



**POLITECNICO**  
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE  
E DELL'INFORMAZIONE

# Bayesian models for early diagnosis and prediction of metabolic syndrome in healthy blood donors

TESI DI LAUREA MAGISTRALE IN  
MATHEMATICAL ENGINEERING - INGEGNERIA MATEMATICA

Author: **Francesca Arrigoni**

Student ID: 103713

Advisor: Prof. Ilenia Epifani

Co-advisors: Prof. Alessandra Guglielmi

Academic Year: 2023-24



## Abstract

Metabolic syndrome is a cluster of conditions that increase the risk of cardiovascular disease, stroke, and type 2 diabetes. This thesis aims to predict the onset of metabolic syndrome in the population of blood donors. The data was provided by AVIS Milano and includes variables related to the donor's lifestyle and the blood test results for each donation. After pre-processing the data we fit it in a two-stage plug-in model. For the first stage, we considered three mixed-effect Bayesian models for longitudinal data to predict future values of the five responses defining the metabolic syndrome. These results have been plugged into a logistic model to estimate the donor's overall probability of developing metabolic syndrome. Three risk zones were identified to classify the donors and we were able to correctly identify up to 90% of at-risk donors. As a result of this thesis, a screening tool was developed and provided to AVIS. The posterior analysis of the Bayesian models reveals a strong correlation between metabolic syndrome and lifestyle-related variables such as BMI, physical activity levels, and smoking habits. The final classification results are very promising and will allow AVIS to more effectively identify at-risk donors before the onset of metabolic syndrome, optimizing resources and improving donors' health.

**Keywords:** Bayesian models, blood donors, classification, longitudinal data, metabolic syndrome, risk assessment



## Abstract in lingua italiana

La sindrome metabolica è un insieme di condizioni che aumentano il rischio di malattie cardiovascolari, ictus e diabete di tipo 2. Questa tesi mira a prevedere l'insorgenza della sindrome metabolica nella popolazione dei donatori di sangue. I dati sono stati forniti da AVIS Milano e includono variabili relative allo stile di vita dei donatori e i risultati delle analisi del sangue di ogni donazione. Dopo aver pulito i dati sono stati inseriti in un modello plug-in a due fasi. Nella prima fase, abbiamo considerato tre modelli a effetti misti bayesiani per dati longitudinali al fine di predire i valori futuri delle cinque variabili necessarie per diagnosticare la sindrome metabolica. Questi risultati sono stati poi utilizzati in un modello logistico per stimare, per ogni donatore, la probabilità di sviluppare la sindrome metabolica. Sono state identificate tre zone di rischio in cui classificare i donatori, e siamo stati in grado di identificare correttamente fino al 90% dei donatori a rischio. Come risultato di questa tesi, è stato sviluppato uno strumento di screening fornito ad AVIS. L'analisi a posteriori dei modelli bayesiani ha rivelato una forte correlazione tra la sindrome metabolica e le variabili legate allo stile di vita come il BMI, i livelli di attività fisica e le abitudini di fumo. I risultati finali della classificazione sono molto promettenti e permetteranno ad AVIS di identificare più efficacemente i donatori a rischio prima dell'insorgenza della sindrome metabolica, ottimizzando così le risorse e migliorando la salute dei donatori.

**Parole chiave:** classificazione, dati longitudinali, donatori del sangue, modelli bayesiani, sindrome metabolica, valutazione del rischio



# Contents

<b>Abstract</b>	<b>i</b>
<b>Abstract in lingua italiana</b>	<b>iii</b>
<b>Contents</b>	<b>v</b>
<b>Introduction</b>	<b>1</b>
<b>1 Dataset description and summary</b>	<b>5</b>
1.1 Description of the problem . . . . .	5
1.2 Data sources description . . . . .	6
1.2.1 Italian donation rules . . . . .	9
1.3 Exploratory data analysis . . . . .	10
1.3.1 Missing values . . . . .	10
1.3.2 Categorical variables . . . . .	12
1.3.3 Correlation and multicollinearity among numerical variables . . . . .	16
1.3.4 Sampling times of target variables . . . . .	17
1.4 Data transformation . . . . .	21
1.4.1 Implausible values . . . . .	21
1.4.2 Target variables . . . . .	22
1.4.3 Creation of the datasets . . . . .	27
<b>2 Bayesian models for the metabolic syndrome</b>	<b>29</b>
2.1 Mixed-effects models for longitudinal data . . . . .	29
2.2 Application of mixed-effects longitudinal models to the AVIS dataset . . . . .	30
2.2.1 Covariates . . . . .	31
2.2.2 Model 1 . . . . .	35
2.2.3 Model 2 . . . . .	36
2.2.4 Model 3 . . . . .	36

2.2.5	Model selection for Bayesian models . . . . .	37
2.3	Prediction model . . . . .	38
2.3.1	Logistic regression model . . . . .	38
2.3.2	Combinatorial model . . . . .	39
2.3.3	Model selection for prediction models . . . . .	40
<b>3</b>	<b>Posterior analysis</b>	<b>43</b>
3.1	Stan Software . . . . .	43
3.2	Comparison of the models . . . . .	43
3.3	Posterior inference for Model 2 . . . . .	44
3.3.1	Fixed effects: $\beta$ regression coefficients . . . . .	44
3.3.2	Random effects $b_i$ . . . . .	55
<b>4</b>	<b>Predictions</b>	<b>63</b>
4.1	Prediction . . . . .	63
4.1.1	Dataset composition . . . . .	63
4.1.2	Comparison of the predictions . . . . .	63
4.1.3	Threshold choice . . . . .	67
4.2	Tool provided to AVIS . . . . .	71
<b>5</b>	<b>Conclusions and future developments</b>	<b>73</b>
	<b>Bibliography</b>	<b>75</b>
<b>A</b>	<b>STAN Code</b>	<b>77</b>
<b>B</b>	<b>Variations of Model 2</b>	<b>79</b>
B.1	Model 2_A . . . . .	79
B.2	Model 2_B . . . . .	79
B.3	Comparison . . . . .	83
<b>C</b>	<b>Horse-shoe prior and feature selection</b>	<b>85</b>
C.1	Horseshoe Prior . . . . .	85
C.2	Feature selection . . . . .	86
<b>D</b>	<b>Convergence diagnostic</b>	<b>87</b>

<b>List of Figures</b>	<b>93</b>
<b>List of Tables</b>	<b>97</b>
<b>Acknowledgements</b>	<b>99</b>



# Introduction

Metabolic syndrome, also called insulin resistance syndrome, is a cluster of conditions that increase the risk of cardiovascular disease, stroke, and type 2 diabetes. It is diagnosed if the individual presents at least three of the following conditions: increased blood pressure, high blood sugar, excess body fat around the waist, low levels of HDL cholesterol, and high triglyceride levels.

In recent years metabolic syndrome has shown a notable increase in cases worldwide, driven by a combination of lifestyle, environmental, and genetic factors. One of the primary contributors to this rise is the significant shift towards more sedentary lifestyles. Additionally, there has been a marked change in dietary habits worldwide, characterized by increased consumption of processed foods high in sugar, unhealthy fats, and refined carbohydrates, which contribute to obesity, one of the major risk factors for metabolic syndrome. The aging population is another factor contributing to the rise in metabolic syndrome, as its prevalence has been shown to be directly correlated with age. These trends are not exclusive to specific regions, as evidenced by the analysis of Scuteri et al. (2015) on the distribution of metabolic syndrome in the European population.

