UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Obesity Genetic Risk Score

Catarina Isabel Nogueira Ribeiro

Mestrado em Bioestatística

Dissertação orientada por:
Prof.ª Doutora Lisete Sousa
Prof. Doutor Miguel Brito

2020

É coisa preciosa, a saúde, e a única, em verdade, que merece que em sua procura empreguemos não apenas o tempo, o suor, a pena, os bens, mas até a própria vida; Tanto mais que sem ela a vida acaba por tornar-se penosa e injusta.

Michel de Montaigne

Resumo

A obesidade é a doença metabólica humana mais comum e mais antiga que foi registada até aos dias de hoje. Desde a pré-história que a obesidade assumiu um papel preponderante na vida do ser humano, sendo referida como símbolo de fertilidade e beleza. Remetendo ao Período Neolítico (cerca de 10.000 A.C.), as "deusas", isto é, as mulheres com características como seios volumosos e coxas bem definidas já eram admiradas neste período. Contudo, nesta época, o ser humano tinha grande dificuldade em obter comida e conseguir *stock* da mesma. Portanto, a natureza foi encarregada de fornecer ao corpo humano um mecanismo para armazenar energia. Esse mecanismo consistia em incentivar o Homem, através da fome, a ingerir uma grande quantidade de calorias e fazer com que seu organismo transformasse o excesso em gordura, armazenando-o por períodos de falta de comida. As sementes, raízes e frutos eram os principais alimentos ingeridos pelo Homem e foi para esse padrão alimentar que a genética preparou o organismo herdado por nós. O problema é que nosso estilo de vida é completamente diferente do estilo de vida levado pelo Homem no Período Neolítico. Atualmente, os alimentos estão facilmente disponíveis nas sociedades modernas e, por outro lado, as mudanças no nosso ambiente ocorrem mais rapidamente do que as modificações no contexto genético. Desta forma, ao considerar o desequilíbrio do nosso estilo de vida moderno e do nosso perfil genético "antigo", é compreensível que muitas pessoas ganhem peso com tanta facilidade. Embora fazendo este paralelismo entre o passado e a atualidade a obesidade não é o simples resultado de indisciplina pela qual o indivíduo ingere uma quantidade excessiva de alimentos ou o facto de não fazer atividade física suficiente. Muitos indivíduos são mais suscetíveis do que outros a aumentarem de peso, ou desenvolverem obesidade, devido aos próprios genes. Na maioria dos casos, os genes envolvidos no aumento de peso aumentam o risco ou a suscetibilidade de um indivíduo para desenvolver a obesidade, quando exposto a fatores ambientais adversos. Em casos raros, a ação direta de certos genes pode causar diretamente aumento de peso ou obesidade.

Assim, podemos afirmar que a obesidade é uma doença crónica grave associada ao excesso de gordura corporal, na medida em que pode ter um efeito negativo na saúde. A escolha de estilos de vida pouco saudáveis ou fatores ambientais contribuem para o aparecimento desta doença que é conhecida por ser hereditária e altamente poligénica. Milhões de variações subtis na sequência de DNA humano, ou genoma, são a chave para uma série de condições, do cancro de mama às doenças cardíacas. Este estudo de caso-controle de associação genética compara a frequência de alelos ou genótipos nos *loci* dos marcadores genéticos, isto é, polimorfismos de nucleotídeo único (SNPs), numa amostra de indivíduos, com e sem uma determinada característica de doença, de uma determinada população. Os sucessos recentes nas descobertas de polimorfismos de nucleotídeo único (SNPs) potencialmente causais para doenças complexas são bastante promissores. Curiosamente, nos dias de hoje, várias empresas oferecem, por taxas relativamente modestas, serviços genómicos personalizados que fornecem estimativas

individualizadas de risco de doença com base na genotipagem do SNP em todo o genoma. A maioria das empresas que oferecem esse perfil deixa claro que não é um serviço clínico e que seus cálculos não se destinam a fins de diagnóstico ou prognóstico. Apenas aconselham os seus clientes a consultar o seu médico para obter mais informações.

Foi recolhida uma amostra de 212 mulheres caucasianas, composta por 112 mulheres obesas e um grupo de controle (peso normal) de 100 mulheres. Para os dois grupos, foram registados o peso corporal total, o índice de massa corporal (IMC), a circunferência da cintura e do quadril, a relação cintura-quadril e a gordura corporal. Além disso, foi também recolhida informação correspondente a 19 SNPs relacionados com a obesidade em 13 genes, também estes relacionados com a doença, para ambos os grupos de mulheres.

Este tipo de estudo de caso-controlo considera métodos e técnicas básicas de análise estatística e tem como objetivo determinar se existe associação entre a característica da doença, isto é, a obesidade e o marcador genético. Um pressuposto fundamental deste tipo de estudo é que os indivíduos selecionados nos grupos de caso/controlo forneçam estimativas imparciais da frequência do alelo. Caso contrário, os resultados encontrados da associação refletirão apenas vieses resultantes do desenho do estudo. Os modelos dominante e recessivo, para cada SNP, são exemplos que são estudados através de tabelas de contingência e nos quais se quer encontrar alguma associação estatística entre a doença e o respetivo modelo, neste trabalho de investigação. Num modelo recessivo são necessárias duas cópias do alelo A para que o risco da doença aumente. Assim, a tabela de contingência, isto é, uma tabela 2×2 é composta pelas observações dos genótipos de aa versus as observações dos genótipos Aa e AA. Enquanto que um modelo dominante, cujo número de cópias do alelo A aumenta o risco de doença, a tabela de contingência pode ser resumida como uma tabela 2×2 da contagem de genótipos de AA versus Aa e aa combinado.

Perfis multi-*locus* de risco genético, os chamados "*scores* de risco genético", podem ser usados para traduzir descobertas de estudos de associação, em todo o genoma, sendo ferramentas para pesquisa da saúde da população. Portanto, os principais objetivos deste trabalho são identificar polimorfismos associados à obesidade em mulheres portuguesas, identificar o *score* de risco genético da obesidade e associar polimorfismos genéticos a traços relacionados à obesidade, usando o *software Excel*, *SPSS* e `Rstudio`.

Como sabemos, fatores como idade, sexo, etnia e massa muscular podem influenciar a relação entre o IMC e a gordura corporal. Além disso, o IMC não faz distinção entre excesso de gordura, músculo ou massa óssea, nem fornece qualquer indicação da distribuição de gordura entre os indivíduos. Apesar dessas limitações, o IMC continua a ser amplamente utilizado como indicador de excesso de peso. Como existem fatores que podem influenciar os valores de IMC, é necessário incluir, neste estudo, covariáveis adicionais para lidar com características complexas, i.e, aplicar modelos de associação de regressão logística e, posteriormente, avaliar a qualidade de um modelo de risco através da curva ROC (Operating Operating receiver Characteristic) e da AUC (Area Under The Curve).

Com base em toda a análise estatística apenas os SNPs PON_1_Q192R, the AdipoQ_G11377C, ACE_I_D and FTO_A_T_SNP demonstraram estar geneticamente associados com a obesidade. Contudo, através do modelo de regressão logístico apenas os SNPs AdipoQ_G11377C, FTO_A_T_SNP e PON_1_Q192R demonstraram ser estatisticamente significativos. Dado este resultado, o *score* de risco genético foi calculado apenas com base nestes 3 SNPs, apresentando um maior risco genético quando uma mulher é portadora de 2 alelos de risco pertencentes

aos SNPs AdipoQ_G11377C e PON_1_Q192R, mas não apresentar nenhum alelo de risco pertencente a FTO_A_T.
Este tipo de estudos de associação genética têm sido amplamente utilizados para melhor entender a patogénese genética de determinadas doenças a fim de melhorar as estratégias preventivas, meios de diagnóstico e terapias.

Palavras-Chave: Obesidade, IMC, Polimorfismos, *Score* de Risco Genético, Regressão Logística

Abstract

The obesity is a serious chronic disease associated with having excess body fat to the extent that it may have a negative effect on your health. Unhealthy lifestyle choices or environmental factors contribute to the development of this disease, which is known to be hereditary and highly polygenic. Millions of subtle variations in the human DNA sequence, or genome, hold the key to a host of conditions, from breast cancer to heart disease.
This genetic association case-control study compares the frequency of alleles or genotypes at genetic marker *locus*, i.e, single-nucleotide polymorphisms (SNPs), in a sample of individuals with and without a particular disease characteristic from a given population.
A sample of 212 Caucasian women, which is composed by 112 obese women and a control group (normal weight) of 100 women, was collected. For both groups was recorded the total body weight, body mass index (BMI), waist and hip circumference, waist-hip ratio and body fat. Moreover, 19 obesity-related SNPs in 13 obesity related genes, where genotyped for all samples. This type of study considers basic methods and techniques of statistical analysis and aims to determine whether there is an association between the disease characteristic and the genetic marker read-based association study. Multi-locus profiles of genetic risk, so-called "genetic risk *score*," can be used to translate discoveries from genome-wide association studies into tools for population health research. Therefore, the main purposes of this dissertation are identify polymorphisms associated with obesity in Portuguese Women, identifying the obesity genetic risk *score* and associate genetic polymorphisms with obesity related traits, using *Excel*, *SPSS* and `Rstudio` software. As we know, factors such as age, sex, ethnicity, and muscle mass can influence the relationship between BMI and body fat. Also, BMI does not distinguish between excess fat, muscle, or bone mass, nor does it provide any indication of the distribution of fat among individuals. Despite these limitations, BMI continues to be widely used as an indicator of excess weight. For this reason, it is necessary to understand if a high risk *score* is a guarantee of being obese, as their data show or if despite the strength of these associations, polygenic susceptibility to obesity is not deterministic.

Keywords: Obesity, BMI, Polymorphisms, Genetic Risk *Score*, Logistic Regression

Index

List of Tables

List of Figures

Abbreviations and Acronyms

AUC - Area Under the Curve;
BH - Benjamini-Hochberg Correction;
BMI - Body Mass Index;
CATT - Cochran-Armitage Trend Test;
DNA - Deoxyribonucleic Acid;
FN - False Negatives;
FP - False Positives;
FPR - False Positive Rate;
GRS - Genetic Risk *Score*;
GWAS - Genome-Wide Association Study;
HL - Hosmer and Lemeshow Goodness of Fit Test;
HWE - Hardy-Weinberg equilibrium;
OR - Odds Ratio;
p - *p-value*;
ROC - Receiver Operating Characteristic;
SNPs - Single Nucleotide Polymorphisms;
TN - True Negatives;
TP - True Positives;
TPR - True Positive Rate;
VIF - Variance Inflation Factor;
WHO - World Health Organization.

Glossary

Allele - A variant of a polymorphism at a *locus*.

B Cells - The human body has millions of different types of B cells every day circulating in blood and lymph which have an important role in immune surveillance. Each cell has a protein receptor (called a B cell receptor or BCR) on its binding scale. BCR is a major protein involved in the B cell, which link between the cell membrane and the immunoglobulin, and this molecule allows the distinction of B cells between other types of lymphocytes. Once the B cell is located, the antigen receives a signal additional T cell assists, it can differentiate into one of the two types of B cells.

Gene - Functional unit of DNA that contains the necessary information for the cell machinery to produce a RNA template that is either functional by itself or can be translated to a protein.

Genotype - Combination of two alleles across both chromosomes at a particular locus in an individual.

Hardy-Weinberg equilibrium - Given a minor allele frequency of q, the probabilities of the three possible genotypes (aa, Aa, AA) at a biallelic locus which is in Hardy-Weinberg equilibrium are ((1-q)^2, 2q(1-q), q^2). In a large randomly mating homogenous population these probabilities should be stable from generation to generation.

Single Nucleotide Polymorphism (SNP) - A genetic variant that consists of a single DNA base pair change, resulting in two possible allelic identities at that position.

The WNT Signaling pathway - The WNT signaling pathway is an ancient and evolutionarily conserved pathway that regulates crucial aspects of cell fate determination, cell migration, cell polarity, neural patterning and organogenesis during embryonic development. The WNTs are secreted glycoproteins and comprise a large family of nineteen proteins in humans hinting to a daunting complexity of signaling regulation, function and biological output.

Chapter 1

Introduction

1.1 Context of the Obesity Health Problem

To understand why obesity is advancing in nowadays it is necessary to take a journey to the past. Our ancestors had great difficulty in getting food and even more so to stock it. Therefore, nature was charged with endowing the human body with a mechanism for storing energy. This mechanism consisted in encouraging the man, through hunger, to ingest a great quantity of calories and to make his organism transform the excess into fat, storing it for periods of lack of food.

However, our ancestors ate mainly seeds, roots and fruits and it was for this food pattern that genetics prepared the organism inherited by us. The problem is that our lifestyle is completely different from that. Today, food is easily available in modern societies and on the other hand, the changes in our environment occurred more rapidly than the modifications in our genetic background. Therefore, when considering the imbalance in our modern lifestyle and our "ancient" genetic profile, it is understandable that many people gain weight so easily.

Obesity is a global public health concern from an excessive fat accumulation that results from a positive balance between total energy intake and fat catabolism.

This disease is associated with a set of hormonal changes affecting the perpetuation of that condition as well as the development of co-morbidities and may contribute for a significant number of diseases including stroke, metabolic syndrome, cardiovascular diseases, type 2 diabetes mellitus, premature death and some cancers.

However, human obesity is not only due to the excessive consumption of foods rich in sugars and fats, but also influenced by genetic factors and the environment in which one lives from the moment of maternal to adult life. A complex mix of genetic, environmental, and psychological factors can increase a person's risk for obesity.

Due in part to evolutionary forces, genetic drift and environmental conditions, changes occurred in the human species. For example, in all subpopulations there have always been obese and non-obese individuals. This difference arises mainly as a consequence of genetic factors, as evidenced by the high heritability of the body mass index (BMI). A characteristic such as eye color, hair color, body size, etc may reflect the activity of a single gene (Mendelian or monogenic) or more than one gene (polygenic). Both cases can be affected by environmental factors. The polygenic multifactorial condition reflects the additive condition of many genes conferring different degrees of susceptibility. Accordingly, we can understand a polygenic trait as the combined action of several genes producing a "continuously variable" phenotype. With

the advent of the Human Genome Project (1990-2003), millions of variants of DNA sequences have been discovered in the human genome.
In Portugal, obesity is a health problem that is affecting more and more the entire population. Almost half of the population is overweight and close to one million adults suffer from obesity. The principal objective of this project is to investigate for the first time, in Portugal, the genetic of common obesity in Portuguese women, which could help in the future to identify a genetic predisposition to obesity and develop possible approaches to treat this condition.

1.2 How Obesity is Classified?

Obesity is a medical condition in which excess body fat has accumulated to the extent that it may have an adverse effect on health and its prevalence during the past years has dramatically increased worldwide (Ogden et al., 2007). Obesity is measured in terms of body mass index (BMI), which is the most commonly measure used to classify overweight and obesity, adiposity, and waist/hip proportions. It is estimated that nearly 500 million people worldwide to have obesity and 1.4 billion are estimated to be overweight. In less than a generation, the total number of people with obesity has doubled.

Body Mass Index (BMI): defined as a person's weight in kilograms divided by the square of his height in meters (Kg/m^2) and is further evaluated in terms of fat distribution via the waist-hip ratio and total cardiovascular risk factors.
The BMI allows, in a quick and straightforward way, to tell if an adult is underweight, normal weight or overweight and has therefore been adopted internationally to classify obesity. Obesity is classified in three classes according to the WHO (World Health Organization):

- Class I (BMI 30.0 Kg/m^2- 34.9 Kg/m^2)
- Class II (BMI 35.0 Kg/m^2- 39.9 Kg/m^2)
- Class III (BMI $\geq$ 40.0 Kg/m^2)

There is a relation between the mentioned classes of obesity and the risk of comorbidities (Table 1.1), can be affected by several factors, including diet and physical activity.

Nutritional Status	BMI (kg/m^2)	Risk of Comorbidities
Underweight	< 18.5	Low (But increased risk of other clinical problems)
Normal range	18.5 - 24.9	Medium
Overweight	25.0 - 29.9	Increased
Obesity Class I	30.0 - 34.9	Moderate
Obesity Class II	35.0 - 39.9	Serious
Obesity Class III	≥ 40	Very Serious

Table 1.1: *WHO (2000)* - Classification for BMI in adults

An important aspect in the evaluation of the obese adult is the distribution of fat body. That is, when adipose tissue accumulates in the upper half of the body, especially in the abdomen, it

is said that obesity is android, abdominal or visceral, being typical in male individuals. When fat is distributed mainly in half the lower part of the body, particularly in the gluteal region and thighs, is said to be of the gynoid type, which is typical of the obese woman.
The identification of these morphological types is very important, since demonstrated today that visceral obesity is associated with metabolic complications, such as type 2 diabetes and dyslipidemia and cardiovascular diseases, such as hypertension, coronary heart disease and cerebrovascular disease. The prevalence of obesity during the past years has dramatically increased worldwide.

1.3 Genetics of Obesity

There is scientific evidence that there is a genetic predisposition in certain individuals, which determine a greater accumulation of fat in the abdominal area due to excessive energy intake and/or decreased physical activity. This visceral fat, located inside the abdomen, is directly related to the development of insulin resistance, metabolic syndrome associated with obesity. This genetic predisposition may be hereditary, that is, there will be a transmission of traces and, occasionally, the risk of suffering from diseases.
Mendelian inheritance is observed for some rare diseases. On the other hand, most common diseases do not present typical Mendelian inheritance. According to the common disease common variant hypothesis, some of those common variants lead to susceptibility to complex polygenic diseases. Each variant of each gene that influences a complex disease will have a small effect on the disease phenotype and susceptibility.

1.4 Syndromic Obesity vs Non-Syndromic Obesity

Syndromic obesity describes obese children and adults with mental retardation, dysmorphic features, organ-specific abnormalities, hyperphagia, and/or other signs of hypothalamic dysfunction. Obesity syndromes may be inherited in either an autosomal or an X-linked pattern. More than 100 syndromes are now associated to obesity, but the most frequent syndromes are Prader-Willi and Bardet. In this case, the patients are clinically severely obese and additionally distinguished with dysmorphic features, organ-specific developmental abnormalities and a mental retardation (Bell et al., 2005).
In Non-syndromic obesity, which will be the focus of this study, both autosomal dominant and recessive forms of obesity have resulted due to several gene mutations.

1.4.1 Non-Syndromic Obesity

1.4.1.1 Monogenic Forms of Obesity

According to Bell et al. (2005), mutations in genes that encode proteins with potential function in regulating appetite are responsible for Mendelian diseases in which obesity is the most obvious phenotype.
Based on genetic and phenotypic characteristics, several types of obesity are seen. Monogenic

forms of obesity result from an alteration in a single gene and follow the Mendelian pattern of inheritance, affecting about 5% of the population. This mutation occurs in genes of the leptin (LEP), Leptin Receptor (LEPR), pro-opiomelanocortin (POMC), Melanocortin 4 receptor (MC4R) and proconvertase 1 (PC1), affect appetite regulation resulting in a severe obesity phenotype due to hyperphagia, indicating that these pathways are explicitly important in regulating weight and adiposity in humans (Barness et al., 2007). Early onset of the disease and an extreme phenotype characterize monogenic obesity.

1.4.1.2 Polygenic Forms of Obesity

Polygenic obesity is the more common clinical situation which is responsible for more than 95% of cases (Bell et al., 2005). A group of alleles responsible for a trait is termed as "polygenic" variants being that these polygenic variants play a role in obesity. Here, the unbalanced lifestyle (stress, overeating, sedentary lifestyle ...) is responsible for obesity in association in genes. The contribution of each gene has only had a small effect on weight and the allelic effects can be additive or nonadditive. Variants of obesity genes show variation in frequency between obese subjects making the study of polygenic obesity more complex.
SNPs (Single Nucleotide Polymorphisms) may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions (regions between genes). Unlike in the case of monogenic obesity where a single mutation in a gene is causal in producing the disease phenotype, in polygenic obesity, each polymorphism confers susceptibility to obesity and the presence of an obesogenic environment leads to the phenotype.
For the detection and analysis of obesity genes and their variants, several molecular genetic approaches are employed to help in unraveling the genomics of obesity. These genetics approaches are linkage studies, candidate gene association study and genome-wide association studies (GWAS).

1.5 The Common *Loci* Associated With Obesity

The most commonly methodology used is the GWAS approach, which are allowing geneticists to scan numerous polymorphisms across the entire genome by using a powerful statistical methods to identify *loci* associated with a particular phenotype. According to Aguilar et al.(2012), recently, new *loci* associated with obesity have been reported, but their function and metabolic implications remain to be elucidated. Advances in genetics have revealed more than 15 *loci* associated with common obesity using hypothesis-free genome-wide association studies (GWAS).
Through GWAS, the first *loci* identified for obesity was the insulin-induced gene 2 (INSIG2). Although, replication studies demonstrated very inconsistent results. In this way, FTO gene was the first *loci* unequivocally associated with obesity and more than 50 genetic *loci* have been identified as being associated with at least one obesity-related trait. Within the 50 *loci* there are 5 *loci* that have entered into several studies related to obesity:

Melanocortin 4 receptor (MC4R)

The Melanocortin 4 receptor (MC4R) gene on chromosome 18q22 encodes 332 amino acid and is mainly responsible for regulating energy balance. It is expressed mainly in the central nervous system contributing to food intake and energy expenditure regulation. According to Mutch et al.(2006), the effects of mutations in the melanocortin-4 receptor gene, for which the obese phenotype varies in degree of severity among individuals, are now thought to be influenced by one's environmental surroundings.

The most frequent form of heredity of obesity is caused by mutations in the melanocortin receptor-4 (MC4R) gene.

The patients have an early linear and hyperphagia growth, low blood pressure and also present hyperfadiga, but not as severe as that observed in LEP deficiency.

The MC4R deficiency represents the common cause (1% - 6%) of morbid obesity in adults and children.

Pro-opiomelanocortin (POMC)

Complete POMC deficiency is caused by homozygous or compound heterozygous loss-of-function mutations in the POMC gene.

The POMC gene, located in the 2p23.3 region, is transcribed into various tissues, including corticotropic cells from the anterior pituitary, neurons from the arcuate nucleus of the hypothalamus and cells into the dermis and lymphatic system.

POMC is regulated by leptin and is cleaved by prohormone convertases to produce the melanocortin receptor (MC-R) ligands adrenocorticotrophin (ACTH) and melanocyte-stimulating hormones (MSH) alpha, beta and gamma. The red hair pigmentation, adrenal insufficiency and obesity are caused by deficiencies in the ligands and subsequent lack of activation of the MC1, MC2, and MC4 receptors, respectively.

