algerian Journal of Education and Society Issues, 1*(1), 124–132.

Sahle, B. W., Chen, W., Melaku, Y. A., Akombi, B. J., Rawal, L. B., & Renzaho, A. M. N. (2020). Association of psychosocial factors with risk of chronic diseases: A nationwide longitudinal study. *American Journal of Preventive Medicine, 58*(2), e39–e50. https://doi.org/10.1016/j.amepre.2019.09.007

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry, 19*(3), 227–229.

Walker, J. (1996). *Methodology application: Logistic regression using the CODES data* (Report No. HS-808 460). National Highway Traffic Safety Administration. https://rosap.ntl.bts.gov/view/dot/4427/dot_4427_DS1.pdf. ROSA P

Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques*. Sage.

Wehby, G. L., Domingue, B. W., & Wolinsky, F. D. (2018). Genetic risks for chronic conditions: Implications for long-term wellbeing. *The Journals of Gerontology: Series A, 73*(4), 477–483.

World Health Organization. (2018). *Noncommunicable diseases (fact sheet)*. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. World Health Organization

World Health Organization. (2023). *Noncommunicable diseases (fact sheet)*. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. World Health Organization

Yoon, P. W., Scheuner, M. T., & Khoury, M. J. (2003). Research priorities for evaluating family history in the prevention of common chronic diseases. *American Journal of Preventive Medicine, 24*(2), 128–135.

عنوان الرسالة: تأثير العوامل الوراثية والبيئية على حدوث الأمراض المزمنة والتنبؤ بها.

اسم الطالب: يوسف حسن حسين وشاحي

أعضاء لجنة المناقشة:

د. صالح عفانة

د. عبد الحليم زيقان

د. محمد دوابشة

ملخص

تُعدُّ الأمراضُ المزمنةُ من أكبر التحديات العالمية وأكثر أسباب الوفاة شيوعًا؛ فهي لا تحدث في فراغٍ، بل ثمة عواملُ قد تُساهمُ في حدوثها. تهدف هذه الدراسةُ إلى دراسةِ تأثيرِ العوامل الوراثية والبيئية ونمطِ الحياة على انتشارِ الأمراض المزمنةِ والتنبؤِ بها باستخدامِ الانحدارِ اللوجستي الثنائي. أُجريت الدراسةُ على عيّنةٍ عشوائيةٍ مكوّنةٍ من 387 فردًا من محافظةِ جنين، وَتَمَّ تحليلُ بياناتهم باستخدامِ البرنامج الإحصائي، حيث أظهر التحليلُ تأثيرَ سبعةِ عواملٍ عند مستوى دلالة 0.05 وهي: العمر، ووجود أشقاء مصابين، والنمطُ العائليُّ للمرض، وممارسةُ الرياضةِ بانتظام، والتعرُّضُ للهواء النقي، والضغوطُ النفسيةُ والاجتماعيةُ، وإدمانُ الكحول أو المخدرات، وإجراءُ الفحوصاتِ الدورية. بلغت دقةُ التنبؤِ للنموذج 80.4%. تُسلِّطُ الدراسةُ الضوءَ على أهميةِ هذه العوامل ودورها في الوقايةِ من الأمراض والحدِّ من انتشارها قدرَ الإمكان من خلالَ التوعيةِ العامةِ وتبنّي نمطِ حياةٍ صحيّ.

الكلمات المفتاحية: الانحدار اللوجستي، طريقة تقدير الاحتمال الأقصى (MLE)، الأمراض المزمنة، العوامل الوراثية والبيئية.