The rise in metabolic syndrome has profound implications for the public health system, placing a growing burden on healthcare resources, affecting individual's quality of life, and having significant economic impacts. Consequently, there is a growing demand for preventive health services, including screening programs, lifestyle interventions, and educational campaigns to raise awareness about metabolic syndrome and its risk factors.

The objective of this thesis is to provide AVIS Milano with a screening tool that leverages a donor's previous blood test results to predict their risk of developing metabolic syndrome. This tool will enable doctors to more accurately identify at-risk donors and provide them with tailored advice on lifestyle changes to prevent the onset of the disease.

The data utilized in this thesis was provided by AVIS Milano. It includes both blood test results from each donation and self-reported information about the donors' lifestyle.

A two-stage plug-in model was developed in this work. In the first stage, a series of mixed-effect Bayesian models for longitudinal data were applied to predict future values

of the five variables used to diagnose metabolic syndrome. Three variations of these models were considered, each incorporating different levels of between-subject variability by using a varying number of random-effects parameters. The optimal model was selected based on goodness-of-fit metrics and considerations of computational efficiency. Posterior analysis of the chosen model revealed that lifestyle-related variables, such as BMI, physical activity levels, and smoking habits, were particularly significant in influencing each target variable and the overall risk of developing metabolic syndrome.

The second stage uses the prediction obtained from the previous model to make inferences on the donor's risk of developing metabolic syndrome. A logistic regression model and a combinatorial model were studied for this stage and the best one was selected based on their predictive capabilities.

This approach integrates advanced statistical techniques to derive insights from longitudinal data and provides a robust framework for predicting and assessing the metabolic syndrome risk among donors.

From these models, a light-stop system was developed to categorize donors into different risk zones based on their likelihood of developing metabolic syndrome. This method has demonstrated the capability to accurately identify up to 90% of at-risk donors and 80% of healthy donors. This effectively reduces the number of donors requiring additional clinical attention before donation to 25% or 12% of the total population, optimizing resource allocation and enhancing preventive healthcare measures. A screening tool incorporating these models was provided to AVIS for initial donor assessments, enhancing their ability to prioritize and intervene early in managing donor health.

My work started by extracting the raw data needed for this thesis from the AVIS's databases using SQL queries. In Chapter 1 the initial data is presented along with a comprehensive description of all variables taken into consideration in this work. The chapter details the rigorous data-cleaning process undertaken and includes an exploratory analysis of the dataset.

Chapter 2 delves into the theoretical foundations of the models adopted in this thesis. It proposes three distinct Bayesian models for predicting target variables and two models specifically designed to forecast the risk of developing metabolic syndrome. This chapter also outlines the criteria used to select the optimal models for each task.

Chapter 3 presents the posterior results derived from the Bayesian models discussed in Chapter 2. It includes the selection of the best model among the proposed Bayesian models. This chapter examines the posterior estimates of fixed-effect and random-effect parameters of the chosen model, analyzing these findings in comparison to existing literature.

Chapter 4 focuses on identifying the most effective model for predicting the overall risk of metabolic syndrome. It provides a detailed analysis and interpretation of the prediction results. Additionally, it explains the development and functionality of the tool provided to AVIS for predicting metabolic syndrome risk based on each donor's historical data. The final chapter concludes with a summary of the key points of the thesis and discusses potential avenues for future research and development in the field.



# 1 | Dataset description and summary

## 1.1. Description of the problem

The metabolic syndrome is a cluster of conditions that occur together, increasing the risk of heart disease, stroke, and type 2 diabetes. A patient is classified as having the syndrome when they have at least three of the following conditions:

- high blood pressure (hypertension): blood pressure consistently higher than 130/85 mm Hg;
- high blood sugar (insulin resistance): fasting glucose level higher than 100 mg/dL;
- excess body fat around the waist: a waist circumference of over 40 inches (101.6 cm) in men and over 35 inches (88.9 cm) in women;
- high triglyceride levels: levels of triglycerides exceeding 150 mg/dL;
- low levels of HDL cholesterol: HDL cholesterol levels below 40 mg/dL in men and below 50 mg/dL in women;

In Europe, metabolic syndrome is present in 24,3% of the population with an age-associated increase as shown by Scuteri et al. (2015). Lifestyle changes like adopting a healthy diet, exercising regularly, losing weight, and managing stress can help prevent or manage metabolic syndrome, see Mohamed et al. (2023).

AVIS (Associazione Volontari Italiani Sangue) is an Italian voluntary association dedicated to blood donation and related activities. It was founded in Milan in 1927 by Vittorio Formentano and has since played a significant role in organizing blood donation campaigns, raising awareness about the importance of blood donation, and mobilizing volunteers to contribute to the healthcare system. Today, it is the largest Italian blood volunteer organization which manages to ensure approximately 70% of the national blood supply. AVIS also focuses on supporting patients who require blood transfusions, advocating for policies that enhance blood safety and availability, and coordinating with

healthcare professionals and institutions to optimize blood donation practices.

One of the objectives of AVIS, as written on their website [Avis](#), is to care for the donors' health and psychophysical well-being, promoting a healthy lifestyle based on proper nutrition, regular physical activity, and taking steps to prevent certain diseases such as sexually transmitted infections or hepatitis. As a result, Dr. Sergio Casartelli, the general director of AVIS Milano, considered utilizing statistical methods on the extensive dataset gathered by AVIS, with the aim of predicting the likelihood of donors developing specific diseases. In particular, we analyze the case of metabolic syndrome, where early detection could facilitate successful intervention through lifestyle modifications.

## 1.2. Data sources description

The data used in this study was provided by the AVIS headquarters of Milan located in Lambrate. It was retrieved from two different databases:

- the EMONET database that contains the information about the donations and the exams of a certain donor,
- the AVIS database that contains information about the lifestyle of the donors.

The original dataset contains information about the donations of 268251 donors gathered between 1992 and now. We have only focused on donations between 2009 and September of 2023, as information about the lifestyle of the donors have been gathered starting from 2009. Furthermore, only habitual donors, i.e. having more than two donations in the period 2009-2023, are considered. The final dataset consists of 8789 habitual donors of which 6651 male and 2138 female.

All variables in the dataset are reported and described in Tables: 1.1 to 1.5. In each table, the first column shows the original Italian name from the corresponding databases, and the second column shows the new English name assigned to facilitate the comprehension of the results. Table 1.1 contains variables that do not change over time (such as gender and blood type) that are retrieved from the EMONET dataset. Table 1.2 contains categorical variables that can change over time retrieved from the EMONET dataset, while Tables 1.3 and 1.4 contain the numerical variables. Table 1.5 contains all variables retrieved from the AVIS dataset, all of which change over time. The waist circumference is measured by a doctor at the time of some of the donations, whereas the variables related to the lifestyle habits of the donor are self-reported data. This means they rely on the individual's own report of their habits and are not medically verified, thus introducing a potential bias.