Proconvertase 1 (PC1))

Proprotein convertase 1, PC1, is an enzyme that in humans is encoded by the PCSK1 gene (Proprotein convertase subtilisin/ kexin type 1). The PC1 is the enzyme largely responsible for the first step in the biosynthesis of insulin. PC1 enzyme performs the proteolytic cleavage of prohormones to their intermediate forms and it is present only in neuroendocrine cells such as brain, pituitary and adrenal and most often cleaves after a pair of basic residues within prohormones.

Deficiency of proconvertase 1, due to PCSK1 gene mutations, is reported as an important cause of obesity.

Leptin (LEP) and Leptin Receptor (LEPR)

The LEP gene is located on the chromosome 7q31.3, while the protein encoded by the leptin receptor (LEPR) gene is located on the chromosome 1P31.3 (Coll et al., 2004). The anorexigenic hormone leptin seems to be the main indicator of adiposity and the signal of the nutritional state, since its plasma levels are highly correlated with the number of adipocytes and the fat content.

When LEP is not detected in the blood, there is a great possibility of diagnosing the congenital deficiency of Leptin, due to homozygosity of the mutated gene that leads to the loss of its function. The mutation in the leptin receptor (LEPR) gene results in abnormal splicing of

mRNA, generating a receptor without the transmembrane and intracellular domains.
Thus, the mutant receptor circulates in high concentration, bound to leptin, leading to a high concentration of leptin in the blood, leading to extreme obesity.

FTO Gene

Recently, in studies conducted by researchers (Gerken et al., 2007), a gene (FTO) on chromosome 16q12.2, Figure 1.1, has been discovered, which is closely associated with body mass index control.
The Fat mass and obesity-associated protein, also known as Alpha-ketoglutarate-dependent dioxygenase FTO.

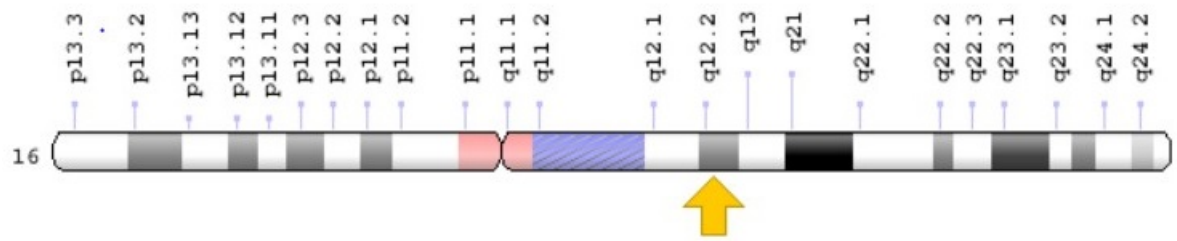

Figure 1.1: Chromossomal Location - FTO Gene. (Source:https://ghr.nlm.nih.gov/gene/FTO#location)

In 2009, variants in the FTO gene were further confirmed to associate with obesity in two very large genome wide association studies of body mass index (BMI).
According to Scuteri et al. (2007), the FTO gene showed the strongest association with BMI ($p - value = 8.6 \times 10^{-7}$), hip circumference ($p - value = 3.4 \times 10^{-8}$), and weight ($p - value = 9.1 \times 10^{-7}$), that in later we will have the capacity to discuss these values, depending on the results of this investigation.

1.6 Main Goals

This work will contribute to the genetic knowledge of obesity in Caucasian women and, in further studies, the genetic diversity that is associated with obesity in the Portuguese population could be compared with other populations.
To achieve the main goal of this work, it is important to establish more detailed objectives:

1. Identify the polymorphisms associated with obesity in Portuguese women, i.e, the association between obesity and 19 SNPs in 13 genes PON-1, AdipoQ, LEP, LEPR, GHRL, MC4R, ACE, ApoA5, FTO, IL6, PPARγ, TCF7L2 and TNF$\propto$ in the previously established sample (case-control study); ;
   (a) The Odds ratios will be calculated for allelic and genotypic (Dominant and Recessive models) contingency tables to compare the prevalence of obesity among persons with normal alleles/genotypes and persons with variant alleles/genotypes.
2. Associate genetic polymorphisms with obesity related traits;

(a) Tests of genetic association will be performed separately for each individual SNP. For alleles the Pearson's Chi-Squared will be applied. During the literature review, in biological research, the methodology that was used in the genotypic association was the Cochran Armitage Test, but the SNPs, in our database, are not an ordinal variables. Nevertheless, after all the previous analysis and after discovering which allele has the disease it was possible to build a *score* for the SNPs under study.

3. Identify the Obesity Genetic Risk Score.

   (a) In order to construct a Genetic Risk *Score*, the Binary Logistic Regression will be used. This methodology allows us to find the most parsimonious model consisting of SNPs that will be associated with obesity, by adding the number of risk alleles (0 or 1) across selected SNPs;

   (b) The ROC curve will be constructed and the AUC will be calculated to understand if the model can distinguish between patients with disease and without disease.

   (c) Finally, the Genetic Risk Score (GRS) will be built based on a combination of obesity-associated polymorphisms.

Chapter 2

Theoretical Framework and Methods

2.1 Study Sample

A sample of 212 Caucasian premenopausal women was selected in 2006, from Curry Cabral Hospital. The sample was composed of two groups. One of the groups was constituted by 112 obese Caucasian premenopausal women, which attended the obesity outpatient clinic. The control (normal-weight) group consisted of 100 Caucasian premenopausal women who either attended a routine health check or belonged to the health care staff of Curry Cabral Hospital. No woman was on any pharmacological regimen (except for oral contraceptives) or took any sporadic drug in the previous 7 days and only women without any previous diagnosis of any acute/chronic health condition (except obesity for the obese group) were selected for this study. A venous blood sample was collected from patients and controls. Genomic DNA was isolated from white blood cells by phenol extraction and the genotyping was done through realtime PCR with TaqMan probes, or PCR and agarose Gel. Each woman was characterized for total body weight, BMI, Waist and hip circumferences, Ratio waist-to-hip ratio and the body fat mass (bioelectrical impedance, Tanita TBF-300A ®).

2.1.1 The analyzed polymorphisms

In recent years, 52 genetic *loci* were identified to be unequivocally associated with obesity-related traits, in source populations (Loos, 2012). According to Frayling et. al (2007), a strong association was detected between common SNPs in the first intron of the fat mass and obesity-associated gene (FTO), on the chromosome 16q12.2 and risk of obesity.
In this case study, there are 13 genes that will be analyzed, of which 5 have already been enumerated in the previous chapter.

ADIPOQ Gene
Adiponectin is a hormone secreted by adipocytes that regulates energy homeostasis and glucose and lipid metabolism, located in the chromosomal region 3q27.3, as shown in Figure 2.1. ADIPOQ gene have been linked, in some SNPs Studies, with obesity and with adiponectin levels in various populations. According the investigation by Apalasamy et al. (2014), whose objective was investigate the association of ADIPOQ rs17366568 and rs3774261 SNPs with obesity and with adiponectin levels in Malaysian Malays. A significant genotypic association

was observed between ADIPOQ rs17366568 and obesity.
In this work, will be studied the AdipoQG (SNP: rs1501299 G/T), the AdipoQ_G11377C (SNP: rs266729 C/G), the AdipoQ_G11391A (SNP: rs17300539 G/A) and the AdipoQ_45T_G (SNP: rs224176 T/G, Intron Variant), all located in the chromosomal region 3q27.3.

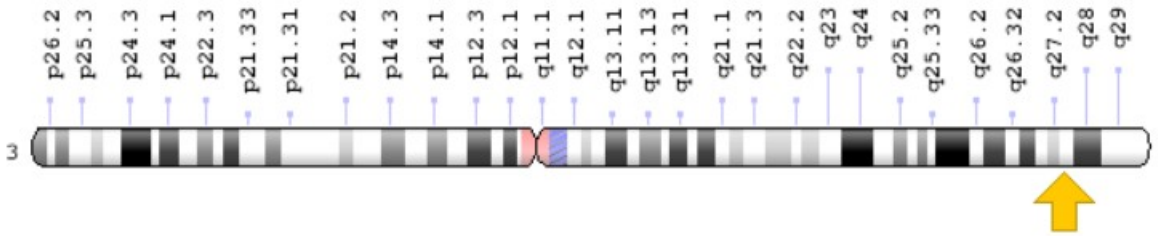

Figure 2.1: Chromossomal Location - ADIPOQ Gene. (Source: https://ghr.nlm.nih.gov/gene/ADIPOQ#location)

GHRL Gene

The GHRL (Ghrelin and Obestatin Prepropeptide) is a Protein Coding gene which encodes the ghrelin-obestatin preproprotein that is cleaved to yield two peptides, ghrelin and obestatin. This gene is located at region 3p25.3, on the short (p) arm of chromosome 3, at position 25.3 and contains five exons, as we can see on Figure 2.2.

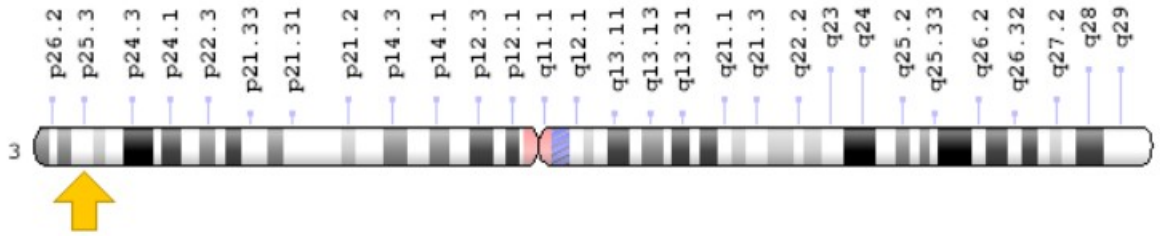

Figure 2.2: Chromossomal Location - GHRL Gene. (Source: https://ghr.nlm.nih.gov/gene/GHRL#location)

The Ghrelin is a powerful appetite stimulant and plays an important role in energy homeostasis. Its secretion is initiated when the stomach is empty, whereupon it binds to the growth hormone secretagogue receptor in the hypothalamus which results in the secretion of growth hormone (somatotropin). Ghrelin is thought to regulate multiple activities, including hunger, reward perception via the mesolimbic pathway, gastric acid secretion, gastrointestinal motility, and pancreatic glucose-stimulated insulin secretion.
The Ghrelin_R51Q, SNP: rs34911341 C/T with a missence mutation (Arg/Gln), and Ghrelin_Leu72Met, SNP: rs696217 (G/T) with a missence mutation (Leu/Met), both located in the same chromosomal region, will be studied in this work.

PON1 Gene

The paraoxonase 1 (PON1) is a protein coding gene, which encodes a member of the paraoxonase family of enzymes and exhibits lactonase and ester hydrolase activity. Succeeding synthesis in the liver and kidney, the enzyme is secreted into the circulation, where it binds to high density lipoprotein (HDL) particles and hydrolyzes thiolactone and xenobiotics, including paraoxon,

a metabolite of the insecticide parathion. This gene is located at region 7q21.3, in long (q) arm of chromosome 7, at position 21.3 (Figure 2.3). The diseases associated with PON1 are Microvascular Complications Of Diabetes 5 (MVCD5) and Amyotrophic Lateral Sclerosis 1. The PON_1_Q192R, SNP: rs662,C/T with a missence mutation (Gln/Arg), and the PON_1_M55L, SNP: rs854560 A/T, with a missence mutation (Met/Leu) will be studied here.

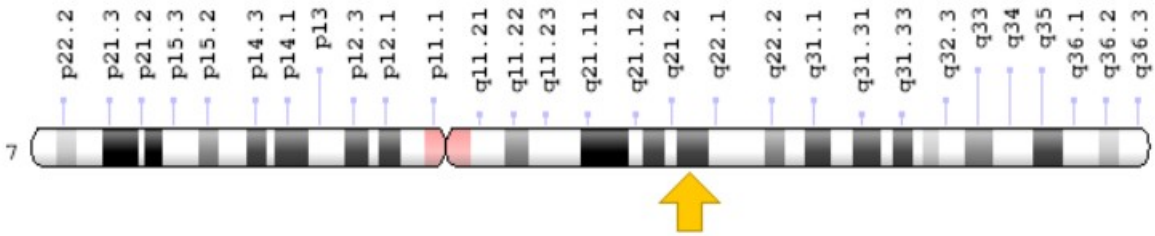

Figure 2.3: Chromossomal Location - PON1 Gene. (Source: https://ghr.nlm.nih.gov/gene/PON1#location)

ACE Gene

The Angiotensin I Converting Enzyme (ACE) gene provides instructions for making the angiotensin-converting enzyme. This enzyme can cut (cleave) proteins and by cutting a protein called angiotensin I at a particular location, the angiotensin-converting enzyme converts this protein to angiotensin II. The Angiotensin II protein causes blood vessels to narrow (constrict), which results in increased blood pressure. The ACE gene located on chromosome 17 at position 23.3, in the long (q) arm 17q23.3 (Figure 2.4) is part of the renin-angiotensin system, which regulates blood pressure and the balance of fluids and salts in the body. There are diseases associated with a certain variation in the ACE gene, as Microvascular Complications Of Diabetes 3 (MVCD3) and Renal Tubular Dysgenesis. The polymorphism ACE_I_D (SNP: rs4646994 (287 bp Ins/ Del)) will be studied.

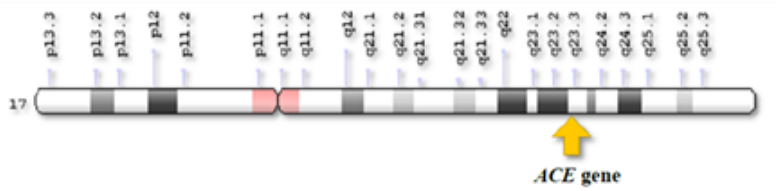

Figure 2.4: Chromossomal Location - ACE Gene. (Source: https://ghr.nlm.nih.gov/gene/ACE#location)

ApoA5 Gene

The ApoA5 gene, exclusively expressed by the liver is located proximal to the apolipoprotein gene cluster, on region 11q23.3, according to the Figure 2.5. The protein encoded by this gene is an apolipoprotein that plays an important role in regulating the plasma triglyceride levels, a major risk factor for coronary artery disease. According Xin et al. (2018), several studies has demonstrated an association between apoA5 and the increased risk of obesity and metabolic syndrome. They verified that apoA5 could significantly reduce plasma triglyceride (TG) level by stimulating lipoprotein lipase (LPL) activity, and the intracellular role of apoA5 has also been proved since apoA5 is associated with cytoplasmic lipid droplets (LDs) and affects intrahepatic

TG accumulation. So, mutations in this gene have been associated with hypertriglyceridemia and hyperlipoproteinemia type 5.
In this work, the ApoA5_T1131C, SNP: rs662799 T/C, Upstream gene variant (-1131), will be investigated.

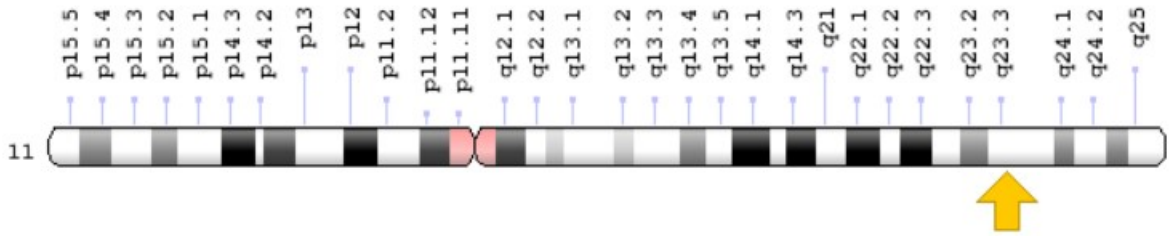

Figure 2.5: Chromossomal Location - ApoA5 Gene. (Source: https://ghr.nlm.nih.gov/gene/APOA5#location)

IL 6 Gene

The Interleukin 6 (IL6) Gene, SNP: rs1800796 (G/C), non coding transcript exon variant (UTR region), according to the Figure 2.6, is located at region 7p15.3. This gene encodes a cytokine that functions in inflammation and the maturation of B cells. The functioning of this gene is implicated in a wide variety of inflammation-associated disease states, including susceptibility to diabetes mellitus and systemic juvenile rheumatoid arthritis. The IL6 is primarily produced at sites of acute and chronic inflammation, where it is secreted into the serum and induces a transcriptional inflammatory response through interleukin 6 receptor, alpha.

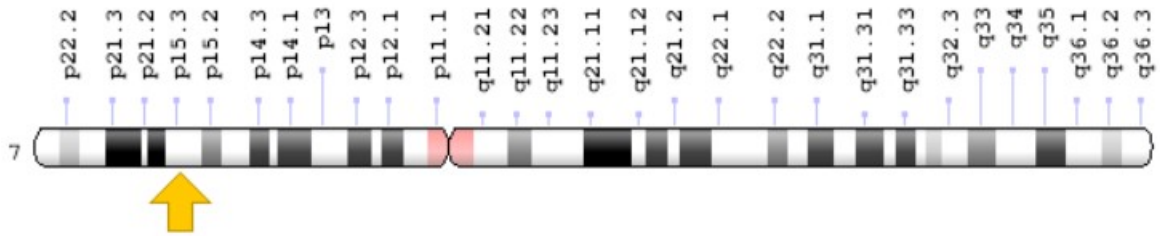

Figure 2.6: Chromossomal Location - IL6 Gene. (Source: https://ghr.nlm.nih.gov/gene/IL6#location)

PPARγ Gene

As we can observe on Figura 2.7, Peroxisome proliferator activated receptor gamma, PPARγ (SNP: rs1801282 (C/G) with a missence mutation (Pro/Ala)) gene is located at 3p25.2.

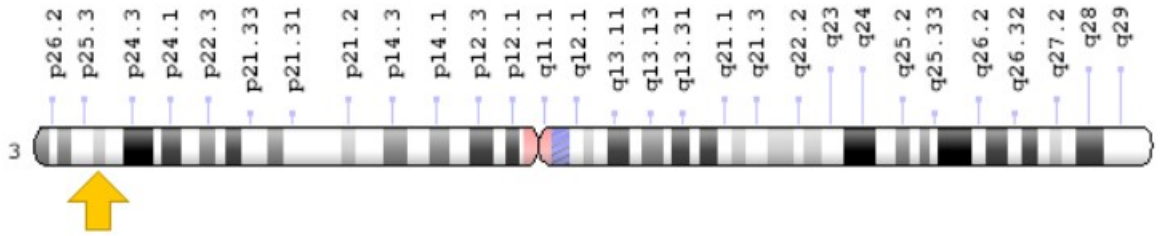

Figure 2.7: Chromossomal Location - PPARγ Gene.(Source: https://ghr.nlm.nih.gov/gene/PPARG#location)

The PPARγ gene encodes a member of the peroxisome proliferator-activated receptor (PPAR) subfamily of nuclear receptors. The PPARs form heterodimers with retinoid X receptors (RXRs) and these heterodimers regulate transcription of various genes. There are three subtypes of PPARs: PPAR-alpha, PPAR-delta, and PPAR-gamma. The PPAR-gamma is the protein encoded by PPARs and is a regulator of adipocyte differentiation. Additionally, PPAR-gamma has been implicated in the pathology of numerous diseases including obesity, diabetes, atherosclerosis and cancer.

TCF7L2 Gene

The Transcription factor 7 like 2, TCF7L2, is a protein coding gene located at region 10q25.2-q25.3, consonant Figure 2.8.

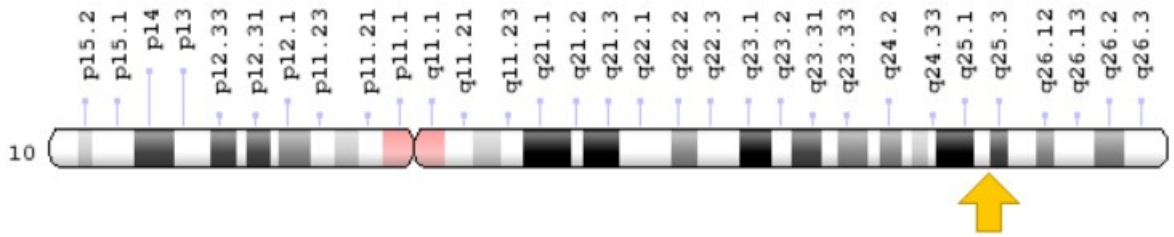

Figure 2.8: Chromossomal Location - TCF7L2 Gene.(Source: https://ghr.nlm.nih.gov/gene/TCF7L2#location)

The TCF7L2 encodes a high mobility group (HMG) box-containing transcription factor that plays a key role in the Wnt signaling pathway. The protein has been implicated in blood glucose homeostasis. Genetic alterations of this gene are associated with increased risk of type 2 diabetes.
In this work will be studied the TCF7L2_rs7903146_C_T (SNP: rs7903146 (C/T), Intron variant) polymorphism.

TNFα Gene

Tumor necrosis factor-alpha (TNFγ), SNP: rs1800629 (G/A),Upstream gene variant (-308), encodes a cytokine with pleomorphic actions and plays a pivotal role in inflammation, according to Russo et al., (2018). This cytokine is mainly secreted by macrophages and belongs to the tumor necrosis factor (TNF). The TNF gene is located at region 6p21.33, as we can see on Figure 2.9. TNF is involved in the regulation of a wide spectrum of biological processes including cell

proliferation, differentiation, apoptosis, lipid metabolism and coagulation. This cytokine has been implicated in a variety of diseases, including autoimmune diseases, insulin resistance, and cancer.

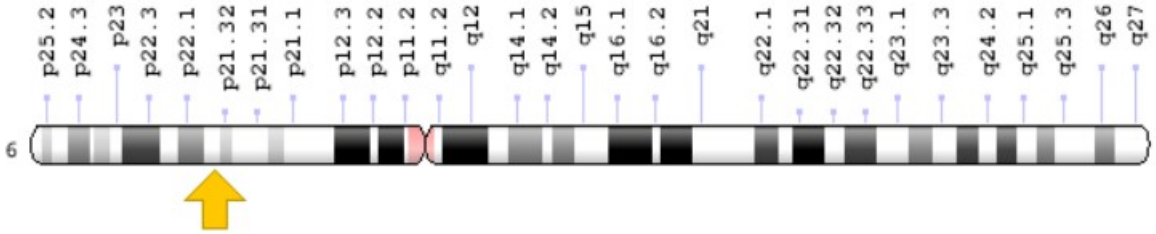

Figure 2.9: Chromossomal Location - TNFα Gene.(Source: https://ghr.nlm.nih.gov/gene/TNF#location)

2.2 Mendelian Genetics - Mendel's Laws of Inheritance

First, it is necessary to understand how genes can be hereditary and what their composition is. In 1860, an Austrian monk named Gregor Mendel introduced a new theory of inheritance based on his experimental work with pea plants.

Mendel believed that heredity is the result of discrete units of inheritance, and every single unit (or gene) was independent in its actions in an individual's genome. According to this Mendelian concept, inheritance of a trait depends on the passing-on of these units. For any given trait, an individual inherits one gene from each parent so that the individual has a pairing of two genes. We now understand the alternate forms of these units as 'alleles'. If the two alleles that form the pair are identical, then the individual is said to be homozygous and if the two genes are different, then the individual is heterozygous for the trait.

Based on his pea plant studies, Mendel proposed that traits are always controlled by single genes. After crossing two plants which differed in a single trait, Mendel discovered that the next generation, the "F1" (first filial generation), was comprised entirely of individuals exhibiting only one of the traits. However, when this generation was interbred, its offspring, the "F2" (second filial generation), showed a 3:1 ratio - three individuals had the same trait as one parent and one individual had the other parent's trait.

Mendel then theorized that genes can be made up of three possible pairings of heredity units, which he called "factors": AA, Aa, and aa. The big 'A' represents the dominant factor and the little 'a' represents the recessive factor. In Mendel's crosses, the starting plants were homozygous AA or aa, the F1 generation were A or a and the F2 generation were AA, Aa, or aa. The interaction between these two determines the physical trait that is visible to us.

Mendel's Law of Dominance predicts this interaction. When mating occurs between two organisms of different traits, each offspring exhibits the trait of one parent only. If the dominant factor is present in an individual, the dominant trait will result. The recessive trait will only result if both factors are recessive.

Therefore, Mendel's observations and conclusions are summarized in the following two principles, or laws.

Law of Segregation

The Law of Segregation states that for any trait, each parent's pairing of genes (alleles) split and one gene passes from each parent to an offspring. Which particular gene in a pair gets passed on is completely up to chance.

Law of Independent Assortment

The Law of Independent Assortment states that different pairs of alleles are passed onto the offspring independently of each other. Therefore, inheritance of genes at one location in a genome does not influence the inheritance of genes at another location.

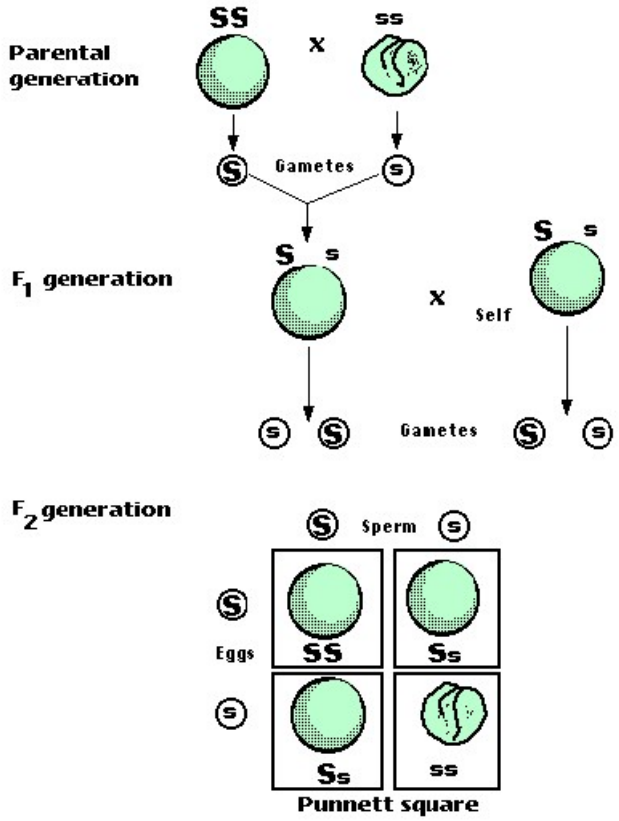

Figure 2.10: Mendel's Laws of Inheritance. (Source: Mendelian Genetics - Genetics Generation (http://knowgenetics.org/mendelian-genetics/) Last access in: 28/07/2019

2.3 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium (HWE) is a principal stating that the genetic variation in a population will remain constant from one generation to the next in the absence of disturbing factors, i.e. no mutation, no migration, no selection, random mating and infinite population size and can be calculated through a mathematical expression. When mating is random in a large population in these circumstances, the law predicts that both genotypic and allelic frequencies will remain constant because they are in equilibrium. In population genetics studies, the Hardy-Weinberg equation can be used to measure whether the observed genotype frequencies in a population differ from the frequencies predicted by the equation.
However, when mutations occur, they disrupt the equilibrium of allele frequencies by introducing

new alleles into a population. Similarly, natural selection and nonrandom mating break the HWE, because they result in changes in gene frequencies.
There are factors that can alter the HWE. One of them is when certain alleles help or harm the reproductive success of the organisms that carry them. Another factor is genetic drift, which occurs when allele frequencies grow higher or lower by chance and typically takes place in small populations. Gene flow, which occurs when breeding between two populations transfers new alleles into a population, can also alter the Hardy-Weinberg equilibrium.

2.3.1 The Hardy-Weinberg equation:

The Hardy-Weinberg law can be used under some circumstances to calculate genotype frequencies from allele frequencies. To explore the Hardy-Weinberg equation, we can examine a simple genetic *locus* at which there are two alleles, A and a. If the p and q allele frequencies are known, then the frequencies of the three genotypes may be calculated using the Hardy-Weinberg equation. The Hardy-Weinberg equation is expressed as:

$$p^2 + 2pq + q^2 = 1 \tag{2.1}$$

Where:

p is the frequency of the A allele, in the population, $0 \leq p \leq 1$;

q is the frequency of the a allele in the population, $0 \leq q \leq 1$.

The distribution of allele frequencies is the same in men and women, i.e.: men (p); women (q) and if they procreate the equality remains in the next generation:

$$(p+q)^2 = p^2 + 2pq + q^2 = 1 \tag{2.2}$$

Where:

$p + q = 1$;

p^2 = frequency of the homozygous genotype AA;

$2pq$ = frequency of the heterozygous genotype Aa;

q^2 = frequency of the homozygous genotype aa.

And these frequencies remain constant in successive generations.

2.4 Odds and Odds Ratio (OR)

The odds of an event are defined as the probability that the event will occur (p, i.e., probability of success) divided by the probability that the event does not occur (1-p, i.e., unsuccess).

$$\text{Odds} = \frac{p}{1-p}. \tag{2.3}$$

Probability p always ranges between 0 and 1.
The Odds Ratio is the measure of association for a case-control study. It tells us how much higher the odds of exposure are among cases of disease compared with controls. The Odds Radio is one of the main ways to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B, in a given population.

2.4.1 OR Calculation from contingency table

If each individual in a population either does or does not have a property "A" (e.g. "obesity"), and also either does or does not have a property "B" (e.g. "allele B") where both properties are appropriately defined, then a ratio can be formed which quantitatively describes the association between the presence/absence of "A" (obesity) and the presence/absence of "B" (allele B) for individuals in the population. For this example, an odds ratio (OR) can be calculated following three steps, which are described below:
Initially, for a given individual or any set of entities that have "B" calculate the odds that the same individual has "A". Secondly, for a given individual that does not have "B" calculate the odds that the same individual has "A". Finally, divide the odds from step 1 by the odds previously calculated to obtain the odds ratio (OR).
For a case-control study, the data look like this:

Exposed	Case	Control	Total
Yes	a	b	a + b
No	c	d	c + d
Total	a + c	b + d	a + b + c + d

Table 2.1: Contingency table

$$\text{Odds of exposure (cases)} = \frac{\text{number of cases with the exposure}}{\text{number of cases without exposure}} = \frac{a}{c} \tag{2.4}$$

$$\text{Odds of exposure (controls)} = \frac{\text{number of controls with the exposure}}{\text{number of controls without exposure}} = \frac{b}{d} \tag{2.5}$$

$$\text{OR} = \frac{\text{Odds of exposure (cases)}}{\text{Odds of exposure (controls)}} = \frac{ad}{bc} \tag{2.6}$$

An odds ratio:
Less than 1 means that the exposure (allele B) is associated with lower odds of outcomes (obesity),
Greater than 1 means that there is a higher odds of obesity happening with exposure to the allele B. In other words, the odds of exposure among cases is greater than the odds of exposure among controls.

Equal to 1 (or close to 1) means that exposure to allele B does not affect the odds of obesity, i.e, means that the odds of exposure among cases is the same as the odds of exposure among controls.

2.5 Association Tests On Contingency Tables

Tests of genetic association are usually performed separately for each individual SNP. The data for each SNP with minor allele a and major allele A can be represented as a contingency table of counts of disease status by either genotype count (e.g., aa, Aa and AA) or allele count (e.g., a and A). A genetic association case-control study compares the frequency of genotypes or alleles at genetic marker *loci*, usually single-nucleotide polymorphisms (SNPs), in individuals from a given population, with and without a given disease trait, in order to determine whether a statistical association exists between the disease trait and the genetic marker (Clarke et al., 2011). Although individuals can be sampled from families ("family-based" association study), the most common design involves the analysis of unrelated individuals sampled from a particular outbred population ("population-based association study").
Disease-related traits are usually the main trait of interest and any of the methods described here to test for genetic association, are generally applicable to any binary trait (exposed / non exposed), as Fisher's Test. Nonetheless, the Cochran-Armitage Trend Test is typically used in categorical data analysis when some categories are ordered and the *score* is chosen as the number of alleles (0, 1, 2).

2.5.1 Cochran Armitage Trend Test (CATT)

In biological research, 2×K genotype contingency tables of N case-control are frequently used for the analysis of ordered categorial data, as we can see through the Table 2.2, which is an example of a 2×3 contingency table. The Cochrane Armitage Trend Test (CATT) has become a standard procedure for association candidate gene testing in large-scale genome-wide association studies (GWAS)(Emily, 2018).

	aa (w_0=0)	Aa (w_1=1)	AA (w_2=2)	Total
Case	n_0	n_1	n_2	n
Controls	m_0	m_1	m_2	m
Total	N_0	N_1	N_2	N

Table 2.2: 2×3 Contingency Table of N case-control by genotype (aa, aA, AA)

Considering a single-marker locus with two possible alleles which are commonly denoted by A and a, each individual has three possible genotypes AA, Aa, and aa. Then we denote the two alleles by 0 and 1 instead of A and a and the genotypes by 0, 1, 2, the sum of the two allele indices involved. Finally, we assume a random sample of n cases and m unrelated controls. The case-control data can then be summarized according to genotypes as shown in Table 2.2.

The Cochran Armitage Test is different from the Pearson Chi-Squared Test. Here, there are K ordered groups in return for the binary response variable. The Cochran Armitage test for trend (CATT) is frequently used to calculate the trend of binominal proportions. These proportions are ordinal or quantitative metric or assignable scores over independent groups in K categories (Tekindal et al., 2016). For example, K groups can be ordered as normal, moderately normal, and abnormal, and 1, 2 and 3 can be assigned to them respectively as scores. This test is widely used in epidemiological and genetic research, in biomedical studies and in toxicological risk assessment (Kpoghomou et al., 2013). The CATT is based on an asymptotic approach and thus shows a poor performance in very small and unbalanced samples.

According to Ghodsi et al., (2016) the power of the test is very often improved as long as the probability of having the disease increases with the number of disease-associated alleles. In genetic association studies in which the underlying genetic model is unknown, the additive version of this test is the most commonly used.

The null hypothesis is the hypothesis of no trend, which means that the binomial proportion is the same for all levels of the explanatory variable.

The test is sensitive to the linearity between independent variable (e.g.: Group case/controls) and dependent variables (e.g: Genotype/ Alleles) and detects trends that would not be noticed by more crude methods, that is, for example, the Pearson Chi-Squared Test.

In order to measure the effect of genotype i and to detect particular types of association, the weights have been introduced, w_i. The special choice (w_0, w_1, w_2) = (0,1,2), represents the additive effect of allele A.

Here, (n_0, n_1, n_2) are the counts of the genotypes in cases and (m_0,m_1,m_2) are counts of the genotypes in controls, (N_0, N_1, N_2) are the counts of the genotypes in case-control samples and (w_0, w_1, w_2) are the number of disease alleles. Let n and m be the total number of cases and controls, respectively, and the total sample size, N=n+m. As cases and controls are independently sampled the genotype counts for cases and controls follow independent multinomial distributions with parameters (p_0,p_1,p_2), and (p'_0,p'_1,p'_2), respectively, where p_i and p'_i, i = 0,1,2, are the genotype probabilities in cases and controls.

$$(n_0,n_1,n_2) \sim \text{Multinomial } (n;\, p_0,p_1,p_2),\ ^{1}$$

$$(m_0,\, m_1,\, m_2) \sim \text{Multinomial } (m;\, p'_0,p'_1,p'_2)^{1}$$

Under the null hypothesis of no genetic association (Homogeneity):

$$H_0 : p_i = p'_i \text{ for i} = 0,1,2.$$

The Cochran-Armitage's trend test statistic for the data in Table 1.3 is given by:

$$T = \frac{N(N(n_1 + 2n_2) - n(N_1 + 2N_2))^2}{n(N-n)(N(N_1 + 4N_2) - (N_1 + 2N_2)^2)} \tag{2.7}$$

and follows the Chi-Squared distribution with one degree of freedom (df) under the null hypothesis.

[1]For simplicity and following the notation at Ghodsi et al. (2016), we decided to keep the random variables with small letters.

However, Agresti (2007) considered that CATT can be set in terms of the Pearson Chi-Squared statistic. Consider a contingency table $2\times J$ with ordered column (Table 2.3).
Let $n_j \sim$ Bin (N_j, p_j), j=0, ... ,J-1, it is of interest to test the following null hypothesis:

$$H_0 : p_0 = p_1 = \cdots = p_{J-1} \quad vs. \quad H_1 : \exists i, j = 1, ..., j-1; i \neq j, p_i \neq p_j \tag{2.8}$$

Score					
	w_0	w_1	...	w_{J-1}	
Case	n_0	n_1	...	n_{J-1}	n
Control	m_0	m_1	...	m_{J-1}	m

Table 2.3: $2\times K$ Contingency Table

It can be carried out by using a linear probability model

$$p_j = \alpha + \beta w_j \tag{2.9}$$

One can use the ordinary least square approach for testing β. Let

$$\bar{w} = \sum N_j w_j / N; \tag{2.10}$$

$$\tilde{p_j} = n_j / N_j \tag{2.11}$$

and

$$\hat{p} = n/N. \tag{2.12}$$

The prediction equation is

$$\hat{p_j} = \hat{p} + \hat{\beta}(w_i - \bar{w}) \tag{2.13}$$

where

$$\hat{\beta} = \frac{\sum N_j(\tilde{p_j} - \hat{p})(w_j - \bar{w})}{\sum N_j(w_j - \bar{w})^2}. \tag{2.14}$$

When the linear probability model holds, the statistic Z^2, under H_0 follows an approximately Chi-Squared distributionon with 1 degree of freedom and tests for a linear trend in the proportions. The trend test may give strong evidence of positive or increasing linear trends, of constant or stable trends over time, or of negative or decreasing trends. Results of the trend test are similar to those obtained by testing that the slope is zero in a linear logit model.

$$Z^2 = \frac{\hat{\beta}^2}{\hat{p}(1-\hat{p})} \sum_j N_j(w_j - \bar{w})^2, \tag{2.15}$$

When the disease model is unknown, there is consensus on the most powerful test to be used between CATT, allelic and genotypic tests. According to Emily (2018), although power for CATT depends on the sample size, the case-to-control ratio and the minor allelic frequency, there is largely influenced by the mode of inheritance and deviation from Hardy-Weinberg Equilibrium (HWE). Furthermore, when compared to other tests, CATT is shown to be the most powerful test

under a multiplicative disease model or when the single-nucleotide polymorphism largely deviates from HWE. In all other situations, CATT lacks in power and differences can be substantial, especially for the recessive mode of inheritance.

2.5.2 Fisher's Test

Fisher's Test is applied to a 2×2 contingency table (Table 2.4) and it is used to test the independence of two variables, where the hypotheses underlying the test, focused in this study are:

H_0: There is no genetic association between disease and alleles

vs.

H_1: There is genetic association between disease and alleles

	Exposed	Non Exposed	Total
Case	A	B	A+B
Control	C	D	C+D
Total	A+C	B+D	n

Table 2.4: 2x2 contingency table

The Fisher's Exact Test is more accurate than the Chi-Squared Test when the expected numbers are small, because the *p-value* is required for all sample sizes, while the results from the Chi-Squared Test that examines the same hypotheses may be imprecise when the number of cells is small. Fisher's exact test is based on the hypergeometric distribution and it is characterized in estimating only *p-value*, that is, no test statistic is used. Therefore, the *p-value* is conditional on the marginal totals of the table.
The *p-value* under H_0 is determined by finding all possible tables, keeping the same marginal totals and varying the lowest observed frequency. For each table, the respective *p-value* is estimated, given by:

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{n!A!B!C!D!} \tag{2.16}$$

Thus, the final *p-value* is the sum of all *p-values* calculated for the tables with a situation equal to or more extreme than observed (according to the direction of H_1). The Hypothesis H_0 is rejected when *p-value* is less or equal than the significance level. It should be noted that this test implies time-consuming calculations, which are now easily surpassed by the use of appropriate software (in this case, the use of `RStudio`).
However, despite the possibility of using Fisher's test to test if there is a genetic association between the allele and obesity, the Chi-square test will be applied, because the cells did not have a small number of observations. The Fisher Test will be applied only for the Odds ratio test.

2.6 The Multiple Logistic Regression Analysis

One of the major reasons the logistic regression model has seen such wide use, especially in epidemiologic research, is the ease of obtaining adjusted odds ratios from the estimated slope

coefficients when sampling is performed conditional on the outcome variables, as in a case-control study. Although this is not the case in this study. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial (Sperandei, 2014). Look at a compilation of m independent variables denoted by the vector $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_m)$ where each of these variables is at least interval scaled. Let Y be the random variable with Bernoulli distribution, so that, the conditional probability that the outcome Y=1 is present is denoted by $\mathrm{P(Y=1|\ x)} = \pi(\mathbf{x})$.
The equation for the logit of the multiple logistic regression model is given by:

$$g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m \tag{2.17}$$

For the multiple logistic regression model the equation is:

$$\pi(\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m)} = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}, \tag{2.18}$$

Where $\beta_i' s$ are the regression coefficients associated with the reference group.

According to Hosmer et al.(2000), if some of the independent variables are discrete, nominal scale variables such as sex, treatment group, race and so forth, it is inappropriate to include them in the model as if they were interval scale variables. The various levels of these nominal scale variables can be represented through the numbers that are merely identifiers and have no numeric significanse. So, in this case, the method of choice is to use *dummy* variables. For example, let's suppose that one of the independent variables is eye color, which has been coded as "blue", "brown" and "other". In this example, two *dummy* variables are necessary. One possible coding strategy is that when the respondent is "blue", the two *dummy* variables, D_1 and D_2 would both be set equal to zero. When the respondent is "brown", D_1 would be set equal to 1 while D_2 would still equal to 0. When the eye color of the respondent is "other", we would use $D_1 = 0$ and $D_2 = 1$, as we can see through the table 2.5.

EYE COLOR	D_1	D_2
Blue	0	0
Brown	1	0
Other	0	1

Table 2.5: Coding of *dummy* variables for *eye color* coded at three levels

Nevertheless, in our case study we only have two levels for the dominant and recessive models, but if a nominal scaled variable has k possible values, then k - 1 *dummy* variables are needed. The reason for using one less than the number of values is that, unless stated otherwise, the models have a constant term (Hosmer et al.,2000).
Imagine that the jth independent variable x_j has k_j levels. The k_j-1 *dummy* variables will be denoted as D_{jl} and the coefficents for these *dummy* will be denoted as β_{jl}, l= 1, 2, ..., k_j -1. In this way, the logit for a model with m variables with the jth variable being discrete is

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + ... + \beta_m x_m \tag{2.19}$$

In agreement with Hosmer et al.(2000), the summation and double subscripting needed to indicate when *dummy* variables are being used are suppressed when discussing the multiple logistic regression model.
In the case of the dependent random variable Y assuming only two possible states (0 or 1) and be a set of m independent variables X_1, X_2, ..., X_m, the logistic regression model can be written as follows:

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}}. \tag{2.20}$$

2.6.1 The Process of Fitting the Multiple Regression Model

The Maximum Likelihood is the method of estimation used in the multivariable case and in the univariable situation.
Assuming a sample of n independent observations ($\mathbf{x}_i$,y_i), i=1, 2, ..., n. Fitting the model requires that we obtain estimates of the vector $\beta = (\beta_0, \beta_1, ..., \beta_m)$.
The likelihood function is identical to (2.18). There will be m+1 likelihood equations that are obtained by differentiating the log-likelihood function with respect to the m+1 coefficients. The resulting likelihood equations may be expressed as follows:

$$\sum_{i=1}^{n}[y_i - \pi(\mathbf{x}_i)] = 0 \tag{2.21}$$

and

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}i)] = 0 \tag{2.22}$$

for j=1, 2, ..., m and $\mathbf{x}_i$=(x_{i1},x_{i2},..., x_{im}) .

Therefore, the fitted values for the multiple logistic regression model are a vector of dimension m, $\widehat{\pi}_i$, with the value of the expression in equation (2.18) being computed using $\widehat{\beta}$, which is a vector of dimension m+1, and $\mathbf{x}_i$.
The variances and covariances of the estimated coefficients can be estimating through the maximum likelihood estimation. The maximum likelihood estimation states that the estimators are obtained from the matrix of second partial derivatives of the log-likelihood function and these partial derivatives have the following general form

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i(1 - \pi_i) \tag{2.23}$$

and

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i(1 - \pi_i) \tag{2.24}$$

for j,l = 0,1,2,...m where π_i denotes $\pi(\mathbf{x}_i)$. The $(p + 1) \times (p + 1)$ is the matrix containing the negative of the terms given in equations (2.22) and (2.23). This matrix is denoted as $\mathbf{I}(\widehat{\beta})$ and is called the "observed information matrix". The variances and covariances of the coefficients

are obtained from the inverse of this matrix, which we denote as $Var(\widehat{\beta}) = \mathbf{I}^{-1}(\widehat{\beta})$. However, in very special cases it is not possible to write down an explicit expression for the elements in this matrix. Thus, it is necessary to use the notation $Var(\widehat{\beta}_j)$ to denote the jth diagonal element of this matrix, which is the variance of $\widehat{\beta}_j$ and $Cov(\widehat{\beta}_j, \widehat{\beta}_l)$ to denote an arbitrary off-diagonal element, which is the covariance of $\widehat{\beta}_j$ and $\widehat{\beta}_l$. The estimators of the variances and covariances, which will be denoted by $\widehat{Var}(\widehat{\beta})$ are obtained by evaluating $Var(\widehat{\beta})$ at $\widehat{\beta}$. The $\widehat{Var}(\widehat{\beta}_j)$ and $\widehat{Cov}(\widehat{\beta}_j, \widehat{\beta}_l)$, j, l =0,1,2,.., m are used to denote the values of the matrix. The estimated standard errors of the estimated coefficients are denoted as

$$\widehat{SE}(\widehat{\beta}_j) = [\widehat{Var}(\widehat{\beta}_j)]^{1/2} \tag{2.25}$$

for j=0,1,2,...,m.