The variables Glucose, HDL\_cholesterol, Triglycerides, Circumference, PMAX and PMIN are used to diagnose the metabolic syndrome in a patient so from now on they will be referred to as target variables.

Variable name	New name	Type	Description
CAI	CAI	num	Donor unique ID
Sesso	Gender	cat	Gender: 1 for men, 0 for women
AB0	AB0	cat	Blood type: A, B, AB, 0
Rh	Rh	cat	Reshus factor: POS or NEG

Table 1.1: Time-fixed covariates from the EMONET dataset

Variable name	New name	Description
Anti hcv	Anti_hcv	HCV (Hepatitis C Virus) antibodies
ANTI Hiv 12	ANTI_Hiv	HIV-1 or HIV-2 (Human immunodeficiency Virus) antibodies
ANTI T. Pallidum	ANTI_T.Pallidum	Treponema pallidum antibodies
P Nat Hbv	P_Nat_Hbv	Nucleic Acid Test result for HBV (Hepatitis B Virus)
P Nat Hcv	P_Nat_Hcv	Nucleic Acid Test result for HBC
P Nat Hiv 12	P_Nat_Hiv	Nucleic Acid Test result for HIV-1/2
S Hbsag	S_Hbsag	HBsAg protein
Sistema Xg	Xg_system	Phenotype of antigen Xg

Table 1.2: Time-dependent categorical covariates from the EMONET dataset

Variable name	New name	Description
Alanina aminotransferasi	ALT	Alanine aminotransferase (ALT)
Albumina	Albumin	Albumin
Albumina perc	Albumin_perc	Percentage of albumin
alfa 1 globuline	A1_globulins	Alpha-1-globulins
alfa 1 globuline perc	A1_globulins_perc	Percentage of alpha-1-globulins
alfa 2 globuline	A2_globulins	Alpha-2-globulins
alfa 2 globuline perc	A2_globulins_perc	Percentage of alpha-2-globulins
Altezza	Height	Donors's height
Basofili	Basophils	Basophils
Basofili perc	Basophils_perc	Percentage of basophils
beta 1 globuline	B1_globulins	Beta-1-globulins
beta 1 globuline perc	B1_globulins_perc	Percentage of beta-1-globulins
beta 2 globuline	B2_globulins	Beta-2-globulins
beta 2 globuline perc	B2_globulins_perc	Percentage of beta-2-globulins
Colesterolo Hdl	HDL_cholesterol	HDL cholesterol
Colesterolo totale	Total_cholesterol	Total cholesterol
Creatinina	Creatinine	Creatinine
Distribuzione di volume	Volume_distribution	Red blood cell distribution width
Ematocrito hct	Hematocrit	Hematocrit
Emoglobina conc media	Hemoglobin_mean_conc	Mean corpuscular hemoglobin concentration
Emoglobina massa media	Hemoglobin_mean_mass	Mean corpuscular hemoglobin
Emoglobina hb	Hemoglobin	Hemoglobin
Eosinofili	Eosinophils	Eosinophils
Eosinofili perc	Eosinophils_perc	Percentage of Eosinophils
Eritrociti rbc	Erythrocytes	Red blood cells
Ferritina	Ferritin	Ferritin
Ferro totale	Total_iron	Total iron
gamma globuline	G_globulins	Gamma globulins
gamma globuline perc	G_globulins_perc	Percentage of gamma globulins
Glucosio	Glucose	Glucose
Leucociti wbc	Leukocytes	Leukocytes
Linfociti	Lymphocytes	Lymphocytes
Linfociti perc	Lymphocytes_perc	Percentage of lymphocytes

Table 1.3: Time-dependent numerical covariates from the EMONET dataset part 1

Variable name	New name	Description
Monociti	Monocytes	Monocytes
Monociti perc	Monocytes_perc	Percentage of monocytes
Neutrofili	Neutrophil	Neutrophil
Neutrofili perc	Neutrophil_perc	Percentage of neutrophils
Peso	Weight	Donor's weight
Piastrine	Platelets	Platelets
PMAX	PMAX	Systolic pressure
PMIN	PMIN	Diastolic pressure
Polso	Heart_rate	Heart rate
Proteine totali	Total_proteins	Total proteins
Transferrina	Transferrin	Transferrin
Trigliceridi	Triglycerides	Triglycerides
Volume medio	Mean_Volume	Mean Cell Volume (MCV)

Table 1.4: Time-dependent numerical covariates from the EMONET dataset part 2

Variable name	New name	Type	Description
Circonferenza vita	Circumference	num	Waist circumference
Fumo	Smoke	cat	Smoking habits
Alcool	Alcohol	cat	Drinking habits
Attività fisica	Activity	cat	Physical activity habits

Table 1.5: Time-dependent covariates from the AVIS dataset

### 1.2.1. Italian donation rules

Since our data comes from the donation process, it is not representative of the general population. Indeed, to donate, candidates must meet certain criteria to protect both the health of the recipient and the donor.

- All donors must be between 18 and 60 years, with the possibility of increasing the maximum age to 65 years and even 70 years if a doctor approves.
- All donors must weigh more than 50 kg.
- The systolic pressure must be under 180 mmHg.
- The diastolic pressure must be under 100 mmHg.

- The resting heart rate must be between 50 and 100 beats/min.
- The hemoglobin must be over 13.5 g/dL for male donors and over 12.5 g/dL for female donors.

Because of these rules, the range of values that some variables can assume is truncated. This has been taken into consideration in this study even if the model for the target variable will not be a truncated one.

## 1.3. Exploratory data analysis

### 1.3.1. Missing values

We analyzed the amount of missing data in the dataset, focusing on the number of patients without any measurements for certain exams. Figure 1.1 shows the proportion of donors in the dataset who do not have measurements for each covariate.

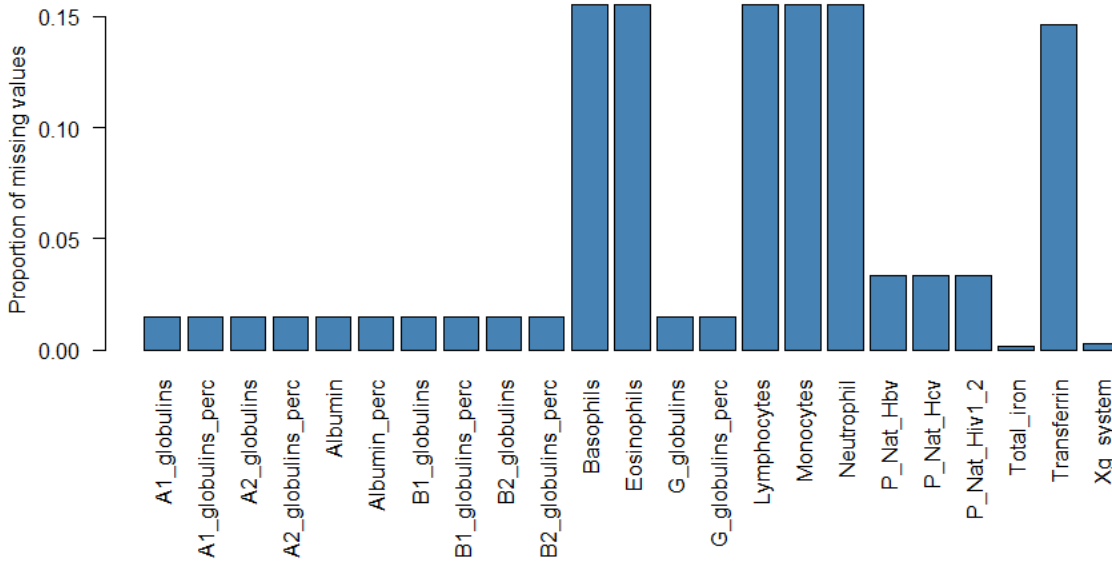


Figure 1.1: Proportion of missing exams

Upon investigating the cause of the missing data, we discovered that variables with a high percentage of missing values were only prescribed during a subset of the chosen time

frame, as shown in Table 1.6. Hence donors who donated only before this period do not have measurements for these specific exams.