2.6.2 The Significance of the Model

In order to fit a particular multiple logistic regression model, the first step is usually to assess the significance of the variables in the model. According to Hosmer et al.(2000), the likelihood ratio test for overall significance of the m coefficients for the independent variables in the model is performed in exactly the same manner as in the univariable case. The test is based on the statistic G given by:

$$G = 2\left\{\sum_{i=1}^{n}[y_i \ln(\widehat{\pi}_i) + (1-y_i)\ln(1-\widehat{\pi}_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n)]\right\} \tag{2.26}$$

Here, the fitted values, $\widehat{\pi}_i$, under the model are based on the fitted model containing m+1 parameters, $\widehat{\beta}$. Under the hypothesis that the m incline coefficients for the covariates in the model are equal to zero, the distribution of G is chi-square with m degrees of freedom.
To check the signicance of the model, we also need to look at the univariable Wald Test statistics:

$$W_j = \frac{\widehat{\beta}_j}{\widehat{SE(\widehat{\beta}_j)}} \tag{2.27}$$

Under the hypothesis that an individual coefficient is zero, these statistics will follow the standard normal distribution. The goal here is obtain the best fitting model while minimizing the number of parameters. So, the next step is to fit a reduced model containing only those variables thought to be significant and compare that reduced model to the full model containing all the variables.

2.6.3 Logistic Regression "Step-by-Step"

The *stepwise* method is a combination of the forward and backward selection techniques, i.e, a step-by-step iterative construction of a regression model that involves automatic selection of explanatory variables.
Stepwise regression can be achieved either by trying out one independent variable at a time and including it in the regression model if it is statistically significant (forward), or by including

all potential independent variables in the model and eliminating those that are not statistically significant (backward), or by a combination of both methods (Sperandei, 2014). The goal is to find a set of independent variables which significantly influence the dependent variable. Conducting these tests automatically can potentially save time for the individual.

2.6.4 ROC Curve and AUC

An incredibly useful tool in evaluating and comparing predictive models (e.g: Logistic Regression) is the ROC curve. The ROC curve or "Receiver Operating Characteristic" is a way to see how any predictive model can distinguish between the true positives and negatives. There are useful statistics that can be calculated from this curve, like the Area Under the Curve (AUC) and the Youden Index. These tell you how well the model discriminate and the optimal cut point for any given model (under specific circumstances).
The ROC curve, when representing the sensitivity and specificity for all possible values for the cutoff point, is one of the most used tools to evaluate and compare different types of diagnostic methodologies. In addition, the AUC is a measure of the performance of the associated test.
The AUC varies between 0.5 and 1, and a classifier with AUC near to the 1 means that the model has a good measure of separability. A poor model has AUC near to the 0.5 which means it has worst measure of separability. When AUC is approximately 0.5, the model has no discrimination capacity to distinguish between positive class and negative class. However in practice, the AUC performs well as a general measure of predictive accuracy.
Although ROC curves are often used for evaluating and interpreting logistic regression models, they are not limited to logistic regression. A common usage in medical studies is to run an ROC to see how much better a single continuous predictor (a "biomarker") can predict disease status compared to chance.

Interpreting the ROC Curve

The ROC curve shows the trade-off between sensitivity (or TPR - *True Positive Rate*) and specificity (1 - FPR (*False Positive Rate*)), according to the Figure 1.3.

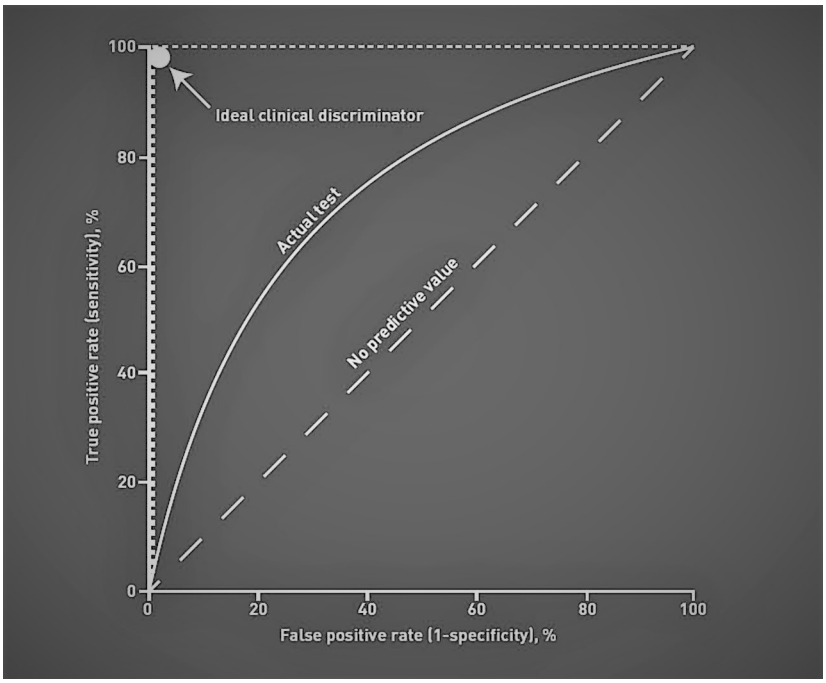

Figure 2.11: ROC Curve. (Source: https://www.theanalysisfactor.com/what-is-an-roc-curve/) Last Access in: 02/08/2019

Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

Note that the ROC curve does not depend on the class distribution. This makes it useful for evaluating classifiers predicting rare events such as rare diseases or disasters. In contrast, evaluating performance by measuring accuracy:

$$\frac{(\mathrm{TP}+\mathrm{TN})}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP})} \tag{2.28}$$

Where:

$$\text{Sensitivity} = \mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \tag{2.29}$$

$$\text{Specificity} = \frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FP})} \tag{2.30}$$

$$\mathrm{FPR} = 1 - \text{Specificity} = \frac{\mathrm{FP}}{(\mathrm{FP}+\mathrm{TN})} \tag{2.31}$$

For Logistic Regression one can create a 2×2 classification table of true (Y) and predicted values (E) from the model for response: E = 0 or 1 versus the true value of Y=0 or 1. The prediction (E) being equal to 1 depends on some cut-off probability, π_0. For example, for some individual i, E=1 if $\widehat{\pi}_i > \pi_0$ and E=0 if $\widehat{\pi}_i \leq \pi_0$. The most common value for π_0 is 0.5. Then Sensitivity is equal to P(E=1| Y=1) and Specificity is P(E=0| Y=0).

Chapter 3

Statistical Analysis - Results

3.1 Exploratory Analysis

To undersrtand the behavior of the data and before proceeding with the analysis of the SNPs it is necessary to make an exploratory and graphic analysis of them. The exploratory analysis of the project was performed using the `RStudio` statistical software `(R version 1.1.456)`. In this case study there are quantitative and qualitative variables.
The quantitative variables are numerical variables: counts, percents, or numbers which are expressed as means, quartiles and standard deviations (SD), as we can see in Figure 4.1, in the Appendix.
Categorical variables are characterized by not having quantitative values and being defined by various categories, that is, they represent a classification of individuals and can be ordinal or nominal. These variables are expressed as absolute and relative frequencies.
As the problem under study is centered on the women's obesity, we started by analyzing the variable *Obesity Class*, which is classified according to the BMI. This variable is an ordinal variable, whose classes are already defined in this database. Within obese women we have 3 obesity classes, ie, Class I, Class II and Class III of Obesity (categories 2, 3 and 4). In the Figure 3.1 and through the Table 3.1, we can observe the majority of obese women, that is about 66.96%, belong to class III of obesity.

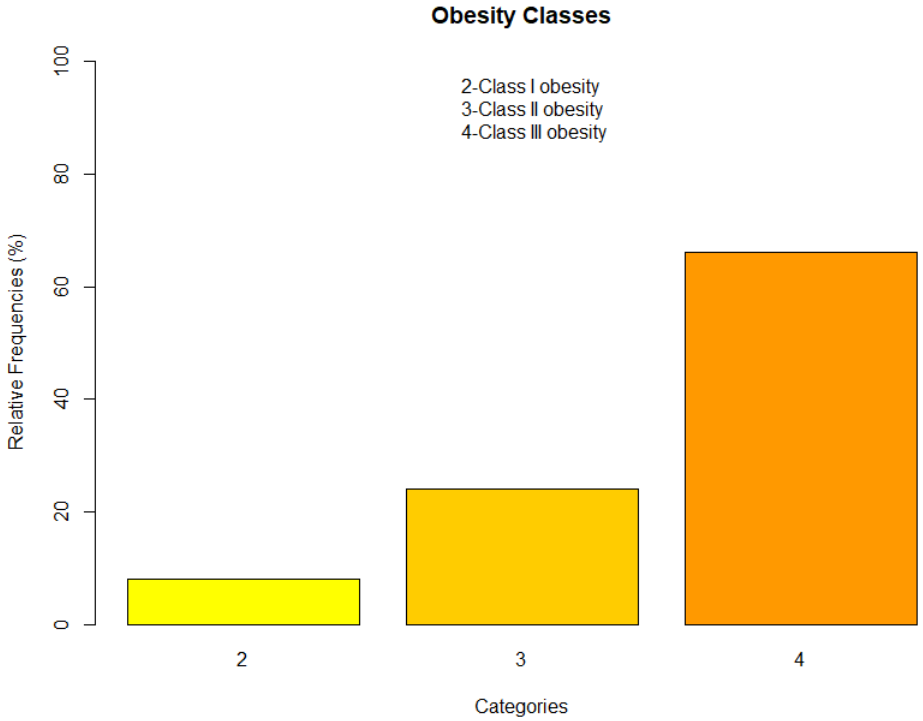

Figure 3.1: Obesity Classes Bar Diagram

Categories	Absolute Frequencies	Relative Frequencies (%)
2-Class I obesity	10	8.93%
3-Class II obesity	27	24.11%
4-Class III obesity	75	66.96%
Total	112	100%

Table 3.1: Absolute and Relative frequencies of *Obesity Classes* variable

There are 10 women whose belong to Class I, which represent 8.93% of the obese sample and in the Class III we have 27 women, about 24.11% of the obese women.
In order to have a better understanding of the data, some variables that will not be considered later in the statistical models and tests, were subjected a descriptive analysis. These variable are *Smoker*, *Surgery*, *Hypertension*, *Contraception*, *type of surgery* and *type of intervention*.

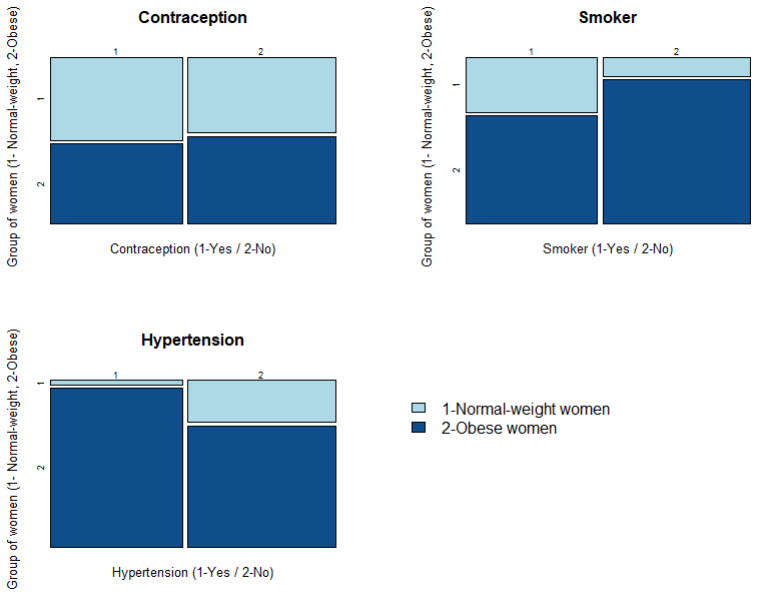

Figure 3.2: Contraception, Smoker and Hyperthension `mosaicplot` for each Group of women

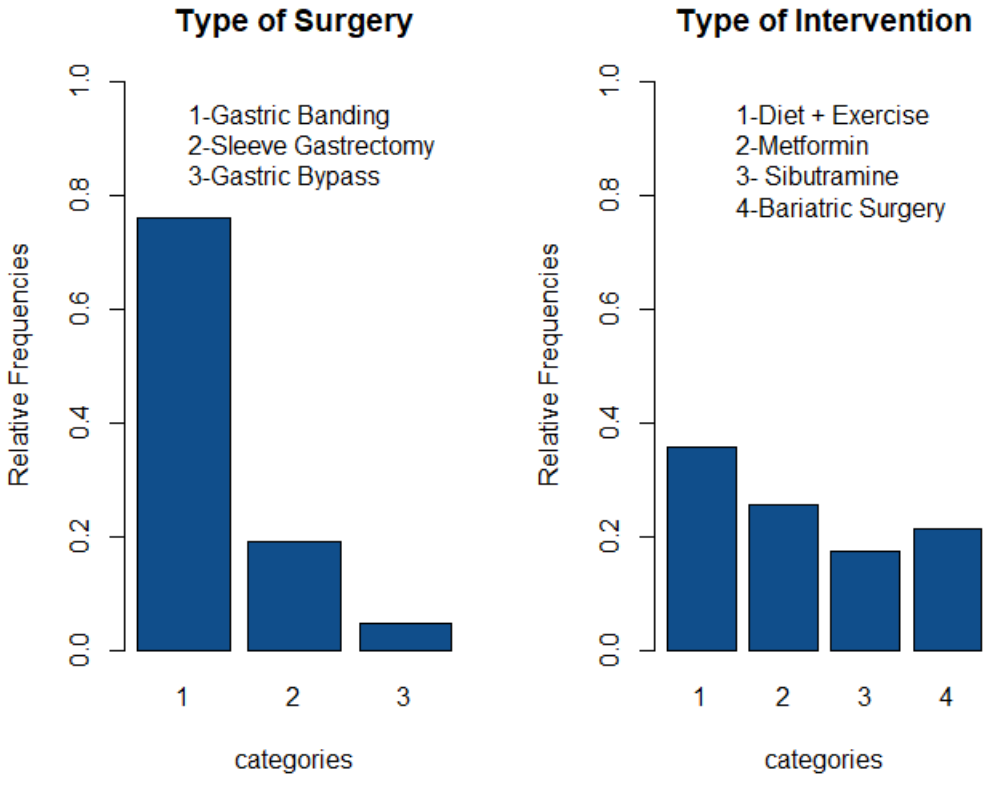

Figure 3.3: Type of Surgery and Type of Intervention Bar Diagrams in Obese Women who had surgery

In general, through the Figure 3.2, we can see that among women who do not take contraceptives, just over half are obese. Among women taking contraceptives, the number is even.
Regarding the Smoker variable, among women who smoke, most are obese. Concerning non-smokers, almost all the women are obese.
According to the hypertension, among women who have hypertension almost all the women are obese, but among the women who do not have hypertension, about three quarters are obese.
According to the Figure 3.3, we can find the several types of surgeries that obese women were subjected, which the most commonly surgery used was Gastric Banding and the least surgery used was Gastric Bypass.
This database contains 19 SNPs (categorical variables), which will be further analyzed, and only 9 quantitative variables. So, a Box Plot analysis of which of them was performed to observe the dispersion of the data. All the information of each SNP, ie, the number of genotypic and allelic observations for both groups (case / control) are present in the tables, on the Appendix. This information are subdivided by the name of each SNP.

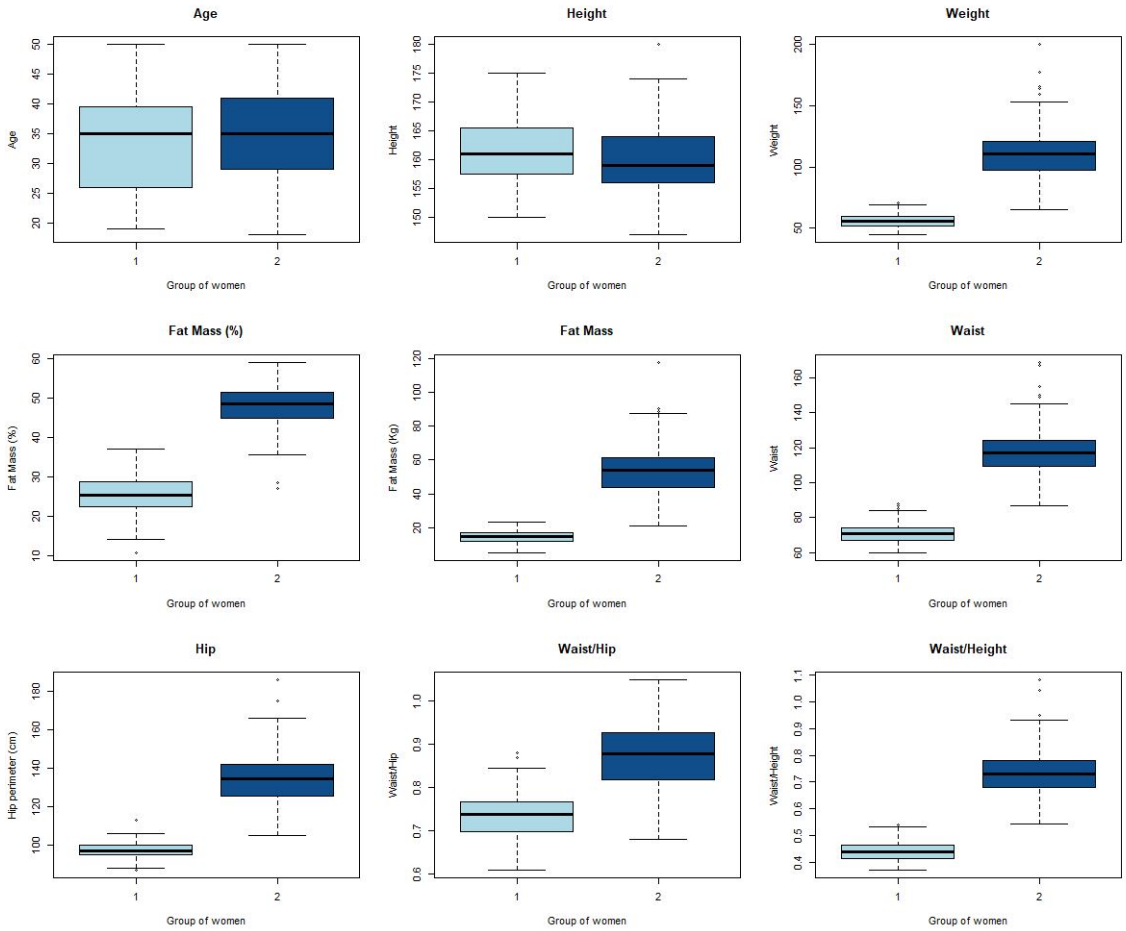

Figure 3.4: Box Plot - Age, Height, Weight, Fat Mass (%), Fat Mass, Waist, Hip, Waist/Hip, Waist/ Height

Considering that all variables have different scales, based on the previous graph (Figure 3.4), we can see that all variables seem to present asymmetry, except for the variables, *Waist/hip*

ratio and *Waist/Height ratio* for both groups of women, because the median value coincides with the mean value, as we can observe in Figure 4.1 (Appendix). On these two Box Plot we can see 2 possible outlier candidates in the group of normal weight, for variable *Waist/Hip*. For the variable *Waist/Height* there are 3 possible outlier candidates in the obese group and only 1 possible outlier candidate in the normal weight group.
For the *Age* variable, the dispersion is almost the same for both groups of women. The Box Plot for normal weight women seems to have a negative asymmetry, while the Box Plot for obese women has a slight positive asymmetry. In the Box Plot of the variable *Height*, we note that obese women have lower heights compared to women in the normal weight group. We can also observe that there is a possible outlier candidate in the group of obese women, that is the value 180. Regarding *Weight*, it is noted that data dispersion is much greater in the obese group than in the normal weight group, as expected, since the weights in the normal weight group are much more moderate than in the obese group. We can see, on the normal weight group a possible outlier candidate. However, on the group of obese women there are 5 possible outlier candidates. In the remaining graphs, i.e, for the variables *Fat Mass (%)*, *Waist* and *Hip*, respectively, there are several possible outlier candidates in both groups, as already expected, due to the women weight differences that exist in this database. Regarding the data dispersion, through table (Figure 4.1), we verify that all variables have some variability because all the value of the standard deviations are different to zero. Nevertheless, the *Waist/Hip* has a standard deviation of 0.07 in the group of obese and 0.05 in the normal weight group, whereas the variable *Waist/Height* has a standard deviation of 0.1 in the obese group and 0.04 in the normal weight group.

3.2 Hardy-Weinberg equilibrium Test

When studying the genetics of a population, one of the first questions that may be of interest is whether the genotype frequencies fit Hardy-Weinberg (HW) expectations. Therefore, statistical tests for Hardy-Weinberg Equilibrium (HWE) are important tools in genetic data analysis (Graffelman et al., 2016). The genotype frequencies will fit HW if the population is behaving like a single randomly mating unit without intense viability selection acting on the sampled *loci*. Besides, testing for HW proportions is often used for quality control in genotyping, as the test is sensitive to misclassifications or undetected null alleles. Traditionally, geneticists have relied on test statistics with asymptotic χ^2 distributions to test for goodness-of-fit with respect to HW proportions. However, as pointed out by several authors (e.g: Rohlfs et al., 2008; Shriner, 2011) these asymptotic tests quickly become unreliable when samples are small or when rare alleles are involved. The latter situation is increasingly common as techniques for detecting large numbers of alleles become widely used.
In this study, the `HWChisq` command from *HardyWeinberg* library on `RStudio` was used to test the goodness-of-fit with respect to HW proportions, on each SPN, i.e, to perform an exact test for HWE (Graffelmane et al., 2016). In other words, the null hypothesis is that the population is at HWE.
There are SNPs that are in the same gene, so in order to decrease the rate of false discoveries, the Table 3.2 shows the *p-values* for each SNP for HW Exact test, with and without the Benjamini-Hochberg (BH) correction. Considering the corrected *p-values*, all the studied

polymorphisms are in HWE except LepG-2548A SNP and MC4R_rs_17782313 SNP. These SNPs are statistically significant, considering a significance level of 5%, due to heterozygotes excess for the LepG-2548A SNP and due to heterozygotes deficit present on MC4R rs 17782313 SNP. When a population does not meet the assumptions, may mean that there was a mutation, the population size is small, mating was not random or there was natural selection. Migration and genetic drift also affect this balance.
In this data base, the genotype table of LepG-2548A SNP (Table 4.37 - Appendix) has one cell without values, i.e, there are no women with geotype GG. So, this could be the reason for the LepG-2548A SNP not being in HWE.