Exam name	Prescription period
Basophils	
Eosinophils	
Lymphocytes	2016-2023
Monocytes	
Neutrophil	
Transferrin	2009-2012
P_Nat_Hcv	2013-2023
P_Nat_Hiv_12	2013-2023
P_Nat_Hbv	2013-2023

Table 1.6: Prescription period associated with variables having a high proportion of missing data

For this reason, we decided to discard the variables of Table 1.6 from the analysis. The remaining percentages of missing data, as shown in Figure 1.2, have been lowered to 1.5%, which is reasonable to deal with by dropping the specific donors.

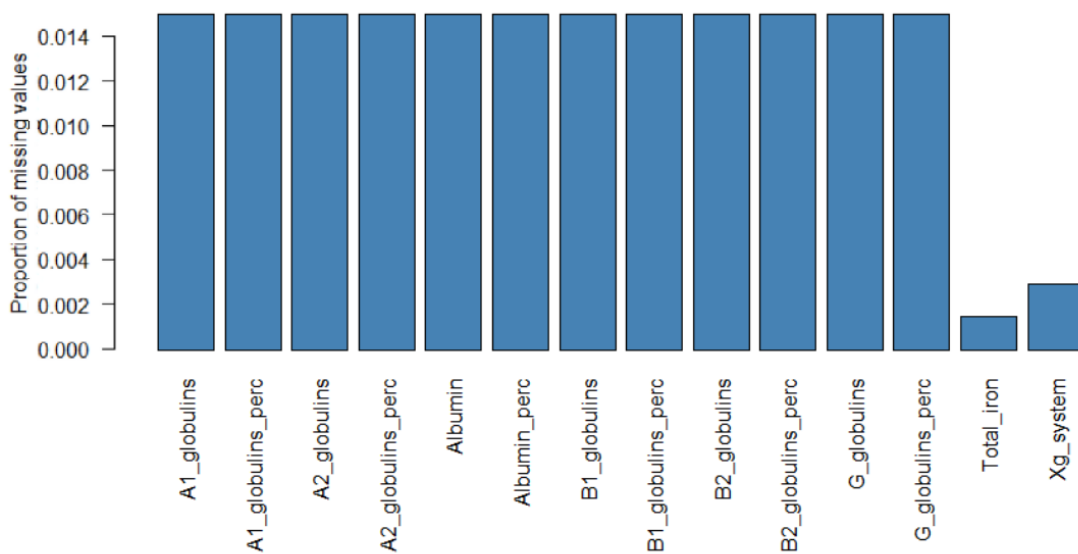


Figure 1.2: Proportion of missing exams on the remaining variables

### 1.3.2. Categorical variables

In this dataset, we have three types of categorical variables:

- time independent categorical variables: gender, blood type, Rhesus factor;
- self-reported time-dependent categorical variables: alcohol consumption, levels of physical activity, smoking habits;
- exams time-dependent categorical variables: Anti hvc, Anti Hiv 1 2,...

We start to summarize the time-independent categorical variables with their sample frequencies.

Variable name	Value	Percentage
Gender	F	24.33%
	M	75.67%
AB0	A	38.67%
	B	11.69%
	AB	4.29%
	0	45.35%
Rh	Positive	85.32%
	Negative	14.68%

Table 1.7: Percentage of levels

As we can see in Table 1.7 we have a majority of male donors; furthermore, the percentage distribution of the blood type is consistent with the Italian distribution reported by the ISS (Istituto Superiore di Sanità) as is the Rhesus factor.

All time-dependent categorical variables present a high number of levels, so some of them have been aggregated to make these variables more understandable. In particular, the variables from Table 1.2 which give the result of the presence of a virus infection have been standardized to only allow NULL/POS/NEG responses by employing the aggregation shown in Table 1.8.

On the other hand, for self-declared variable we aggregated the possible responses as shown in Table 1.9, Table 1.10 and Table 1.11, we can see that as these variables are inputted by hand they have a greater variability in their value.

Value	New value
NULL	
NV	NULL
IR	
DUM	
POSITIVO	
POS	POS
RR	
NEG	NEG

Table 1.8: Aggregated responses for the exams' categorical variables

## Smoking habits

Value	New value
No- di 10 sig/die	
No5-10 sigarette/die	
- di 10 sig/die	
+ di 20 sig/die	
< 5 sigarette/die	
> 30 sigarette/die	
10-20 sig/die	Smoker
10-20 sigarette/die	
20-30 sigarette/die	
5-10 sigarette/die	
fumo passivo	
pipa o sigaro	
Sigaretta elettronica	
ex da 1 anno	
ex da 10 anni	Non-smoker
ex da 3 anni	
No	

Table 1.9: Aggregated responses for the variable Smoke

### Drinking habits

Value	New value
Assunzione saltuaria	
No	Non habitual drinker
No < 25 g/die	
NoNo	
25-50 g/die	Habitual drinker
50-100 g/die	
> 100 g/die	
oltre 100 g/die	

Table 1.10: Aggregated responses for the variable Alcohol

### Physical activity

Value	New value
Sedentaria	
Vita sedentaria	Sedentary lifestyle
Scarsa	
Saltuaria	
Vita moderatamente attiva	Moderate lifestyle
Moderata (30-60 minuti/die)	
Moderata 30-60 minuti/die	
Vita attiva	Active lifestyle
Superiore ad 1 ora/die	
Attività sportiva non agonistica	
Attività sportiva agonistica	
> 60 min /die ( non sportiva ) > 1 h/die	

Table 1.11: Aggregated responses for the variable Activity

All categorical variables are listed with their new encoding and corresponding sample frequencies in Table 1.12. As we can see the percentages of the positive values for all the exam categorical variables are very low, as the viruses sought in these tests afflict a very small part of the population. Furthermore, when you are diagnosed with a virus such as HIV or Hepatitis you are not allowed to donate blood anymore, so for each donor there is at most one positive value for these exams.

For this reason, as they are not representative, we have not considered all the variables listed previously in Table 1.2.

On the other hand, the self-reported variables are more balanced and we can see that most donors have a healthy lifestyle. Indeed there are more non-smokers than smokers, and most donors report doing some kind of physical activity, there does not seem to be a significant difference between habitual and non habitual drinkers intake in the population.

Variable name	Value	Percentage
Alcohol	Habitual drinker	48.12%
	Non habitual drinker	51.88%
Smoke	Smoker	34.56%
	Non smoker	65.44%
Activity	Active lifestyle	24.23%
	Moderate lifestyle	49.73%
	Sedentary lifestyle	26.04%
Anti_hcv	NEG	99.88%
	POS	0.023%
	NULL	0.097%
Anti_Hiv_12	NEG	99.96%
	POS	0.001%
	NULL	0.039%
Anti_T.pallidum	NEG	99.97%
	POS	0.013%
	NULL	0.017%
S_Hbsag	NEG	99.97%
	POS	0.003%
	NULL	0.027%
Xg_system	NEG	97.28%
	POS	2.56%
	NULL	0.16%

Table 1.12: Percentage of levels

### 1.3.3. Correlation and multicollinearity among numerical variables

The correlation between numerical variables was studied to identify redundant features and prevent an overfitting of the model.

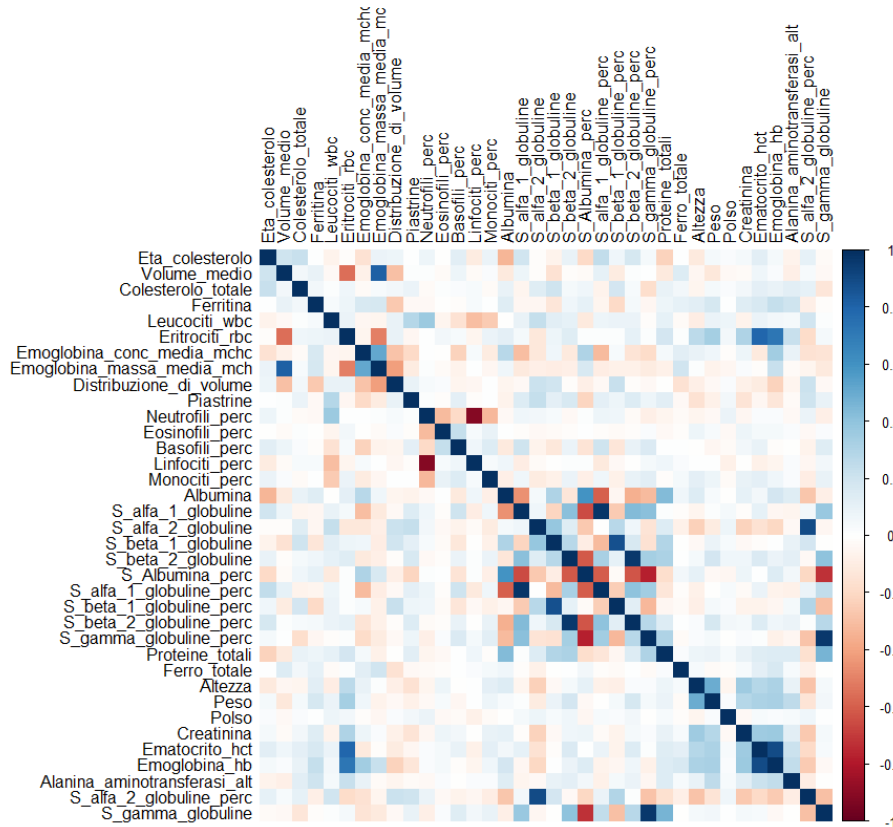


Figure 1.3: Correlation between numeric covariates

Based on the correlation values presented in Figure 1.3 and the multicollinearity values obtained through the calculation of the Variance Inflated Factor (VIF), variables with the highest degree of collinearity were identified. Consequently, the following variables were deemed redundant and dropped:

- Neutrophil\_perc
- Albumin\_perc
- G\_globulins\_perc
- A1\_globulins\_per
- A2\_globulins\_perc
- B1\_globulins\_perc
- B2\_globulins\_perc

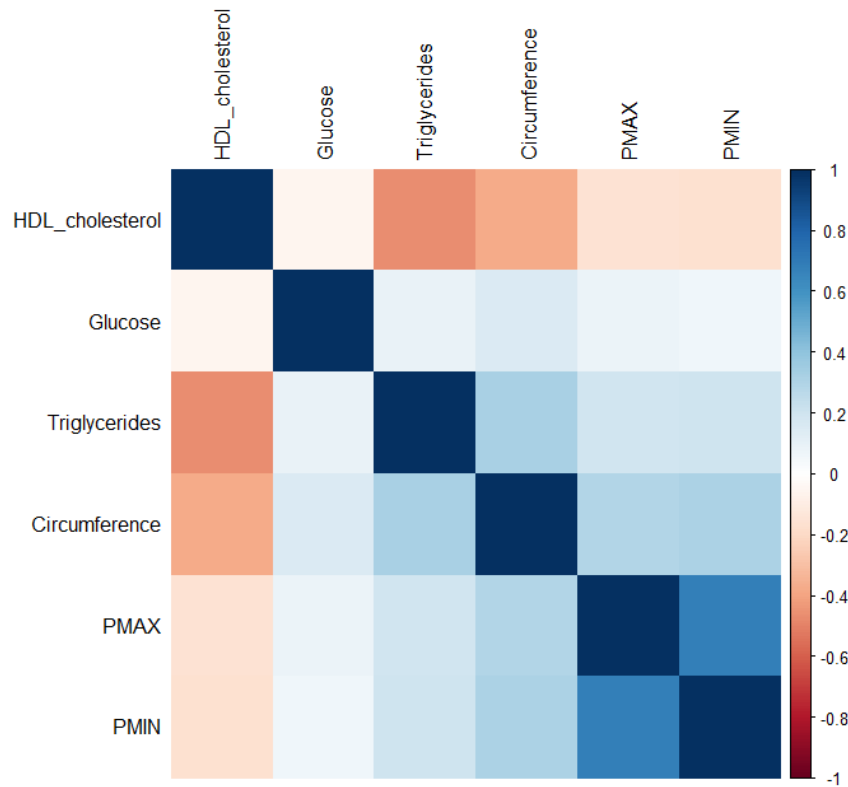


Figure 1.4: Correlation between target variables

From Figure 1.4, it can be seen that the variables PMAX and PMIN are highly correlated. It is known in literature that systolic and diastolic blood pressure have a linear relationship (see Gavish et al. (2008)). Hence, it was decided to consider only the variable PMAX as a target variable as PMIN would follow a similar trend.

#### 1.3.4. Sampling times of target variables

Donations do not happen at fixed time intervals, since men are allowed to donate 4 times per year and must wait at least 90 days between donations, while women can donate 2 times with a waiting period of 180 days. A small tolerance on these thresholds is possible after clinical evaluation, and women in menopause can start donating as frequently as men.

Due to these rules, the personal commitments of donors, their health status, and the fluctuating demand for blood during specific periods, significant variations in the intervals between donations can occur.

Another source of variability is introduced by the fact that different target variables are measured at different time intervals. Some variables such as Glucose and PMAX are

measured at all donations, some others as Triglycerides are only measured once a year.

Table 1.13 presents a summary of these gap times for each target variable, whereas Figures 1.5 to 1.9 depict graphs of the frequency distributions. The red horizontal lines correspond to the logarithms of 90 and 180, which denote the minimum waiting times for men and women, respectively.