Polymorphisms	**HWE (*p-value*)**	**HWE (corrected *p-value*)**
PON_1_Q192R	0.7418	1.0000
PON_1_M55L	0.2450	0.7350
AdipoQG	0.8620	1.0000
AdipoQ_G11377C	0.1023	0.4604
AdipoQ_G11391A	0.7722	1.0000
AdipoQ_45T_G	0.7722	1.0000
FTO_A_T	0.0357	0.2142
PPARG_Pro12Ala	1.0000	1.0000
ApoA5_T1131C	0.2108	0.7350
ACE_I_D	0.4210	1.0000
IL_6_G572C	1.0000	1.0000
TNFa_G308A	1.0000	1.0000
Leptin_G2548A	<0.0001	<0.0001
LeptinR_K109R	0.5295	1.0000
Ghrelin_Leu72Met	1.0000	1.0000
MC4R_V103I	1.0000	1.0000
MC4R_rs17782313	<0.0001	<0.0001
TCF7L2_rs7903146_C_T	0.6093	1.0000

Table 3.2: HWE of each SNP (Pearson's Chi-squared (χ^2)) - *p-values* with and without Benjamini-Hochberg correction

3.3 Likelihood of suffering from the disease in each SPN

Based on Oliveira (1996) when we are dealing with databases that are genetically heterogeneous, the first task is to determine the gene frequencies at each locus of interest. The frequency of each allele in a population is determined by the proportion of chromosomes containing that allele. Since this determination is made in diploid gametes, each individual has two chromosomes containing the locus. Thus, there is a total of $2N$ chromosomes to consider for a number of N individuals and the genetic frequency for a given allele A, for example, is given by the formulas:

$$p = f(\mathrm{A}) = \frac{2 \times \mathrm{obs(AA)} + \mathrm{obs(Aa)}}{2\mathrm{N}} \tag{3.1}$$

and

$$q = f(\text{a}) = \frac{2 \times \text{obs(aa)} + \text{obs(Aa)}}{2\text{N}} \tag{3.2}$$

where obs(AA), obs(aa) and obs(Aa) represent the number of women with a given genotype, for example AA/aa/Aa, for each SNP, in each group (case/ control).
The allele frequencies were obtained for each SNP, for both groups (case/control) (Table 3.3). Within the 19 SNPs, the allelic frequency of PON_1_Q192R , PON_1_M55L, AdipoQ_G11377C, PPARG_Pro12Ala, ApoA5_T1131C and TNFa_G308A, do not differ much between case and control groups. The non-allelic difference may be a disadvantage when investigating a risk allele, as the values are very identical in both alleles. So, these values may influence the fact whether or not they are associated with the disease, when the association tests will be applied. In opposite, TCF7L2_rs7903146_C_T is the SNPs which have major differences in the allelic frequency level between the two groups, i.e, have a great allelic diversity, from one group to the other. The allele C, in obese group, has a large allelic difference (f(C)=0.402) compared to the allelic frequency in the control group (f(C)=0.293). The allele T is the allele with the highest frequency, in both groups. However, it has a higher allele frequency in the control group (f(T)=0.707). This difference in allele frequencies, may justify a risk of obesity if, throughout the work, the allele T is considered a risk allele.

Polymorphisms	Frequencies			
	Case (Obese)		Control (Normal Weight)	
PON_1_Q192R	f(Q) = 0.311	f(R) = 0.689	f(Q) = 0.317	f(R) = 0.683
PON_1_M55L	f(L) = 0.352	f(M) = 0.648	f(L) = 0.273	f(M) = 0.727
AdipoQG	f(G) = 0.369	f(T) = 0.631	f(G) = 0.305	f(T) = 0.695
AdipoQ_G11377C	f(C) = 0.409	f(G) = 0.591	f(C) = 0.383	f(G) = 0.617
AdipoQ_G11391A	f(G) = 0.450	f(A) = 0.550	f(G) = 0.402	f(A) = 0.598
AdipoQ_45T_G	f(T) =0.469	f(G) = 0.531	f(T) = 0.397	f(G) = 0.603
FTO_A_T	f(A) = 0.262	f(T) = 0.738	f(A) = 0.205	f(T) = 0.795
PPARG_Pro12Ala	f(C) = 0.468	f (G) = 0.532	f(C) = 0.434	f(G) = 0.566
ApoA5_T1131C	f(A) = 0.475	f(G) = 0.525	f(A) = 0.458	f(G) = 0.542
ACE_I_D	f(D) = 0.443	f(I) = 0.557	f(D)= 0.283	f(I) = 0.717
IL_6_G572C	f(C) = 0.502	f(G) = 0.498	f(C) = 0.445	f(G) = 0.555
TNFa_G308A	f(G) = 0.458	f(A) = 0.542	f(G) =0.433	f(A) = 0.567
Leptin_G2548A	f(A) = 0.320	f(G) = 0.680	f(A) =0.276	f(G) =0.724
LeptinR_K109R	f (K) = 0.412	f (R) = 0.588	f(K) =0.351	f(R) =0.649
Ghrelin_R51Q	f(R) = 0.541	f(Q) = 0.459	f(R) =0.401	f(Q) =0.599
Ghrelin_Leu72Met	f(L) = 0.477	f(M) = 0.523	f(L) = 0.437	f(M) =0.563
MC4R_V103I	f(I) = 0.481	f(V) = 0.519	f(I) = 0.458	f(V) =0,542
MC4R_rs17782313	f(T) = 0.396	f(C) = 0.604	f(T) =0.355	f(C) = 0.645
TCF7L2_rs7903146_C_T	f (C) = 0.402	f(T) = 0.598	f(C) = 0.293	f(T) = 0.707

Table 3.3: SNPs - Allele Frequencies (Case and Control Groups)

3.4 Odds Ratio

A case-control study compares the prevalence of a specific disease among individuals with normal alleles and individuals with variant alleles, which generates an odds ratio (OR). The most common type of allele variation, single nucleotide polymorphism, consists of a major allele (for example, A) and a minor allele (a). The genotype can be a major allele homozygote (AA - Dominant), a heterozygote (Aa) or a minor allele homozygote (aa - Recessive). There are several types of genetic models with different effects, for example, the dominant model, the recessive model, the over-dominant model that assumes the heterozygote has the strongest impact and the co-dominant models including additive and multiplicative models. However, only two of them will be studied here. A dominant model compares AA vs. Aa+aa and a recessive model compares aa vs. AA+Aa. According to Horita et al. (2015), researchers used to calculate ORs using many models and then select the best model according to the obtained ORs. This may increase the possibility of type I error due to multiple comparisons (Bagos, 2013).

Allelic and Genotypic ORs

In this case control study, for each SNP a uppercase letter will be used to a certain allele present in that polymorphism. So, we start by introducing the Odds Ratio with the notation for the allello "A" and for the allele "B".
Allelic OR describes the association between disease and allele by comparing the odds of disease in an individual carrying allele A with the odds of disease in an individual carrying allele B. Genotypic ORs describe the association between disease and genotype by comparing the odds of disease in an individual carrying a genotype with the odds of disease in an individual carrying another genotype. For this reason, there are usually two genotypic ORs, one comparing the odds of disease between individuals carrying genotype AA and those carrying BB, and the other comparing the odds of disease between individuals carrying genotype AB and those carrying genotype BB (Clarke et al. 2011).
In order to understand if there is a genetic association between one of the genetic models and obesity it is necessary to test the dominant and recessive models for each SNP. If the alleles of the gene of interest are A and B in haploid cells, and A is the "increasing"/ "risk" allele, i.e, the one causing an effect, the three genotype groups would then be AA, AB and BB. This dichotomization of the SNP genotypes can be done as follows:

- ◇ Dominant: "AA + AB" vs. "BB",
- ◇ Recessive: "AA" vs. "AB + BB".

By performing the contingency tables for the recessive and dominant models, it was observed that there were cells without observations. Therefore, it was decided that only one (dominant or recessive) genotypic tests would be performed if any of the cells in the genotype count case/control table contains 0 observations, i.e, missing values.
In each SNP there is an imbalance between the case/control groups, because the number of observations in each cells is quite distinct, i.e, the number of observations is very high in one group compared to the other group that contains values close to zero. Due to this, that may exist genetic influence.
The genotypic OR were calculated for both dichotomization of each SNP genotypes and all the

ORs greater than 1 were marked in the Table 3.4. This means that there is a higher odds of obesity happening with exposure to the risk allele. In other words, the odds of exposure among cases is greater than the odds of exposure among controls.

Polymorphisms	Models (Dominant and Recessive)	ORs	CI	Fisher's Test (*p-value*)
PON_1_Q192R	QQ vs. QR+RR	0.448	(0.23843; 0.84252)	0.0170
	QQ+QR vs.RR	0.454	(0.17751; 1.16192)	0.1212
PON_1_M55L	LL vs. LM+MM	0.859	(0.46078; 1.60170)	0.6376
	LL+LM vs. MM	2.006	(0.86718; 4.64163)	0.1379
AdipoQG	GG vs. GT+TT	0.863	(0.47121; 1.57929)	0.6467
	GG+GT vs. TT	0.562	(0.20293; 1.55570)	0.3303
AdipoQ_G11377C	CC vs. CG+GG	0.283	(0.14163; 0.56681)	0.0003
	CC+CG vs. GG	0.716	(0.20167; 2.54412)	0.7574
AdipoQ_G11391A	GG vs. GA+AA	0.696	(0.36517 ; 1.32664)	0.3301
	GG+GA vs. AA	0.581	(0.05182; 6.52327)	0.6564
AdipoQ_45T_G	TT vs. TG+GG	1.587	(0.78676; 3.20087)	0.2161
	TT+TG vs. GG	1.514	(0.32909; 6.96896)	0.7087
FTO_A_T	AA vs. AT+TT	2.266	(1.06775; 4.80737)	0.0446
	AA+AT vs. TT	1.337	(0.70502; 2.53531)	0.4174
PPARG_Pro12Ala	CC vs. CG+GG	1.394	(0.68630; 2.83010)	0.3743
ApoA5_T1131C	AA vs. AG+GG	0.563	(0.23616; 1.33982)	0.2069
	AA+AG vs. GG	1.072	(0.06614; 17.37896)	0.9998
ACE_I_D	DD vs. ID+II	1.742	(0.91932; 3.30063)	0.1060
	DD+ID vs. II	2.638	(0.84110; 8.27333)	0.0978
IL_6_G572C	CC vs. GC+GG	0.936	(0.38587 ; 2.27207)	0.9999
TNFa_G308A	GG vs. GA+AA	0.571	(0.28815 ; 1.13319)	0.1276
Leptin_G2548A	AA vs. AG+GG	0.584	(0.28104; 1.21512)	0.1903
LeptinR_K109R	KK vs. KR+RR	0.693	(0.37305; 1.28853)	0.2748
	KK+KR vs. RR	0.458	(0.11729; 1.79106)	0.3510
Ghrelin_R51Q	RR vs. QR+QQ	1.905	(0.73402; 4.94278)	0.2324
Ghrelin_Leu72Met	LL vs. LM+MM	0.6	(0.27711 ; 1.29914)	0.2514
MC4R_V103I	II vs. VI+VV	0.749	(0.28371; 1.97732)	0.6291
MC4R_rs17782313	TT vs. CT+CC	0.949	(0.53316; 1.68828)	0.8841
	TT+CT vs. CC	0.609	(0.26605; 1.39448)	0.3078
TCF7L2_rs7903146_C_T	CC vs. CT+TT	1.312	(0.73566; 2.33845)	0.3804
	CC+CT vs. TT	1.184	(0.45746; 3.06214)	0.8091

Table 3.4: Recessive and Dominant Models for each SNP - ORs

In the Table 3.5, we can see all the *p-value* with and without BH correction. The Benjamini correction was applied to each SNP pair of models, separately, because adjusts probability values due to increased risk of a type I error, when making multiple statistical tests for the same SNP. The Adaptive Group Benjamini Hochberg was tried, which corrects all *p-values* while maintaining these groups, but did not work.

Polymorphisms	**Models (Dominant and Recessive)**	**Fisher's Test (*p-value*)**	**Fisher's Test (*p-value* adj.)**
PON_1_Q192R	QQ vs. QR+RR	0.0170	0.0340
	QQ+QR vs.RR	0.1212	0.2424
PON_1_M55L	LL vs. LM+MM	0.376	1.0000
	LL+LM vs. MM	0.379	0.2758
AdipoQG	GG vs. GT+TT	0.6467	1.0000
	GG+GT vs. TT	0.3303	0.6606
AdipoQ_G11377C	CC vs. CG+GG	0.0003	0.0006
	CC+CG vs. GG	0.7574	1.0000
AdipoQ_G11391A	GG vs. GA+AA	0.3301	0.6602
	GG+GA vs. AA	0.6564	1.0000
AdipoQ_45T_G	TT vs. TG+GG	0.2161	0.4322
	TT+TG vs. GG	0.7087	1.0000
FTO_A_T	AA vs. AT+TT	0.0446	0.0892
	AA+AT vs. TT	0.4174	0.8348
ApoA5_T1131C	AA vs. AG+GG	0.2069	0.4138
	AA+AG vs. GG	0.9998	1.0000
ACE_I_D	DD vs. ID+II	0.1060	0.2120
	DD+ID vs. II	0.0978	0.1956
LeptinR_K109R	KK vs. KR+RR	0.2748	0.5496
	KK+KR vs. RR	0.3510	0.7020
MC4R_rs17782313	TT vs. CT+CC	0.8841	1.0000
	TT+CT vs. CC	0.3078	0.6156
TCF7L2_rs7903146_C_T	CC vs. CT+TT	0.3804	0.7608
	CC+CT vs. TT	0.8091	1.0000

Table 3.5: Recessive and Dominant Models for each SNP - *p-values* with and without Benjamini-Hochberg correction

Considering the Table 3.5 and significance level of 10%, within the 19 SNPs, only 4 SNPs, i.e, PON_1_Q192R, AdipoQ_G11377C, FTO_A_T, ACE_I_D are statistically significant. The PON_1_Q192R, in the model QQ vs QR+RR ($p = 0.017 < 0.10$; p.adj$= 0.0340 < 0.10$), OR $= 0.448 < 1$, that means that the exposure (QQ genotype) is associated with lower odds of obesity, because the chance of being obese in presence of the QQ genotype is a bit less than the chance of being obese when this genotype is not present. However, through the *p-value* the allele Q is the recessive allele for non-obesity, and R is the dominant allele for obesity.

The model CC vs CG + GG, in AdipoQ_G11377C, (p = 0.0003 < 0.10 ; p.adj = 0.0006 < 0.10), has an OR = 0.283 < 1, it means that the CC genotype is associated with lower odds of obesity, because the chance of being obese in the presence of the genotype CC is less than the chance of being obese when this genotype is not present. So, C is the recessive allele for non-obesity, and G is the dominant allele for obesity. The FTO_A_T SNP, in the model AA vs AT + TT (p = 0.0446 < 0.10 ; p.adj = 0.0892 < 0.10), the OR is equal to 2.266, i.e, greater than 1. For this reason, the odds of exposure (AA) among the cases is greater than the odds of exposure to the AA genotype among controls. So, A is the recessive allele for obesity and T is the dominant allele for non-obesity.
In the SNP ACE_I_D, the model DD + ID vs II (p = 0.0978 < 0.10 ; p.adj = 0.1956 > 0.10), OR = 2.638> 1, has a significant *p-value* only without BH correction. However, since this SNP has been shown to be important in the study of obesity, I decided to include it in the conclusions taking from this analysis. Since the OR is greater than one, it means that there is a higher odds of obesity happening with exposure to the II genotype, because the chance of being obese in the presence of these genotypes is greater than the chance of being obese when these genotypes are not present. So, D is the dominant allele for obesity and I is the recessive allele for non-obesity. Note that the *p-value* is very close to 0.10 (if we consider the adjusted *p-value*, this polymorphism is no longer significant).
After that, the allelic odds ratios were calculated for each SNP and all the *p-values* proved to be significant.

Polymorphisms	**OR**	**CI**	χ^2**obs.**	χ^2***(p-value)***	χ^2***(p-value adj.)***
PON 1 Q192R	0.54	(0.28, 1.03)	3.541	0.06	0.08
AdipoQ G11377C	0.39	(0.22, 0.70)	10.639	<0.01	0.04
FTO A T	1.41	(0.95, 2.08)	2.977	0.08	0.08
ACE I D	1.71	(1.04, 2.83)	4.509	0.03	0.06

Table 3.6: Allelic OR for each significant SNP

We can see, through the Table 3.6, that only FTO_A_T (OR = 1.41) and ACE_I_D (OR=1.71) have an odds ratio greater than 1. For example, for FTO_A_T the women with the A allele have 1.41 chances to have obesity. However, the women with alelle D have 1.71 chances to suffer from obesity. Otherwise, the women that have the allele C in the AdipoQ_G11377C SNP (OR=0.39),is associated with lower odds of obesity. Also, the chance of not being obese are 2.6 (1/0.39 = 2.564) times more than being obese.

3.5 Tests for Association

The tests of genetic association compares the frequencies of alleles and genotypes at genetic marker *loci*, usually single-nucleotide polymorphisms (SNPs), in individuals from a given population, with and without a given disease trait, in order to determine if there is a statistical association between the disease trait and the genetic marker. The data for each SNP with minor allele a and major allele A can be represented as a contingency table of counts of disease status by either genotype count (e.g., aa, Aa and AA) (Table 2.2, section 2.5.1) or allele count (e.g., a and A). Under the null hypothesis of no genetic association with the disease, we expect the

relative allele or genotype frequencies to be the same in case and control groups. To test the genetic association between the alleles of each SNP and the obesity, i.e., to test if they are statistically significant the Pearson Chi-square Test was applied and the results can be observed in the Table 3.7.

Polymorphisms	χ^2 **obs.**	χ^2(***p-value***)	χ^2(***p-value Adj.***)
PON_1_Q192R	2.9614	0.0853	0.1559
PON_1_M55L	0.76704	0.3811	0.3811
AdipoQ_G11377C	9.7812	0.0018	0.0108
FTO_A_T	2.6442	0.1039	0.1559
ACE_I_D	3.982	0.0460	0.1380

Table 3.7: Allelic Association Test - χ^2 Test - *p-values* with and without Benjamini-Hochberg correction

Through the *p-values* (without BH correction), the PON_1_Q192R (p=0.0853), the AdipoQ_G11377C (p=0.0018) and ACE_I_D (p=0.0460) are statistically significant considering the level of 10% of significance. However, with BH correction, i.e., through the *p-values* adjusted, only the AdipoQ_G11377C (p = 0.0108) is statistically significant. So, as the null hypothesis is rejected only for this SNP, we can say that there is a genetic association between the allele and the disease, where the allele G seems to manifest in a dominant way.
Oppositely, the remaining SNPs are not statistically significant, for the Allelic Association Tests, because all the *p-value* are higher than the 10% significance level. For this reason, the allele of each SNP is not associated with the obesity.
In the previous section, the OR values calculated are not much higher than 1, so the *p-values* are accordingly.
One of the inherent problems is the fact there is an imbalance in the number of observations for each allele, in each SNP, i.e, sometimes there are cells with values zero in an allele contingency table and anothers values between groups of case/control are uneven.
The CATT test is sensitive to the linearity between independent variable (e.g.: Group case/controls) and dependent variables (e.g: Genotype/ Alleles) and detects trends that would not be noticed by more crude methods, that is, for example, the Pearson Chi-Squared Test. In a first literary review, the method to be used would be the Cochran Armitage Test to test whether there is an association between disease and genotype. However, the SNPs, in our database, are not ordinal variables, i.e, an explanatory variable without ordered levels. For this reason and given all the previous analysis it was possible to establish a *score* for each Genotype of each SNP that proved to be significant. The special choice was (w_0, w_1, w_2) = (0,1,2) that represents the presence of the risk allele.

Polymorphisms	CATT (Zobs)	CATT *(p-value)*	CATT (*p-value*) Adj.
PON 1 Q192R	2.628	0.0086	0.0301
PON 1 M55L	-0.521	0.6021	0.6021
AdipoQ G11377C	3.07	0.0021	0.0147
FTO A T	-1.873	0.0610	0.1068
ApoA5 T1131C	1.16	0.2461	0.2871
ACE I D	-2.052	0.0402	0.0938
LeptinR K109R	1.385	0.166	0.2324

Table 3.8: Genotype Association Test - CATT(Z)- *p-value* with and without Benjamini-Hochberg correction

In the Tale 3.8, we can see the results obtained in CATT. Observing the *p-values* with BH correction, the only SNPs with *p-values* below than the usual significance levels, especially at the 10% significance level are PON_1_Q192R ($p = 0.0301$), AdipoQ_G11377C ($p = 0.0147$), FTO_A_T ($p = 0.1068$) and ACE_I_D ($p = 0.0938$). So, the H_0 is rejected for these 4 SNPs, which concludes that there is a genetic association between obesity and the genotype of each SNP.

3.6 Genetic Risk Score

3.6.1 The Multiple Logistic Regression Models of Association

When there is a need to include additional covariates to handle complex traits, more complicated logistic regression models of association are used. Examples of this are situations in which we expect disease risk to be modified by environmental effects such as epidemiological risk factors (e.g., smoking and gender), clinical variables (e.g., disease severity and age at onset) and population stratification (e.g., principal components capturing variation due to differential ancestry), or by the interactive and joint effects of other marker *loci*. In Logistic Regression Models, the logarithm of the odds of disease is the response variable, with linear (additive) combinations of the explanatory variables (genotype variables and any covariates) entering into the model as its predictors (Clarke et al., 2011).
Despite the results obtained previously, we decided to include all SNPs in order to understand if we obtain the same results through the logistic regression. So, the explanatory variables that are used are all the SNPs with 2 categories (Recessive and Dominant models) and the independent variable is the women's group (normal weight and obese).
Although the selection is performed partly by software and partly by hand, the stepwise and best subset approaches are automatically performed by the software.
First, a visual analysis of the missing values might be helpful to check for missing values and look how many unique values there are for each variable using.
One of the most complex problems in data analysis is that of missing values occurrence. It may happen in several situations as consequence of different causes such as the type of study, the sampling procedures and the goal of the inquiry. The `Amelia` package has a special plotting function `missmap` that will plot our dataset and highlight missing values (Figure 3.5).