Variable name	Sex	Min	1st Qu	Median	Mean	3rd Qu	Max
Glucose	F	4	184	212	291.5	294	3406
	M	1	99	122	180.8	181	3953
HDL_cholesterol	F	4	227	404	525.8	645	3922
	M	2	148	301	390.4	481	4036
Triglycerides	F	4	223	393	497.7	612	3922
	M	1	146	287	363.4	455	3815
Circumference	F	1	183	222	335.3	381.5	3293
	M	1	102	133	227.6	246	4254
PMAx	F	1	183	213	302.1	321	3406
	M	1	99	124	192.6	193.3	3953

Table 1.13: Summary of gap times between donations

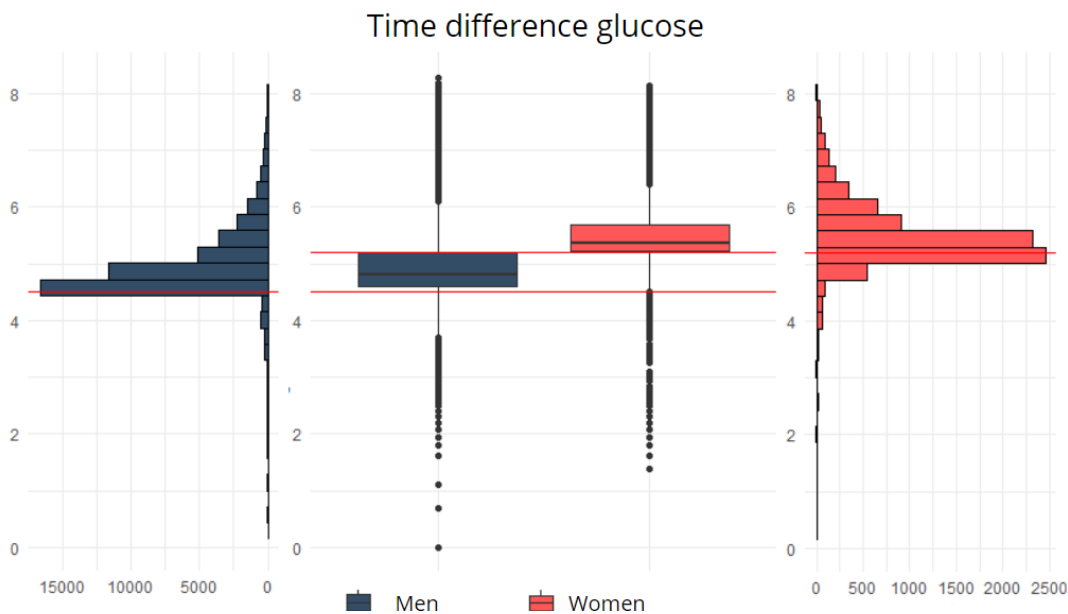


Figure 1.5: Gaptimes on log scale between successive measurements of Glucose.

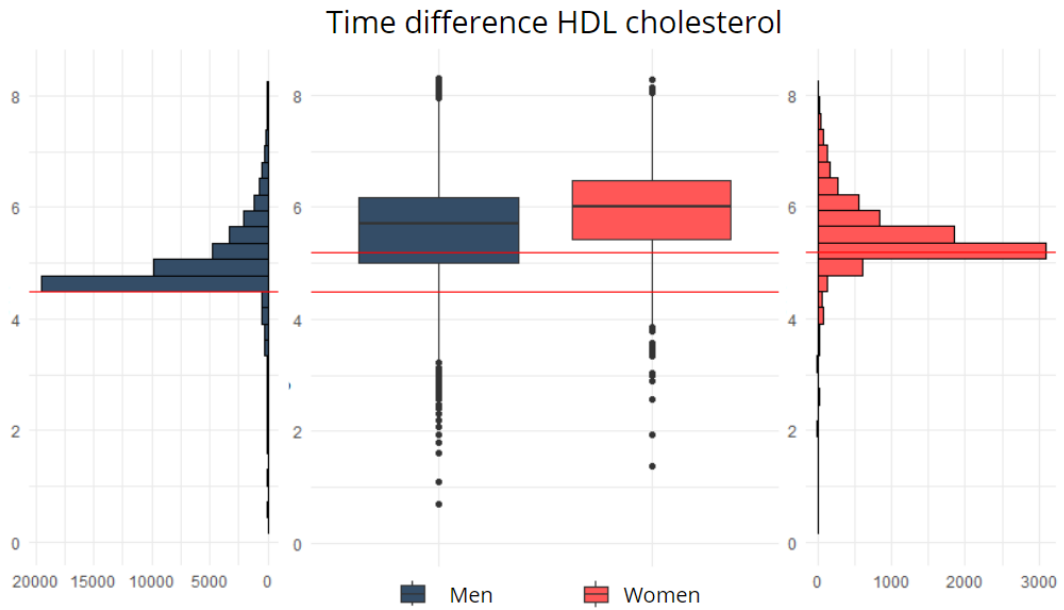


Figure 1.6: Gaptimes on log scale between successive measurements of HDL\_cholesterol.

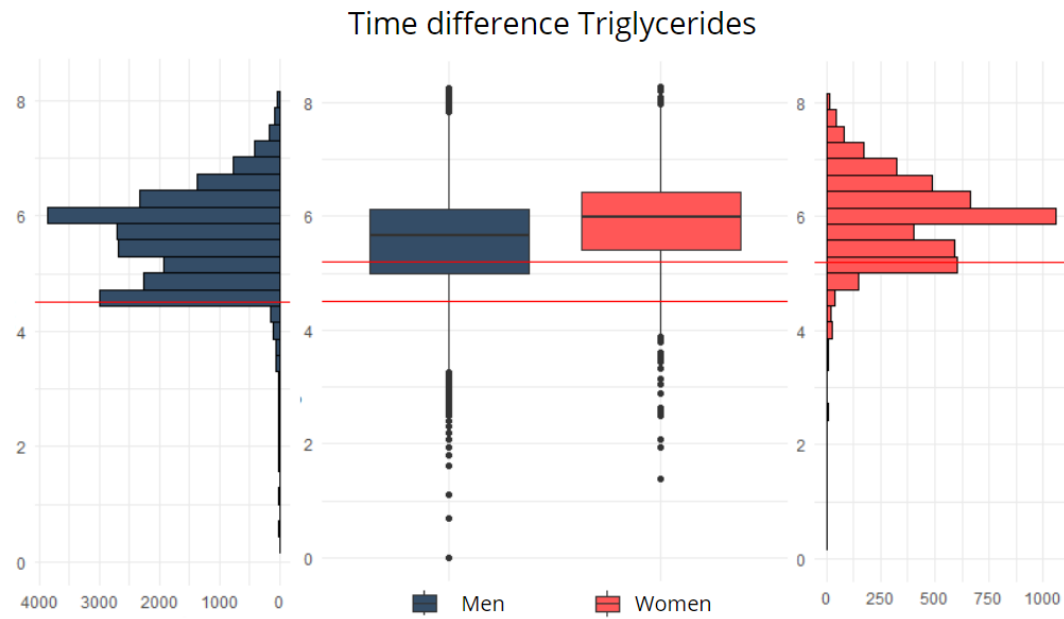


Figure 1.7: Gaptimes on log scale between successive measurements of Triglycerides.

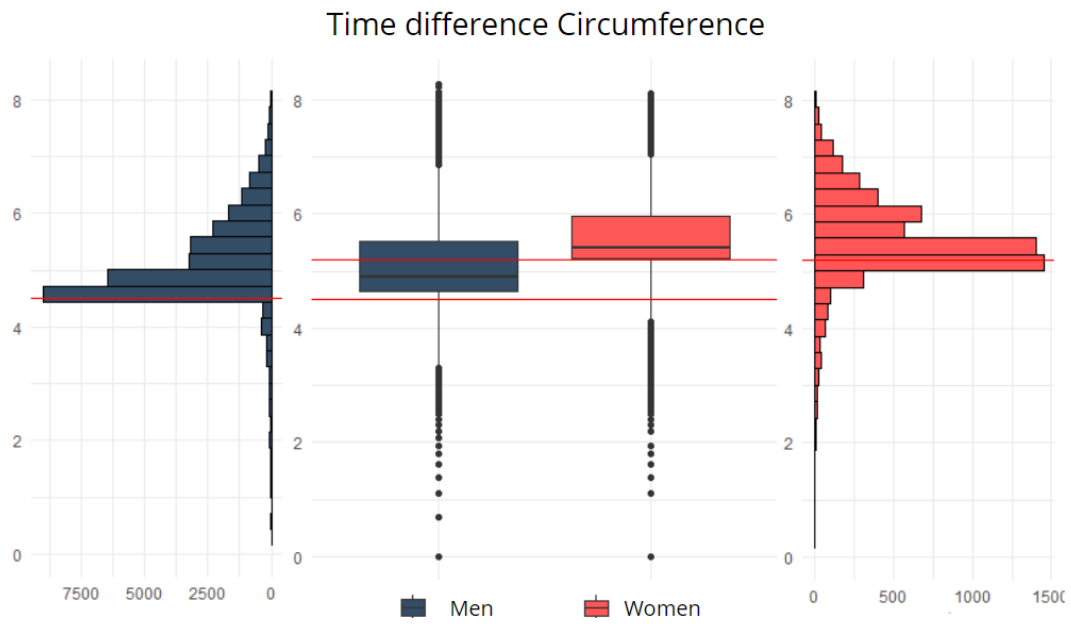


Figure 1.8: Gaptimes on log scale between successive measurements of Circumference.

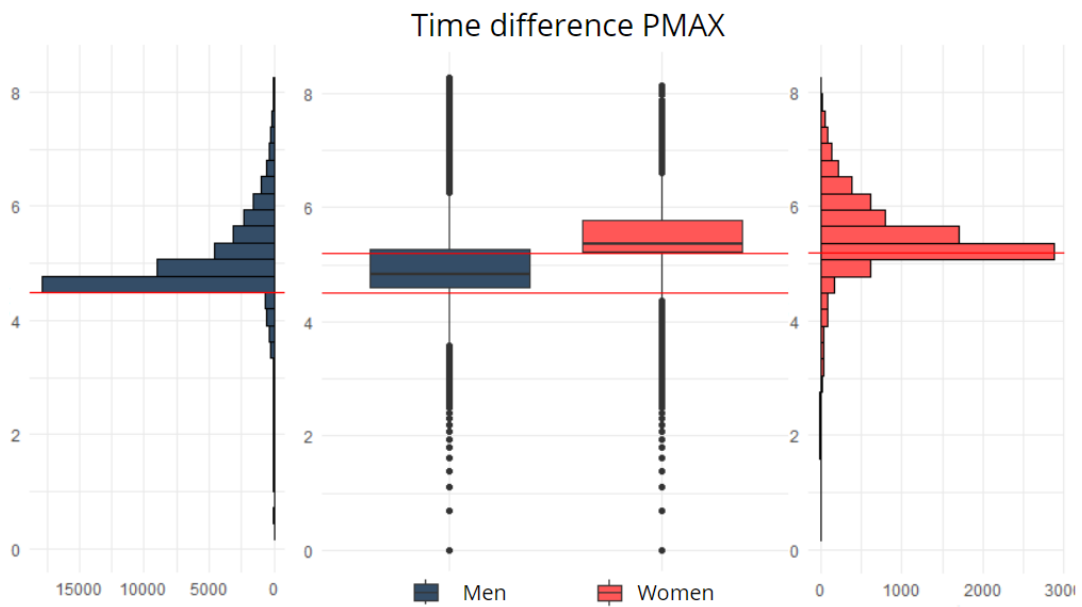


Figure 1.9: Gaptimes on log scale between successive measurements of PMAX.

## 1.4. Data transformation

The categorical features that were retained have been converted into dummy variables to fit a statistical model. Additionally, the Body Mass Index (BMI) variable has been included in the features, calculated as  $BMI = \text{Weight}/\text{Height}^2$ , with weight in *kg* and height in *m*, to offer a more comprehensive representation of the individual donor's body composition. Furthermore, the variable age has been incorporated, indicating the age of the donor at the time the target variable was measured. All target variables and numerical covariates underwent a logarithmic transformation to normalize them and have been scaled.

### 1.4.1. Implausible values

Among the time-dependent covariates, some are manually entered by doctors during the patient's pre-donation visit. Consequently, input errors can occur, leading to some implausible values. These values were identified through manual inspection and knowledge of reasonable bounds. They were replaced with the correct value if there was only a unit of measurement error, or with "NA" if the value was nonsensical.

The target variables were also searched for unreasonable values. Table 1.14 shows the total number of such values detected, along with the thresholds used to identify a normal value.

Variable name	Number of implausible values	Acceptable range
Glucose	35	20-300 mg/dL
HDL_cholesterol	5	15-165 mg/dL
Triglycerides	67	10-600 mg/dL
Circumference	23	50-200 cm
Height	124	120-260 cm
Weight	46	40-210 kg
PMAX	143	60-200 mmHg
PMIN	56	40-185 mmHg
Heart_rate	110	30-180 beats/min

Table 1.14: Number of implausible values

### 1.4.2. Target variables

Table 1.15 shows the summary statistics of the target variables after absurd values removal and in Figure 1.11 the distributions of the target variables are depicted after the logarithmic transformation.