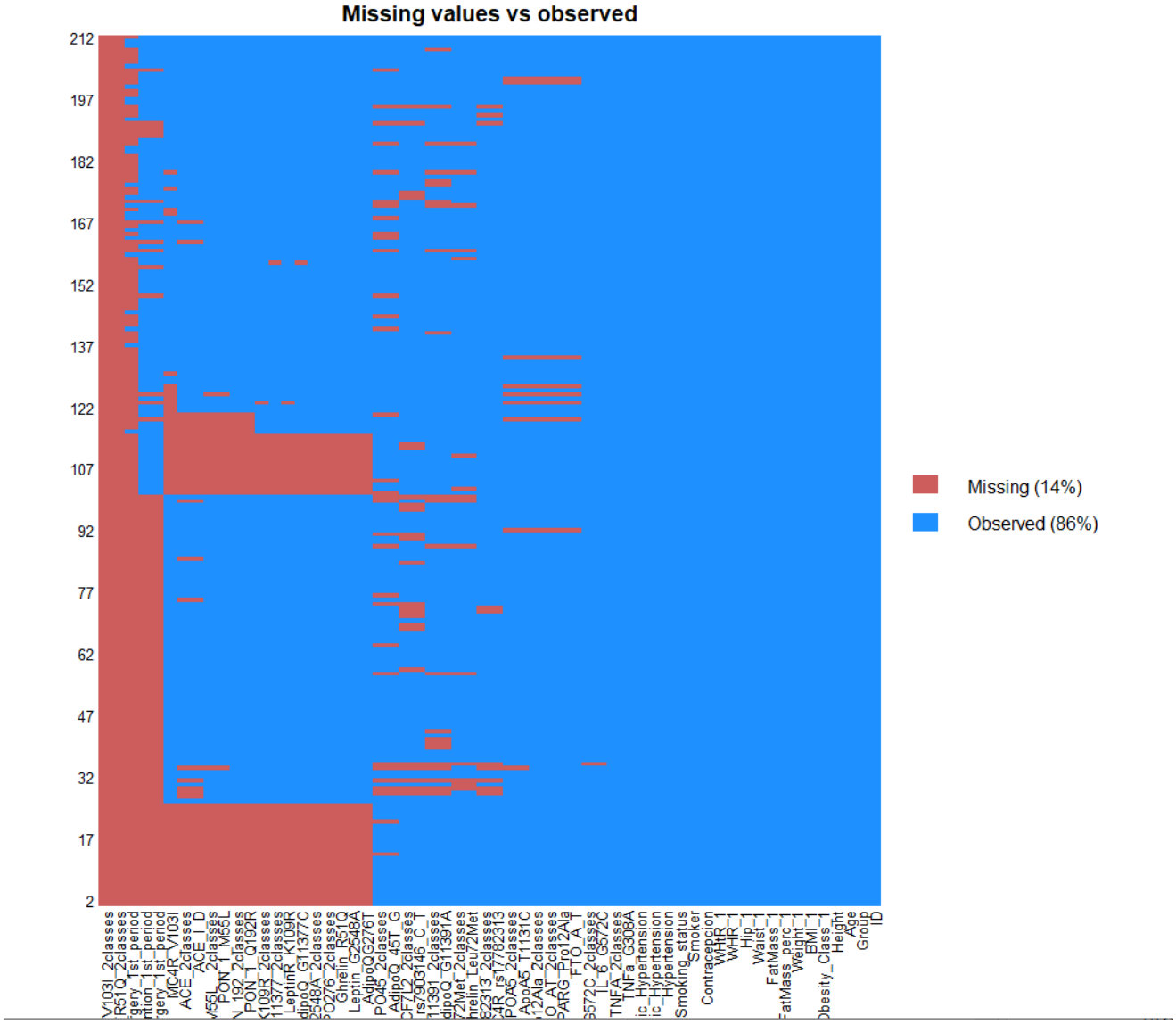

Figure 3.5: Graphical Representation of Missing Values (NA)

There are many variables with missing values, which are coded as NA. For this reason, we will build manually the logistic regression model.

To build the model by hand, we can choose 2 main approaches, i.e., the forward selection and the backward elimination, using two models fit of criterion. The model can be chosen through the *p-value* criterion or through the AIC criterion. Thus, both models of criterion were considered, for forward and backward methods.

Despite the terms used in both methods (forward and backward), we always obtained the same final model and, for this reason we chose to comment only the model obtained by the forward method.

To build the logistic regression model it was created a model only with an independent variable (Group), i.e., without explanatory variables, in order to build a model using the forward method. After that, we started to build the regression model based on the *p-value*, adding, one by one, the SNP (if any) whose inclusion gives the most statistically significant fit improvement, and repeating this process until none improves the model to a statistically significant extent. The

final model obtained can be found in the Figure 4.21, in Appendix. Based on the `output`, we can observe that all SNPs included (AdipoQ_G11377C_SNP (p= 0.0026), FTO_A_T_SNP (p= 0.0158) and PON_1_Q192R_SNP (p= 0.0178) in the model are statistically significant at the level of significance of 5%, with an AIC = 203.6. Comparatively, with the final model obtained through the AIC criterion (Figure 4.23 - Appendix), we have a much smaller AIC value (AIC = 189.28) than the model obtained previously. However, there is a SNP that is not statistically significant (ACE_I_D (p= 0.1987)). Besides that, the first model (obtained through the *p-value* criterion), has less missing values that have been removed (53 observations deleted), than the last model (63 observations deleted). For this reason and as we want the most parsimonious model, i.e., the model that involves the minimum of possible parameters to be estimated and that explains the behavior of the response variable well, the model we chose is the model obtained through the stepwise forward method, through the *p-value* criterion:

$$\text{Group\_fator} \sim AdipoQ\_G11377C\_SNP + FTO\_A\_T\_SNP + PON\_1\_Q192R\_SNP$$

With this removal, the database consists of 159 observations, being composed by 85 obese woman and 74 women with normal weight.
In order to diagnose the multicollinearity of the model, the *Variance Inflation Factor* (VIF) was calculated for the saturated model. According to Hair et al. (1962), large VIF values indicate a high degree of collinearity or multicollinearity among the independent variables and the suggested cutoff for the tolerance value correspond a VIF of 10.0.
In our model, all values are between 1.0030 and 1.0115 (Figure 4.22 - Appendix). Since the VIF has very low values, there are no multicollinearity problems.
The Hosmer and Lemeshow of Godness of Fit Test (HL) calculates if the observed event rates match the expected event rates in population subgroups. So, the HL was applied to understand the goodness-of-fit of the model and the results can be observed at Figure 4.25 - Appendix. Since the *p-value* is high ($p = 0.9129$), the model is considered to be well adjusted.
Finally, the ROC curve was constructed and the AUC calculated to understand if the model can distinguish between patients with disease and without disease, i.e. the discriminatory capacity of the model (Figure 3.6). As the xx and yy axis vary between 0 and 1, it was expected that the graph had the shape of a square. However, there is a graphical limitation since by placing a scale of 1.1 on the *latex* software, the image is stretched.

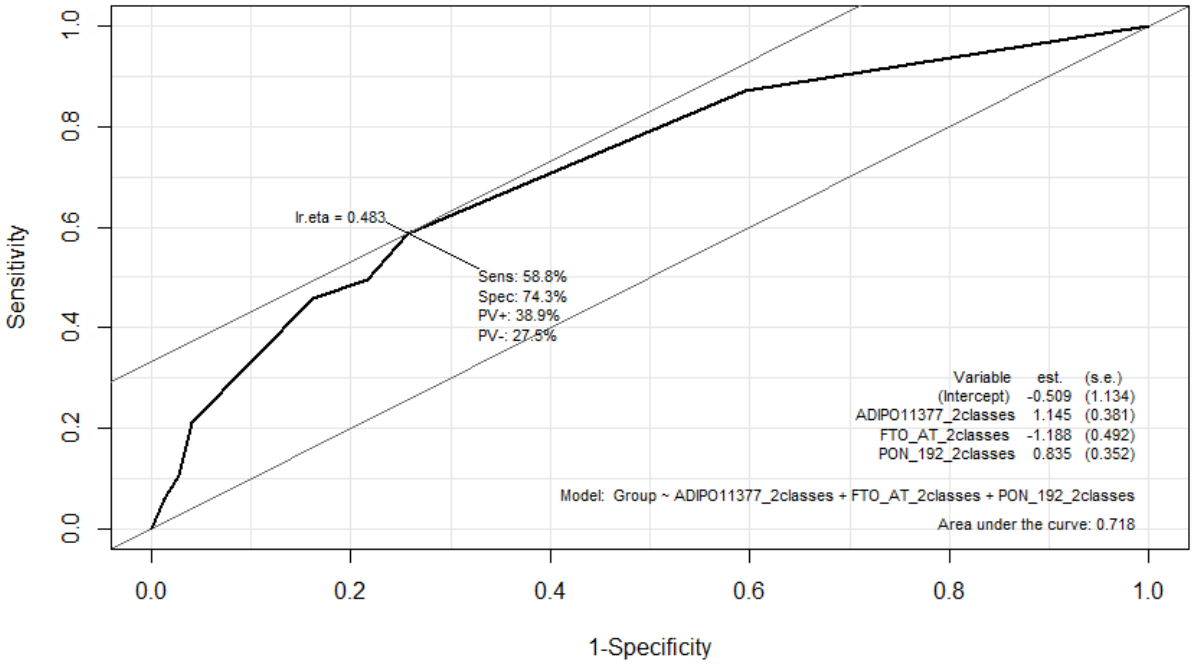

Figure 3.6: ROC Curve with AUC - `Epi` package, `ROC` function (R Version 1.1.463)

The AUC is a measure of discriminative ability across all possible thresholds. The value of AUC can be interpreted as the probability that a random individual who will develop obesity has a higher (genetic and/or non-genetic) obesity risk than a random individual who will not develop obesity (Loos et al., 2017).
The value obtained for the area under the curve (AUC) was 0.718, which according to Hosmer et al.(2000), indicates that the model has a moderate discriminatory capacity. It is also verified that it has a sensitivity value (58.8%) lower than the specificity value (74.3%), that means, it discriminates the true negatives (Y = 0) better than true positives (Y = 1), at the optimal cutoff point. However, we would expect a cutoff producing sensibility higher than specificity, identifying more than 58.8% of obese women.
In a future study with a larger sample, specificity and sensitivity values may be compared with the obtained values in order to understand how the collected sample may influence these values.
Therefore, the multiple logistic regression model that estimates the association of the obesity between the SNPs is:

$$\log\left(\frac{\pi}{1-\pi}\right) = 0.2837 + 1.1451 \times (AdipoQ\_G11377C\_SNP) - 1.1881 \times (FTO\_A\_T\_SNP) + 0.8353 \times (PON\_1\_Q192R\_SNP)$$

3.6.2 Estimated Risk Score Through Logistic Regression Model

The risk prediction models have included an increasing number of BMI-associated *loci*, typically combined into a genetic risk *score* (GRS). The genetic risk *score* represents the number of risk alleles across all included genetic variants (SNPs). A negative *score* indicates a protection against disease.
The higher the *score*, the higher the genetic susceptibility to becoming obese, according to Loos

et al., (2017). The risk *score* can be expressed as a weighted sum of an event Y = 1, given a vector with explanatory variables, X, containing the measures of the relevant risk factors:

$$E(\mathbf{Y}|\mathbf{X}) = P(Y = 1|\mathbf{X}) = g^{-1}(X^T\beta) \tag{3.3}$$

Where g^{-1} :IR$\rightarrow$[0,1], $i.e$,

$$Score = X^T\widehat{\beta} = \sum_{j=1}^{p} X_j\widehat{\beta}_j \tag{3.4}$$

Where $\widehat{\beta}_j$ represents the estimated weights obtained through the multiple logistic regression model, excluding β_0, and X_j represents the risk SNP_j. In such way, the genetic risk *score* (GRS) for the obese women was built based on a combination of 3 SNPs, AdipoQ_G11377C_SNP, FTO_A_T_SNP and PON_192_2_SNP, that increase the risk of obesity.

$Score = 1.1451 \times AdipoQ\_G11377C\_SNP - 1.1881 \times FTO\_A\_T\_SNP + 0.8353 \times PON\_192\_2\_SNP$

The baseline of genetic risk *score* is 0, which corresponds to an individual not having the risk allele. For this reason all the SNPs have the value 0, on the equation.
Through this expression we can understand which is the highest risk *score*. For that we just need to add a *score* (0,1,2) which corresponds to the presence of risk alleles (0- none risk allele, 1 - if 1 risk allele is present and 2 - if 2 risk alleles are present), for each SNP (AdipoQ_G11377C_SNP, FTO_A_T_SNP and PON_192_2_SNP). Through the Figure 4.26, in the Appendix, we can see some combinations for each SNP, regarding the number of risk alleles present for each SNP. For example, if a woman has all the risk alleles in each SNP, the genetic risk *score* will be equal to 1.58. On the other side, the lowest risk *score* is -2.38, it means to be carrying 2 risk alleles of the FTO_A_T_SNP, but not having any risk alleles corresponding to the other SNPs the risk of obesity decreases.
If a woman has two risk alleles for AdipoQ_G11377C_SNP and PON_192_2_SNP, but has a protective allele for FTO_A_T_SNP, the genetic risck *score* increases to 3.96.
In conclusion, we can understand that the genetic risk *score* varies greatly depending on the number of risk alleles that are carrying for each SNP.

Chapter 4

Discussion and Conclusion

After analyzing the database that was selected in Curry Cabral Hospital, in 2006, for a case-control study of obese women, several conclusions were reached.
Regarding the HWE, all the SNPs are in HWE except LepG-2548A SNP and MC4R rs 17782313 SNP. While MC4R rs 17782313 SNP has heterozygotes deficit, on LepG-2548A SNP, there are no women with genotype GG, in this database. So, this could be the reason for the LepG-2548A SNP are not in HWE.
Based on the calculation of the allele frequency for each SNP, the allelic frequency of PON_1_Q192R, PON_1_M55L, AdipoQ_G11377C, PPARG_Pro12Ala, ApoA5_T1131C and TNFa_G308A, do not differ much between case and control groups.
In opposite, the TCF7L2_rs7903146_C_T is the SNP which have a great allelic diversity between the groups (case/control). The allele differency of allele C is higher in the obese group compared to the control group. However, the T allele is the most frequent allele compared to the C allele frequency in both groups.
Looking to answer the biggest scientific question: "What are the polymorphisms associated with the obesity?", through genetic odds ratio, only 4 SNPs the PON_1_Q192R (For **Recessive Model** : p.adj= 0.0340< 0.05; OR = 0.448< 1), AdipoQ_G11377C (For the **Dominant Model**: p.adj = 0.0006< 0.05; OR = 0.283< 1), FTO_A_T (For the **Recessive Model**: p.adj = 0.0892< 0.10; OR= 2.266), ACE_I_D (For the **Dominant Model** : p.adj = 0.1956 > 0.10; OR= 2.638> 1) are statistically signicant.
In the SNP ACE_I_D, the model DD + ID vs II (p = 0.0978 < 0.10 ; p.adj = 0.1956 > 0.10), OR = 2.638> 1, has a significant *p-value* only without BH correction. The PON_1_Q192R and AdipoQ_G11377C have an OR less than 1, on both models (dominant and recessive) and in some studies related to PON_1_Q192R activity in obesity, no significant differences were found for PON1 activity between normal and obese women (Veiga et al., 2010).
According to Leoska et al. (2018), the ADIPOQ gene influence the effect of lifestyle on obesity-related traits. Nevertheless, the studies have reported inconsistent results, as occurs in this study, where the fact may be related to the type of population, gender, age, the degree of metabolic risk levels, and physical activity interactions.
According to the results of CATT for the Genotype Association Test, only the AdipoQ_G11377C, is statistically significant, where the allele G seems to manifest itself in a dominant way.
Oppositely, the other SNPs are not statistically significant, because all the *p-value* are higher than the 10% significance level. Nonetheless, the OR values calculated in the previous section are not much higher than 1, so the *p-values* are accordingly. One of the inherent problems is

the fact there is an imbalance in the number of observations for each allele, in each SNP, i.e, sometimes there are cells with zero values in an allele contingency table and anothers values between groups of case/control are very disparate.

After this, a parsimonious model was found and the genetic risk *score* was calculated through it. The genetic risk *score* (GRS) for the obese women was built based on a combination of 3 SNPs, AdipoQ_G11377C_SNP, FTO_A_T_SNP and PON_192_2_SNP. In agreement with AUC, the model that was found has a moderate discriminatory capaticy, but only identifies 58.8% of obese women. So, the model that estimates the association of the obesity between the SNPs may not be the best and, for this reason, the Logistic Regression may not be the best method to build a genetic risk *score*.

Since, this model was found only with 159 women, it is necessary additional studies with larger samples to clarify which polymorphisms will be associated with obesity. In future studies, other methods (e.g: Bayesian (Vilhjálmsson et al.,2015) or Naive methods (Ware et al., 2017)) can also be used to create a genetic risk *score*.

The genetic risk *score* was built and adding a *score* (0,1,2) which corresponds to the presence of risk alleles (0- none risk allele, 1 - if 1 risk allele is present and 2 - if 2 risk alleles are present) we can understand the susceptibility to become obese. For example, an women who has two risk alleles for AdipoQ_G11377C_SNP and PON_192_2_SNP, but has a protective allele for FTO_A_T_SNP, the genetic risk *score* increases to 3.96 and has the greatest genetic risk of suffering from obesity. Opposite, the lowest risk *score* is -2.38, i.e., a women carrying 2 risk alleles of the FTO_A_T_SNP, but not carrying any risk alleles of the others SNPs.

Throughout the statistical analysis several limitations were found. Many SNPs have not been shown to have alleles associated with the obesity and many of the women did not have certain genotypes of each SNP. So, a larger and more specific sample of this SNP will help contribute to possible future studies.

This work has contributed significantly to the genetic knowledge of obesity in Caucasian women and may help in future meta-analysis studies by clarifying which variants are actually associated with the tendency to develop an obese phenotype. It also provided a better understanding of the genetic diversity that is associated with obesity in the Portuguese population and, in further studies, they can be compared with other populations.

Bibliography

Abreu, D. (2013). *Síndrome Coronário Agudo: Análise do impacto das variáveis sócio-demográficas, ambientais e clínicas na demora média entre o início da sintomatologia e o restabelecimento do fluxo*. PhD thesis, Faculdade de Ciências da Universidade de Lisboa.

Agresti, A. (2007). *An introduction to categorical data analysis (2nd edition)*, volume 28. John wiley edition.

Albuquerque, D., Manco, L., González, L. M., Gervasini, G., Benito, G. M., González, J. R., and Rodríguez-López, R. (2017). Polymorphisms in the SNRPN gene are associated with obesity susceptibility in a Spanish population. *Journal of Gene Medicine*, 19(5).

Albuquerque, S. (2015). *Study of genetic variants associated with obesity in Portuguese children Study of genetic variants associated with obesity in Portuguese children Estudo de variantes genéticas associadas à obesidade em crianças de origem Portuguesa*. PhD thesis.

Alice, M. (2014). How to Perform a Logistic Regression in R. `https://datascienceplus.com/perform-logistic-regression-in-r/`. Last accessed on Sep 14,2019.

Bagos, P. G. (2013). Genetic model selection in genome-wide association studies: Robust methods and the use of meta-analysis. *Statistical Applications in Genetics and Molecular Biology*, 12(3):285–308.

Barness, L. A., Opitz, J. M., and Gilbert-Barness, E. (2007). Obesity: Genetic, molecular, and environmental aspects. In *American Journal of Medical Genetics*.

Bell, C. G., Walley, A. J., and Froguel, P. (2005). The genetics of human obesity.

Belsky, D. W., Moffitt, T. E., Sugden, K., Williams, B., Houts, R., McCarthy, J., and Caspi, A. (2013). Development and evaluation of a genetic risk score for obesity. *Biodemography and Social Biology*, 59(1):85–100.

Charlesworth, B. (2013). Population Genetics. *Encyclopedia of Biodiversity: Second Edition*, pages 182–198.

Chen, Z. and Ng, H. K. T. (2012). A robust method for testing association in genome-wide association studies. *Human Heredity*, 73(1):26–34.

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature Protocols*, 6(2):121–133.

Cochran, W. G. (1954). The Combination of Estimates from Different Experiments. *Biometrics*, 10(1):101.

Coll, A. P., Farooqi, I. S., Challis, B. G., Yeo, G. S., and O'Rahilly, S. (2004). Proopiomelanocortin and energy balance: Insights from human and murine genetics. In *Journal of Clinical Endocrinology and Metabolism*, volume 89, pages 2557–2562.

Cooke Bailey, J. N. and Igo, R. P. (2016). Genetic risk scores. *Current Protocols in Human Genetics*, 2016(October):1.29.1–1.29.9.

Eça, J. (2010). *Validação da associação entre SNPs e a Doença Coronária Validação da associação entre SNPs e a Doença Coronária.* PhD thesis, Faculdade de Ciências da Universidade de Lisboa.

El-Hani, C. N. (2007). Between the cross and the sword: The crisis of the gene concept.

Emily, M. (2018). Power comparison of Cochran-Armitage trend test against allelic and genotypic tests in large-scale case-control genetic association studies. *Statistical Methods in Medical Research*, 27(9):2657–2673.

Farooqi, I. S. and O'Rahilly, S. (2000). *The Genetics of Obesity in Humans.*

Farooqi, I. S. and O'Rahilly, S. (2005). Monogenic obesity in humans. *Annual review of medicine*, 56:443–58.

Farooqi, I. S. and O'Rahilly, S. (2014). Genetic obesity syndromes. In *The Genetics of Obesity.*

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R. B., Elliott, K. S., Lango, H., Rayner, N. W., Shields, B., Harries, L. W., Barrett, J. C., Ellard, S., Groves, C. J., Knight, B., Patch, A.-M., Ness, A. R., Ebrahim, S., Lawlor, D. A., Ring, S. M., Ben-Shlomo, Y., Jarvelin, M.-R., Sovio, U., Bennett, A. J., Melzer, D., Ferrucci, L., Loos, R. J. F., Barroso, I., Wareham, N. J., Karpe, F., Owen, K. R., Cardon, L. R., Walker, M., Hitman, G. A., Palmer, C. N. A., Doney, A. S. F., Morris, A. D., Smith, G. D., Hattersley, A. T., and McCarthy, M. I. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science (New York, N.Y.)*, 316(5826):889–94.

Gerken, T., Girard, C. A., Tung, Y. C. L., Webby, C. J., Saudek, V., Hewitson, K. S., Yeo, G. S., McDonough, M. A., Cunliffe, S., McNeill, L. A., Galvanovskis, J., Rorsman, P., Robins, P., Prieur, X., Coll, A. P., Ma, M., Jovanovic, Z., Farooqi, I. S., Sedgwick, B., Barroso, I., Lindahl, T., Ponting, C. P., Ashcroft, F. M., O'Rahilly, S., and Schofield, C. J. (2007). The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science*, 318(5855):1469–1472.

Ghodsi, M., Amiri, S., Hassani, H., and Ghodsi, Z. (2016). An enhanced version of Cochran-Armitage trend test for genome-wide association studies. *Meta Gene*, 9:225–229.

Graffelman, J. and Weir, B. S. (2016). Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome. *Heredity*, 116(6):558–568.

Graffelman, J. and Weir, B. S. (2018). Multi-allelic exact tests for HardyâWeinberg equilibrium that account for gender. *Molecular Ecology Resources*, 18(3):461–473.

Grens, K. (2019). Genetic Risk Score Developed for Obesity — The Scientist Magazine®.

Guglielmi, G. (2019). New genetic ârisk score' could predict obesity odds. *Science*.

Gutierrez-Aguilar, R., Kim, D.-H., Woods, S. C., and Seeley, R. J. (2012). Expression of new loci associated with obesity in diet-induced obese rats: from genetics to physiology. *Obesity (Silver Spring, Md.)*, 20(2):306–12.

Hair, J., William, B., and Babin, Barry and Anderson, R. (1962). *Multivariate Data Analysis*, volume 16. Seventh ed edition.

Haupt, A., Thamer, C., Heni, M., Ketterer, C., Machann, J., Schick, F., Machicao, F., Stefan, N., Claussen, C. D., Häring, H. U., Fritsche, A., and Staiger, H. (2010). Gene variants of TCF7L2 influence weight loss and body composition during lifestyle intervention in a population at risk for type 2 diabetes. *Diabetes*, 59(3):747–750.

Herrera, B. M. and Lindgren, C. M. (2010). The Genetics of Obesity. *Current Diabetes Reports*, 10(6):498–505.

Horita, N. and Kaneko, T. (2015). Genetic model selection for a case-control study and a meta-analysis. *Meta Gene*, 5:1–8.

Hosmer, David W., Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley-inte edition.

Huen, K., Yousefi, P., Street, K., Eskenazi, B., and Holland, N. (2015). PON1 as a model for integration of genetic, epigenetic, and expression data on candidate susceptibility genes. *Environmental Epigenetics*, 1(1).

Hung, C. F., Breen, G., Czamara, D., Corre, T., Wolf, C., Kloiber, S., Bergmann, S., Craddock, N., Gill, M., Holsboer, F., Jones, L., Jones, I., Korszun, A., Kutalik, Z., Lucae, S., Maier, W., Mors, O., Owen, M. J., Rice, J., Rietschel, M., Uher, R., Vollenweider, P., Waeber, G., Craig, I. W., Farmer, A. E., Lewis, C. M., Müller-Myhsok, B., Preisig, M., McGuffin, P., and Rivera, M. (2015). A genetic risk score combining 32 SNPs is associated with body mass index and improves obesity prediction in people with major depressive disorder. *BMC Medicine*, 13(1):1–10.

Huvenne, H. and Dubern, B. (2014). Monogenic Forms of Obesity. In *Molecular Mechanisms Underpinning the Development of Obesity*, pages 9–21. Springer International Publishing, Cham.

Huvenne, H., Dubern, B., Clément, K., and Poitou, C. (2016). Rare Genetic Forms of Obesity: Clinical Approach and Current Treatments in 2016.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434–446.

Jewell, N. P. (2004). *Statistics for epidemiology*. Chapman & Hall/CRC.

Karl, F. (2015). R - Studio , Package "allelic". pages 1–3.

Kpoghomou, M.-A., Soatiana, J. E., Kalembo, F. W., Bishwajit, G., and Sheng, W. (2013). UGT2B17 Polymorphism and Risk of Prostate Cancer: A Meta-Analysis. *ISRN Oncology*, 2013:1–7.

Krude, H., Biebermann, H., Schnabel, D., Tansek, M. Z., Theunissen, P., Mullis, P. E., and Grüters, A. (2003). Obesity due to proopiomelanocortin deficiency: three new cases and treatment trials with thyroid hormone and ACTH4-10. *The Journal of clinical endocrinology and metabolism*, 88(10):4633–40.

Lachance, J. (2016). Hardy-weinberg equilibrium and random mating. `https://www.sciencedirect.com/topics/neuroscience/hardy-weinberg-principle`. Last accessed on Sep 24, 2019.

Leońska-Duniec, A., Grzywacz, A., Jastrzebski, Z., Jazdzewska, A., Lulińska-Kuklik, E., Moska, W., Leźnicka, K., Ficek, K., Rzeszutko, A., Dornowski, M., and Cieszczyk, P. (2018). ADIPOQ polymorphisms are associated with changes in obesity-related traits in response to aerobic training programme in women. *Biology of Sport*, 35(2):165–173.

Loos, R. J. and Janssens, A. C. J. (2017). Predicting Polygenic Obesity Using Genetic Information. *Cell Metabolism*, 25(3):535–543.

Loos, R. J. F. (2012). Genetic determinants of common obesity and their value in prediction. *Best practice & research. Clinical endocrinology & metabolism*, 26(2):211–26.

M. Szumilas (2010). Explaining Odds Ratio. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19(3):227–9.

McClave, J. T., Dietrich, F. H., and Sincich, T. (1997). *Statistics*. Prentice Hall.

Mutch, D. M. and Clément, K. (2006). Unraveling the genetics of human obesity.

Narkhede, S. (2018). Understanding Logistic Regression - Towards Data Science.

NIH (2019a). ADIPOQ gene. `https://ghr.nlm.nih.gov/gene/ADIPOQ`. Last accessed on Sep 14,2019.

NIH (2019b). APOA5 gene. `https://ghr.nlm.nih.gov/gene/APOA5`. Last accessed on Sep 14,2019.

NIH (2019c). FTO gene. `https://ghr.nlm.nih.gov/gene/FTO`. Last accessed on Sep 14,2019.

NIH (2019d). IL6 gene. `https://ghr.nlm.nih.gov/gene/IL6`. Last accessed on Sep 14,2019.

NIH (2019e). PON1 gene. `https://ghr.nlm.nih.gov/gene/PON1`. Last accessed on Sep 14,2019.

NIH (2019f). PPARG gene. `https://ghr.nlm.nih.gov/gene/PPARG`. Last accessed on Sep 14,2019.

NIH (2019g). TCF7L2 gene. `https://ghr.nlm.nih.gov/gene/TCF7L2`. Last accessed on Sep 14,2019.

Ogden, C. L., Yanovski, S. Z., Carroll, M. D., and Flegal, K. M. (2007). The Epidemiology of Obesity. *Gastroenterology*, 132(6):2087–2102.

Oliveira, P. (1996). Frequências dos genes â lei de Hardy-Weinberg. `http://home.dbio.uevora.pt/~oliveira/Bio/Manual/51.htm`. Last accessed on Sep 24,2019.

Phillips, N. (2017). YaRrr! The Pirate's Guide to R. `https://bookdown.org/ndphillips/YaRrr/comparing-regression-models-with-anova.html`. Last accessed on Sep 22, 2019.

Portela, R. (1965). Integrated Ecological Economic Modeling of ecosystem Services from the Brazilian Amazon Rainforest. *The British Journal of Psychiatry*, 111(479):1009–1010.

Research, U. I. f. D. (2016). Logit Regression: R Data Analysis Examples.

Robert Kalmes, J.-L. H. (2001). Hardy-Weinberg Model. `http://atlasgeneticsoncology.org/Educ/HardyEng.html`. Last accessed on Sep 15,2019.

Rohlfs, R. V. and Weir, B. S. (2008). Distributions of hardy-weinberg equilibrium test statistics. *Genetics*, 180(3):1609–1616.

Russo, R. A. G. and Katsicas, M. M. (2018). Takayasu Arteritis. *Frontiers in Pediatrics*, 6.

Saigi-Morgui, N., Vandenberghe, F., Delacrétaz, A., Quteineh, L., Gholamrezaee, M., Aubry, J. M., Von Gunten, A., Kutalik, Z., Conus, P., and Eap, C. B. (2016). Association of genetic risk scores with body mass index in Swiss psychiatric cohorts. *Pharmacogenetics and Genomics*, 26(5):208–217.

Scuteri, A., Sanna, S., Chen, W. M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orrú, M., Usala, G., Dei, M., Lai, S., Maschio, A., Busonero, F., Mulas, A., Ehret, G. B., Fink, A. A., Weder, A. B., Cooper, R. S., Galan, P., Chakravarti, A., Schlessinger, D., Cao, A., Lakatta, E., and Abecasis, G. R. (2007). Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genetics*, 3(7):1200–1210.

Shkarupa, V. M., Mishcheniuk, O. Y., Henyk-Berezovska, S. O., Palamarchuk, V. O., and Klymenko, S. V. (2016). Polymorphism of DNA repair gene xpd lys751gln and chromosome aberrations in lymphocytes of thyroid cancer patients exposed to ionizing radiation due to the chornobyl accident. *Experimental Oncology*, 38(4):257–260.

Shriner, D. (2011). Approximate and exact tests of Hardy-Weinberg equilibrium using uncertain genotypes. *Genetic Epidemiology*, 35(7):632–637.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 24(1):12–18.

Su, X., Kong, Y., and Peng, D. Q. (2018). New insights into apolipoprotein A5 in controlling lipoprotein metabolism in obesity and the metabolic syndrome patients.

Tekindal, M. A., Gullu, O., Yazici, A. C., and Yavuz, Y. (2016). The Cochran-Armitage test to estimate the sample size for trend of proportions for biological data. *Turkish Journal of Field Crops*, 21(2):286–297.

Veiga, L., Silva-Nunes, J., Melao, A., Oliveira, A., Duarte, L., and Brito, M. (2010). Q192r polymorphism of the paraoxonase-1 gene as a risk factor for obesity in portuguese women. *European journal of endocrinology / European Federation of Endocrine Societies*, 164:213–8.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American Journal of Human Genetics*, 97(4):576–592.

Ware, E. B., Schmitz, L. L., Faul, J., Gard, A., Mitchell, C., Smith, J. A., Zhao, W., Weir, D., and Kardia, S. L. (2017). Heterogeneity in polygenic scores for common human traits. *bioRxiv*, (5):106062.

Yukio (2018). Regressão Logística no R — EstatSite.com. `https://estatsite.com/2018/08/26/regressao-logistica-no-r/`. Last accessed on Sep 8,2019.

Zhang, Y., Cao, Y., Xin, L., Gao, N., and Liu, B. (2018). Association between rs1800629 polymorphism in tumor necrosis factor-alpha gene and dilated cardiomyopathy susceptibility Evidence from case control studies. *Medicine (United States)*, 97(50).

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine*, 4(7).

Zhao, F., Song, M., Wang, Y., and Wang, W. (2016). Genetic model. *Journal of Cellular and Molecular Medicine*, 20(4):765.

Appendix

Descriptive Statistics Analysis

	Obese (n=112)								Normal weight (n=100)								Total (n=212)	
	Measures of Location						Measures of Dispersion		Measures of Location						Measures of Dispersion		Measures of Location	Measures of Dispersion
	Min	1st Qu.	Median	Mean	3rd Qu.	Max	St. Dev.	Var.	Min	1st Qu.	Median	Mean	3rd Qu.	Max.	St. Dev.	Var.	Mean	St. Dev.
Age (years)	18	29	35	34.59	41	50	8.31	69.06	19	26	35	34.15	39.25	50	8.35	69.72	34.38	8.31
Weight (Kg)	65.2	97.7	110.3	111.4	120.5	199.8	21.26	451.93	44.60	52.15	56.10	55.99	59.42	70.80	5.26	27.69	85.26	31.93
Height (cm)	147	156	159	159.8	164	180	6.61	43.72	150	157.8	161	161.6	165.2	175	5.38	28.95	160.6	6.12
BMI (Kg/m2)	30.20	38.80	42.85	43.64	47.20	82.10	7.87	61.99	18.50	20.20	21.25	21.44	22.73	24.80	1.71	2.91	33.17	12.54
Fat mass (Kg)	21.11	43.93	54.02	53.87	61.42	118.08	14.8	219.1	5.13	11.59	14.48	14.34	16.69	23.15	3.59	12.9	35.23	22.64
Fat mass (%)	27.10	44.90	48.50	47.73	51.27	59.10	5.11	26.1	10.80	22.35	25.40	25.33	28.70	37	4.68	21.94	37.16	12.23
Waist perimeter (cm)	87	109.8	117	117.54	124.2	169	15.06	226.88	60	67	71	71.73	74	88	5.84	34.11	95.94	25.71
Waist/hip ratio	0.68	0.82	0.88	0.88	0.93	1.05	0.07	0.01	0.61	0.70	0.74	0.74	0.77	0.88	0.05	0	0.81	0.09
Hip perimeter (cm)	105	125.8	134.5	134.4	142	186	13.3	176.79	87	95	97	97.43	100	113	4.47	20.03	117.0	21.08
Waist/height ratio	0.54	0.68	0.74	0.74	0.78	1.08	0.1	0.01	0.37	0.41	0.44	0.44	0.46	0.54	0.04	0	0.60	0.16

Figure 4.1: Descriptive analysis of continuous quantitative variables

PON_1_Q192R

PON_1_Q192R	QQ	QR	RR	Totals
Case	29	46	17	92
Controls	38	30	7	75
Total	67	76	24	167

Table 4.1: Genotype count - PON_1_Q192R

	QQ	QR+RR	Totals
Case	29	63	92
Control	38	37	75
Totals	67	100	167

Table 4.2: Genotypic count for PON_1_Q192R - QQ vs. QR+RR

	QQ+QR	RR	Totals
Case	75	17	92
Control	68	7	75
Totals	143	24	167

Table 4.3: Genotypic count for PON_1_Q192R - QQ+QR vs. RR

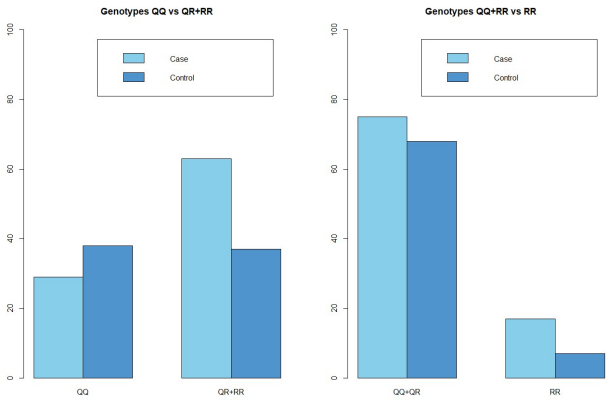

Figure 4.2: Bar Chart of Genotype Crossings for PON_1_Q192R

PON_1_M55L

PON_1_M55L	LL	ML	MM	Totals
Case	36	44	11	91
Controls	32	26	16	74
Total	68	70	27	165

Table 4.4: Genotype count - PON_1_M55L

	LL	LM+MM	Totals
Case	36	55	91
Control	32	42	74
Totals	68	97	165

Table 4.5: Genotypic count for PON_1_M55L - LL vs. LM+MM

	LL+LM	MM	Totals
Case	80	11	91
Control	58	16	74
Totals	138	27	165

Table 4.6: Genotypic count for PON_1_M55L - LL+LM vs. MM

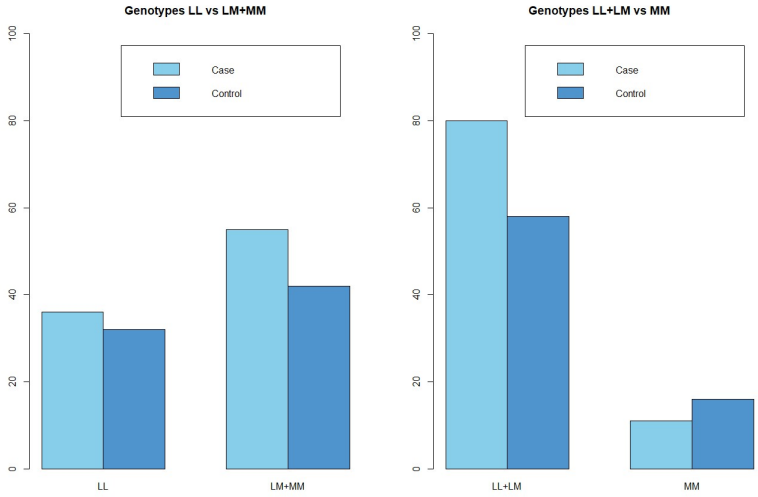

Figure 4.3: Bar Chart of Genotype Crossings for PON_1_M55L

AdipoQG276T

AdipoQG276T	GG	GT	TT	Totals
Case	43	41	13	97
Controls	36	33	6	75
Total	79	74	19	172

Table 4.7: Genotype count - AdipoQG276T

	GG	GT+TT	Totals
Case	43	54	97
Control	36	39	75
Totals	79	93	172

Table 4.8: Genotypic count for AdipoQG276T - GG vs. GT+TT

	GG+GT	TT	Totals
Case	84	13	97
Control	69	6	75
Totals	153	19	172

Table 4.9: Genotypic count for AdipoQG276T - GG+GT vs. TT

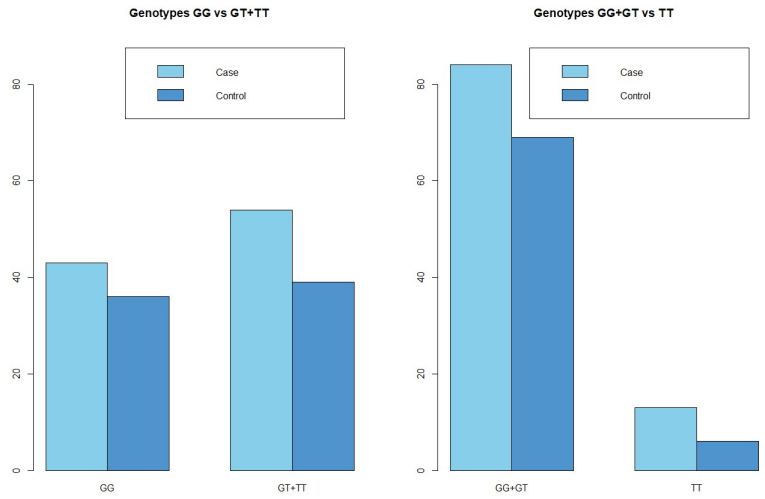

Figure 4.4: Bar Chart of Genotype Crossings for AdipoQG276T

AdipoQ_G11377C

AdipoQ_G11377C	CC	CG	GG	Totals
Case	51	38	7	96
Controls	60	11	4	75
Total	111	49	11	171

Table 4.10: Genotype Count - AdipoQ_G11377C

	CC	CG+GG	Totals
Case	51	45	96
Control	60	15	75
Totals	111	60	171

Table 4.11: Genotypic Count for AdipoQ_G11377C - CC vs. CG+GG

	CC+CG	GG	Totals
Case	89	7	96
Control	71	4	75
Totals	160	11	171

Table 4.12: Genotypic Count for AdipoQ_G11377C - CC+CG vs. GG

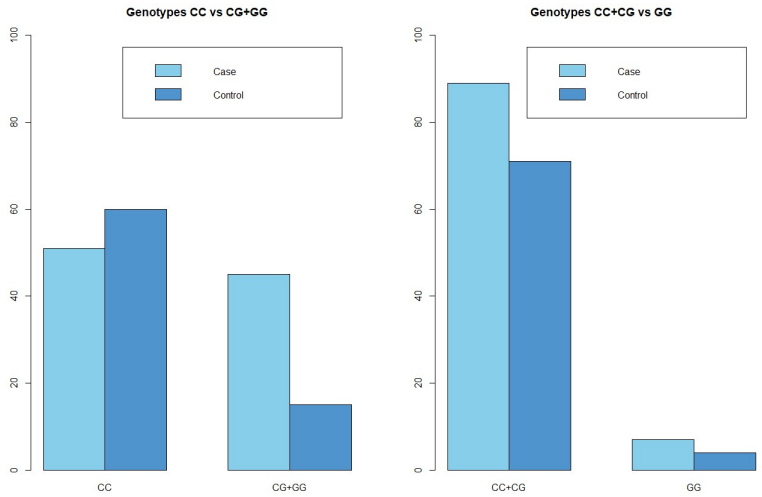

Figure 4.5: Bar Chart of Genotype Crossings for AdipoQ_G11377C

AdipoQ_G11391A

AdipoQ_G11391A	GG	GA	AA	Totals
Case	70	30	2	102
Controls	66	20	1	87
Total	136	50	3	189

Table 4.13: Genotype Count - AdipoQ_G11391A

	GG	GA+AA	Totals
Case	70	32	102
Control	66	21	87
Totals	136	53	189

Table 4.14: Genotypic Count for AdipoQ_G11391A - GG vs. GA+AA

	GG+GA	AA	Totals
Case	100	2	102
Control	86	1	87
Totals	186	3	189

Table 4.15: Genotypic Count for AdipoQ_G11391A - GG+GA vs. AA

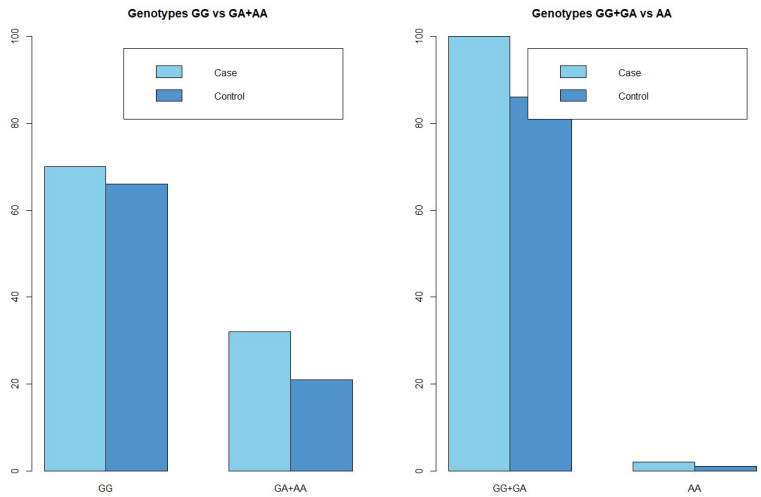

Figure 4.6: Bar Chart of Genotype Crossing for AdipoQ_G11391A

AdipoQ_45T_G

AdipoQ_45T_G	TT	TG	GG	Totals
Case	77	15	3	95
Controls	62	19	4	85
Total	139	34	7	180

Table 4.16: Genotype Count - AdipoQ_45T_G

	TT	TG+GG	Totals
Case	77	18	95
Control	62	23	85
Totals	139	41	180

Table 4.17: Genotypic Count for AdipoQ_45T_G - TT vs. TG+GG

	TT+TG	GG	Totals
Case	92	3	95
Control	81	4	85
Totals	173	7	180

Table 4.18: Genotypic Count for AdipoQ_45T_G - TT+TG vs. GG

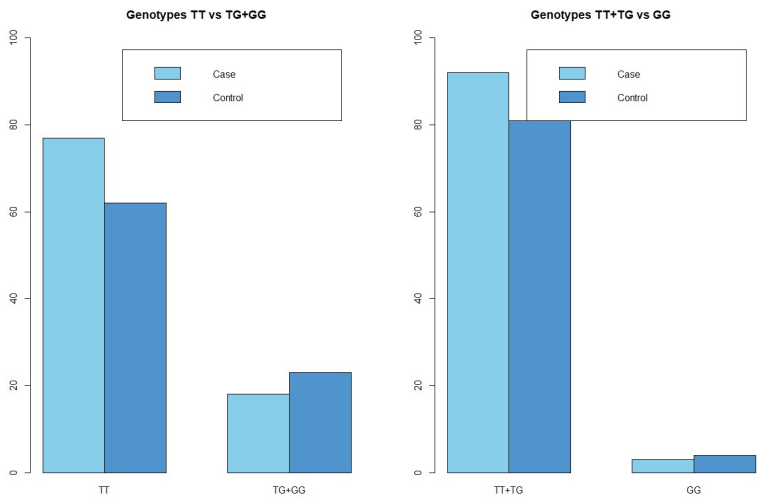

Figure 4.7: Bar Chart of Genotype Crossing for AdipoQ_45T_G

FTO_A_T

FTO_A_T	AA	AT	TT	Totals
Case	25	57	23	105
Controls	12	60	27	99
Total	37	117	50	204

Table 4.19: Genotype Count - FTO_A_T

	AA	AT+TT	Totals
Case	25	80	105
Control	12	87	99
Totals	37	167	204

Table 4.20: Genotypic Count for FTO_A_T - AA vs. AT+TT

	AA+AT	TT	Totals
Case	82	23	105
Control	72	27	99
Totals	154	50	204

Table 4.21: Genotypic Count for FTO_A_T - AA+AT vs. TT

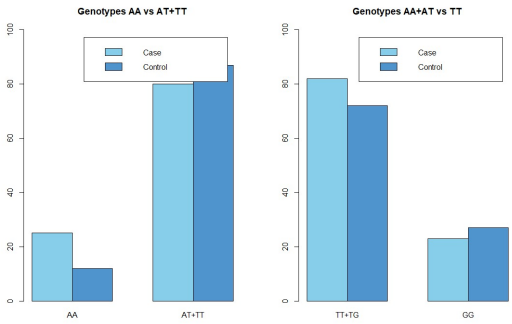

Figure 4.8: Bar Chart of Genotype Crossing for FTO_A_T

PPARG_Pro12Ala

PPARG_Pro12Ala	CC	CG	GG	Totals
Case	88	15	2	105
Controls	78	21	0	99
Total	166	36	2	204

Table 4.22: Genotype Count - PPARG_Pro12Ala

	CC	CG+GG	Totals
Case	88	17	105
Control	78	21	99
Totals	166	38	204

Table 4.23: Genotypic Count for PPARG_Pro12Ala - CC vs. CG+GG

	CC+CG	GG	Totals
Case	103	2	105
Control	99	0	99
Totals	202	2	204

Table 4.24: Genotypic Count for PPARG_Pro12Ala - CC + CG vs. GG

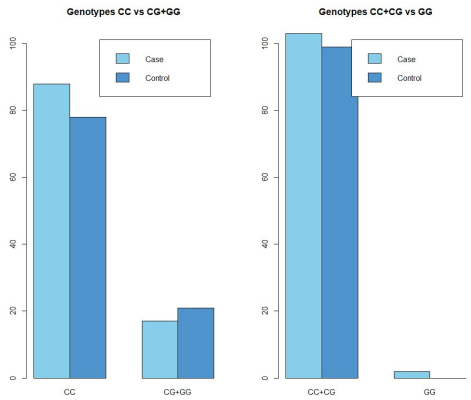

Figure 4.9: Bar Chart of Genotype Crossing for PPARG_Pro12Ala

ApoA5_T1131C

ApoA5_T1131C	AA	AG	GG	Totals
Case	89	15	1	105
Controls	89	8	1	98
Total	178	23	2	203

Table 4.25: Genotype Count - ApoA5_T1131C

	AA	AG+GG	Totals
Case	89	16	105
Control	89	9	98
Totals	178	25	203

Table 4.26: Genotypic Count for ApoA5_T1131C - AA vs. AG+GG

	AA+AG	GG	Totals
Case	104	1	105
Control	97	1	98
Totals	201	2	203

Table 4.27: Genotypic Count for ApoA5_T1131C - AA+AG vs. GG

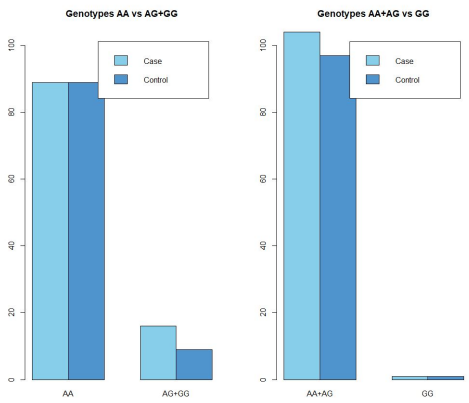

Figure 4.10: Bar Chart of Genotype Crossing for ApoA5_T1131C

ACE_I_D

ACE_I_D	DD	ID	II	Totals
Case	54	31	5	90
Controls	31	27	9	67
Total	85	58	14	157

Table 4.28: Genotype count - ACE_I_D

	DD	ID+II	Totals
Case	54	36	90
Control	31	36	67
Totals	85	72	157

Table 4.29: Genotypic Count for ACE_I_D - DD vs. ID+II

	DD+ID	II	Totals
Case	85	5	105
Control	58	9	67
Totals	143	14	157

Table 4.30: Genotypic Count for ACE_I_D - DD+ID vs. II

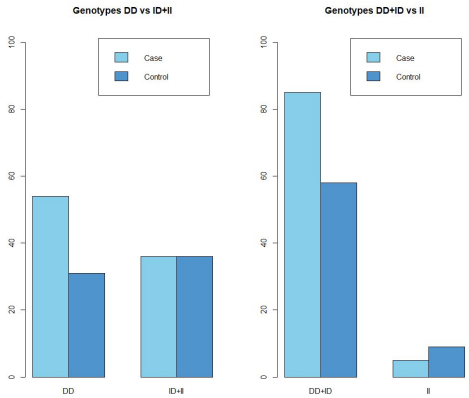

Figure 4.11: Bar Chart of Genotype Crossing for ACE_I_D

IL_6_G572C

IL_6_G572C	CC	GC	GG	Totals
Case	100	12	0	112
Controls	89	10	0	99
Total	189	22	0	211

Table 4.31: Genotype Count - IL_6_G572C

	CC	GC+GG	Totals
Case	100	12	112
Control	89	10	99
Totals	189	22	211

Table 4.32: Genotypic Count for IL_6_G572C - CC vs. GC+GG

	CC+GC	GG	Totals
Case	112	0	112
Control	99	0	99
Totals	211	0	211

Table 4.33: Genotypic Count for IL_6_G572C - CC+GC vs. GG

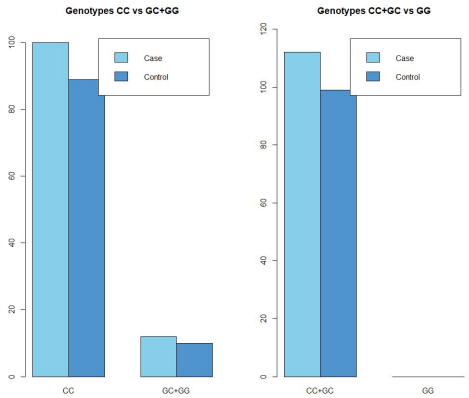

Figure 4.12: Bar Chart of Genotype Crossing for IL_6_G572C

TNFa_G308A

TNFa_G308A	GG	GA	AA	Totals
Case	84	26	2	112
Controls	84	16	0	100
Total	168	42	2	212

Table 4.34: Genotype Count - TNFa_G308A

	GG	GA+AA	Totals
Case	84	28	112
Control	84	16	100
Totals	168	44	212

Table 4.35: Genotypic Count for TNFa_G308A - GG vs. GA+AA

	GG+GA	AA	Totals
Case	110	2	112
Control	100	0	100
Totals	210	2	212

Table 4.36: Genotypic Count for TNFa_G308A - GG+GA vs. AA

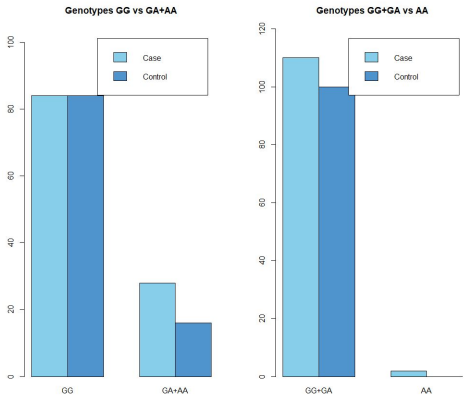

Figure 4.13: Bar Chart of Genotype Crossing for TNFa_G308A

Leptin_G2548A

Leptin_G2548A	AA	AG	GG	Totals
Case	17	76	4	97
Controls	20	55	0	75
Total	37	131	4	172

Table 4.37: Genotype Count - Leptin_G2548A

	AA	AG+GG	Totals
Case	17	80	97
Control	20	55	75
Totals	37	135	172

Table 4.38: Genotypic Count for Leptin_G2548A - AA vs. AG+GG

	AA+AG	GG	Totals
Case	93	4	97
Control	75	0	75
Totals	168	4	172

Table 4.39: Genotypic Count for Leptin_G2548A - AA+AG vs. GG

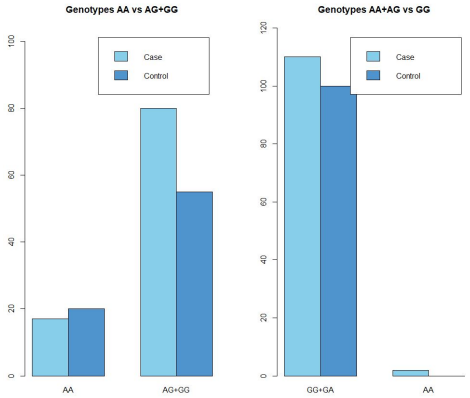

Figure 4.14: Bar Chart of Genotype Crossing for Leptin_G2548A

LeptinR_K109R

LeptinR_K109R	KK	KR	RR	Totals
Case	53	35	8	96
Controls	48	24	3	75
Total	101	59	11	171

Table 4.40: Genotype Count - LeptinR_K109R

	KK	KR+RR	Totals
Case	53	43	96
Control	48	27	75
Totals	101	70	171

Table 4.41: Genotypic Count for LeptinR_K109R - KK vs. KR+RR

	KK+KR	RR	Totals
Case	88	8	96
Control	72	3	75
Totals	160	11	171

Table 4.42: Genotypic Count for LeptinR_K109R - KK+KR vs. RR

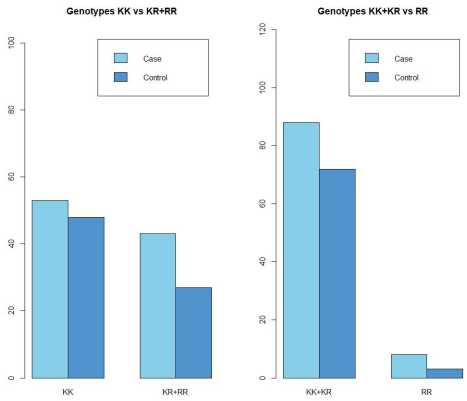

Figure 4.15: Bar Chart of Genotype Crossing for LeptinR_K109R

Ghrelin_R51Q

Ghrelin_R51Q	RR	QR	QQ	Totals
Case	89	8	0	97
Controls	63	12	0	75
Total	152	20	0	172

Table 4.43: Genotype Count - Ghrelin_R51Q

	RR	QR+QQ	Totals
Case	89	8	97
Control	63	12	75
Totals	152	20	172

Table 4.44: Genotypic Count for Ghrelin_R51Q - RR vs. QR+QQ

	RR+QR	QQ	Totals
Case	97	0	97
Control	75	0	75
Totals	172	0	172

Table 4.45: Genotypic Count for Ghrelin_R51Q - RR+QR vs. QQ

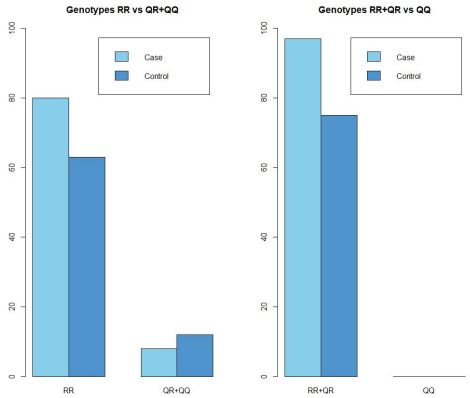

Figure 4.16: Bar Chart of Genotype Crossing for Ghrelin_R51Q

Ghrelin_Leu72Met

Ghrelin_Leu72Met	LL	LM	MM	Totals
Case	84	20	1	105
Controls	80	12	0	92
Total	164	32	1	197

Table 4.46: Genotype Count - Ghrelin_Leu72Met

	LL	LM+MM	Totals
Case	84	21	105
Control	80	12	92
Totals	164	33	197

Table 4.47: Genotypic Count for Ghrelin_Leu72Met - LL vs. LM+MM

	LL+LM	MM	Totals
Case	104	1	105
Control	92	0	92
Totals	196	1	197

Table 4.48: Genotypic Count for Ghrelin_Leu72Met - LL+LM vs. MM

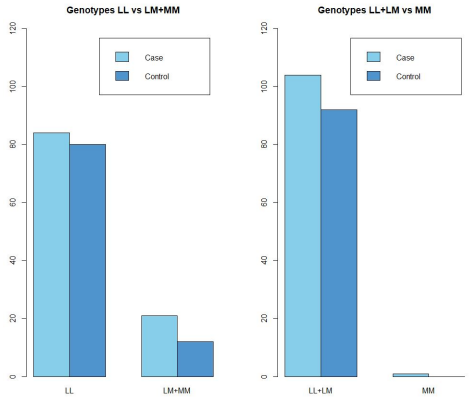

Figure 4.17: Bar Chart of Genotype Crossing for Ghrelin_Leu72Met

MC4R_V103I

MC4R_V103I	II	VI	VV	Totals
Case	69	11	0	80
Controls	67	8	0	75
Total	136	19	0	155

Table 4.49: Genotype Count - MC4R_V103I

	II	VI+VV	Totals
Case	69	11	80
Control	67	8	75
Totals	136	19	155

Table 4.50: Genotypic Count for MC4R_V103I - II vs. VI+VV

	II+VI	VV	Totals
Case	80	0	80
Control	75	0	75
Totals	155	0	155

Table 4.51: Genotypic Count for MC4R_V103I - II+VI vs. VV

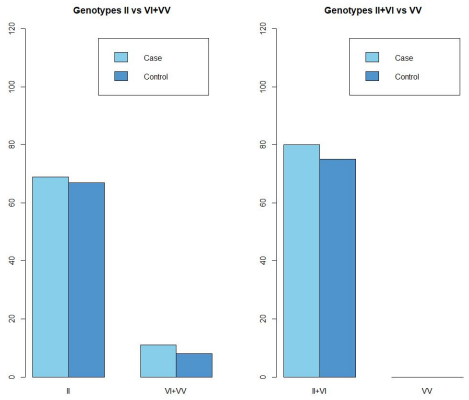

Figure 4.18: Bar Chart of Genotype Crossing for MC4R_V103I

MC4R_rs17782313

MC4R_rs17782313	TT	CT	CC	Totals
Case	69	22	18	109
Controls	60	23	10	93
Total	129	45	28	202

Table 4.52: Genotype Count - MC4R_rs17782313

	TT	CT+CC	Totals
Case	69	40	109
Control	60	33	93
Totals	129	73	202

Table 4.53: Genotypic Count for MC4R_rs17782313 - TT vs. CT+CC

	TT+CT	CC	Totals
Case	91	18	109
Control	83	10	93
Totals	174	28	202

Table 4.54: Genotypic Count for MC4R_rs17782313 - TT+CT vs. CC

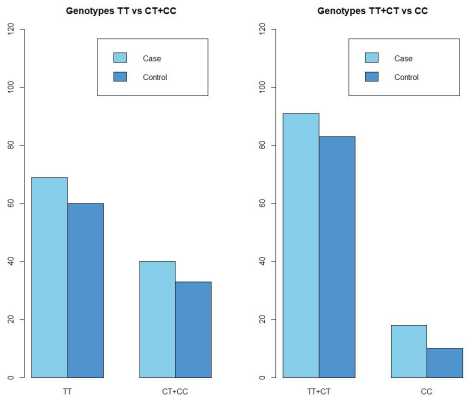

Figure 4.19: Bar Chart of Genotype Crossing for MC4R_rs17782313

TCF7L2_rs7903146_C_T

TCF7L2_rs7903146_C_T	CC	CT	TT	Totals
Case	55	41	10	106
Controls	37	36	9	82
Total	92	77	19	188

Table 4.55: Genotype Count - TCF7L2_rs7903146_C_T

	CC	CT+TT	Totals
Case	55	51	106
Control	37	45	82
Totals	92	96	188

Table 4.56: Genotypic Count for TCF7L2_rs7903146_C_T - CC vs. CT+TT

	CC+CT	TT	Totals
Case	96	10	106
Control	73	9	82
Totals	169	19	188

Table 4.57: Genotypic Count for TCF7L2_rs7903146_C_T - CC+CT vs. TT

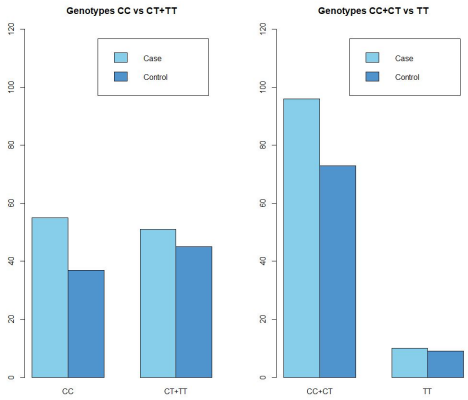

Figure 4.20: Bar Chart of Genotype Crossing for TCF7L2_rs7903146_C_T

Logistic Regression

```
Call:
glm(formula = Group_fator ~ ADIPO11377_SNP + FTO_AT_SNP + PON_192_2_SNP,
    family = binomial(), data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1739  -1.1482   0.6554   1.0770   1.5775

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.2837     0.4935   0.575  0.56534
ADIPO11377_SNP2   1.1451     0.3807   3.008  0.00263 **
FTO_AT_SNP2      -1.1881     0.4922  -2.414  0.01578 *
PON_192_2_SNP2    0.8353     0.3523   2.371  0.01775 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 219.66  on 158  degrees of freedom
Residual deviance: 195.60  on 155  degrees of freedom
  (53 observations deleted due to missingness)
AIC: 203.6

Number of Fisher Scoring iterations: 4
```

Figure 4.21: Output of `Rstudio` code - Stepwise Method - Forward (through the *p-value*)

```
> vif(model_3Ad.Fto.Pon)
ADIPO11377_SNP     FTO_AT_SNP  PON_192_2_SNP
      1.003040       1.008925       1.011470
```

Figure 4.22: Output of `Rstudio` code - Stepwise Method - Forward (through the *p-value*) with VIF values

```
Call:
glm(formula = Group_fator ~ ACE_SNP + FTO_AT_SNP + ADIPO11377_SNP +
    PON_192_2_SNP, family = binomial(), data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3934  -1.0362   0.4688   0.9198   1.6181

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.9814     0.5896   1.664  0.09602 .
ACE_SNP2        -0.4643     0.3612  -1.285  0.19873
FTO_AT_SNP2     -1.5114     0.5543  -2.727  0.00640 **
ADIPO11377_SNP2  1.1715     0.4046   2.896  0.00378 **
PON_192_2_SNP2   0.6526     0.3721   1.754  0.07947 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 204.61  on 148  degrees of freedom
Residual deviance: 179.28  on 144  degrees of freedom
  (63 observations deleted due to missingness)
AIC: 189.28

Number of Fisher Scoring iterations: 4
```

Figure 4.23: Output of `Rstudio` code - Stepwise Method - Forward (through the *AIC*)

```
> vif(model_4Ace.Fto.Ad.Pon)
      ACE_SNP     FTO_AT_SNP ADIPO11377_SNP  PON_192_2_SNP
     1.006724       1.013780       1.014269       1.017479
```

Figure 4.24: Output of `Rstudio` code - Stepwise Method - Forward (through the AIC) with VIF values

```
> #Teste de Hosmer-Lemeshow
> m<-model_3Ad.Fto.Pon
> y<-Group_fact
> hoslem.test(m$y, fitted(m))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  m$y, fitted(m)
X-squared = 3.3176, df = 8, p-value = 0.9129
```

Figure 4.25: Output of `Rstudio` code - Hosmer and Lemeshow of Fit Test

```
> #### SCORE ###
> #ADIPO = 0 ; FTO = 2; PON = 0
> SCORE_riskonlywithFTO<-1.1451*0 - 1.1881*2 + 0.8353*0;SCORE_riskonlywithFTO
[1] -2.3762
> #ADIPO = 1 ; FTO = 1; PON = 1
> SCORE_riskalleles1<-1.1451*1 - 1.1881*1 + 0.8353*1;SCORE_riskalleles1
[1] 0.7923
> #ADIPO = 2 ; FTO = 0; PON = 2
> SCORE_riskithoutFTO<-1.1451*2 - 1.1881*0 + 0.8353*2;SCORE_riskithoutFTO
[1] 3.9608
> #ADIPO = 2 ; FTO = 1; PON = 2
> SCORE_riskiwithFTO1<-1.1451*2 - 1.1881*1 + 0.8353*2;SCORE_riskiwithFTO1
[1] 2.7727
> #ADIPO = 1 ; FTO = 2; PON = 1
> SCORE_riskonlywithFTO2<-1.1451*1- 1.1881*2 + 0.8353*1;SCORE_riskonlywithFTO2
[1] -0.3958
> #ADIPO = 2 ; FTO = 2; PON = 2
> SCORE_riskalleles2<-1.1451*2- 1.1881*2 + 0.8353*2;SCORE_riskalleles2
[1] 1.5846
```

Figure 4.26: Output of `Rstudio` code - Genetic Risk Score for some combinations, in each SNP