Variable name	Mean	Standard deviation	Min	Max
Glucose	91.33	11.41	31	293
HDL_cholesterol	57.66	14.53	17	161
Triglycerides	97.82	52.69	14	596
Circumference	91.62	10.39	52	187
PMAX	120.2	9.86	62	190
PMIN	76.66	7.82	50	140

Table 1.15: Summary statistics after absurd values' removal

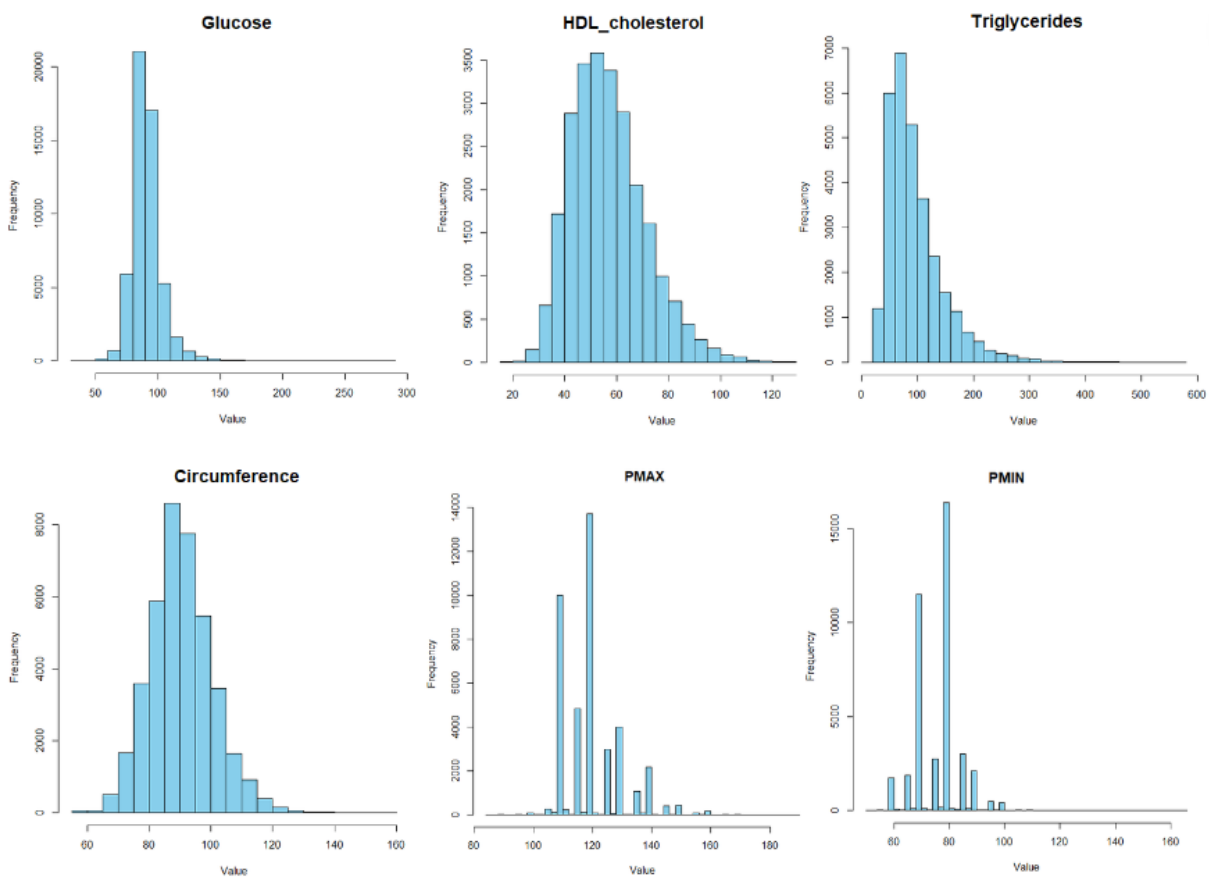


Figure 1.10: Histograms of the target variables

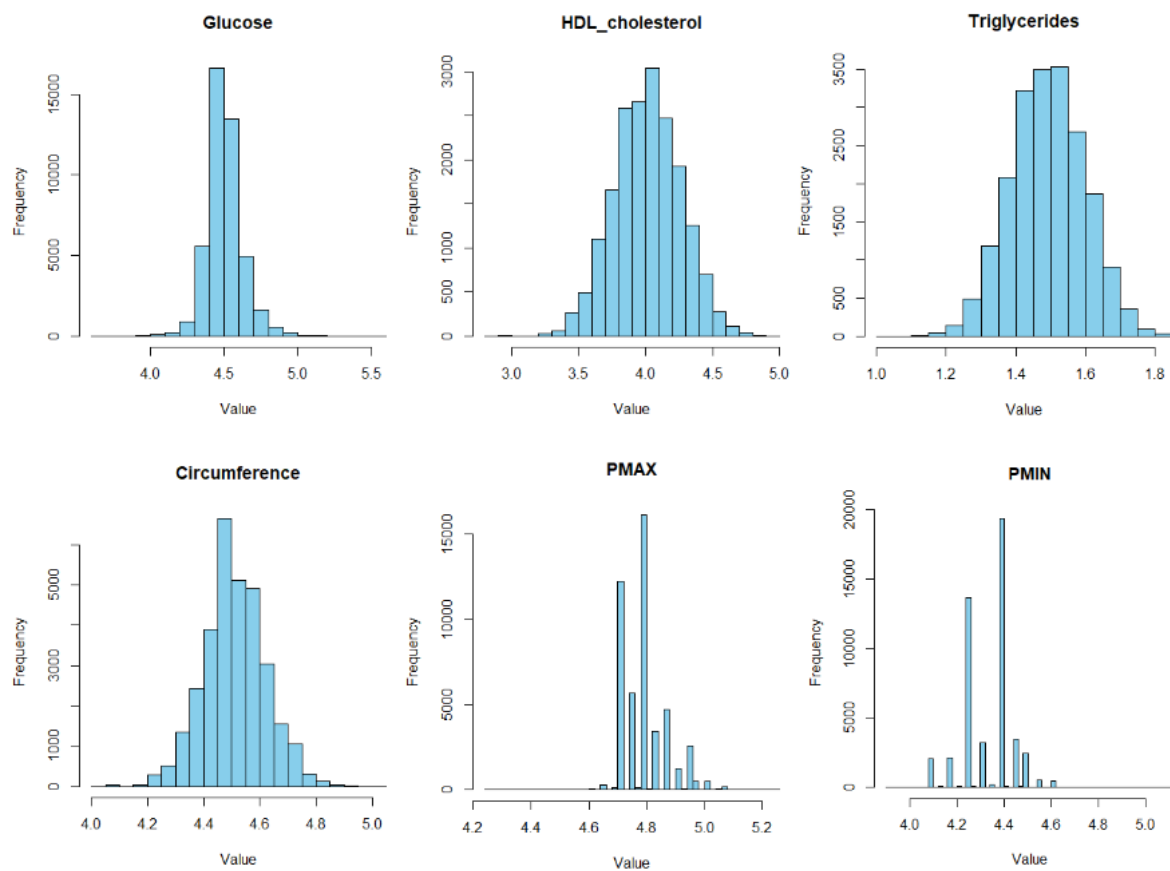


Figure 1.11: Histograms of the target variables on log scale

As the histogram of the triglycerides value is still skewed after a logarithmic transformation, another transformation was applied.

For the pressure variables, it can be seen that some values are more frequent than others, as it is a discrete variable instead of continuous and we can assume that when inputting the data the doctors round the value to the nearest multiple of five. To obtain a smoother distribution, a jitter was added to both variables.

Figures 1.12 to 1.17 display the distributions of the target variables grouped by gender, following the previously mentioned transformations. We can see that Glucose is the only variable that does not seem to depend on the gender of the donor. HDL cholesterol levels are higher for women while triglycerides levels, waist circumference, systolic and diastolic pressure levels are higher for men. Even divided by gender, the distributions are mostly symmetric with PMAX being the less symmetric variable .

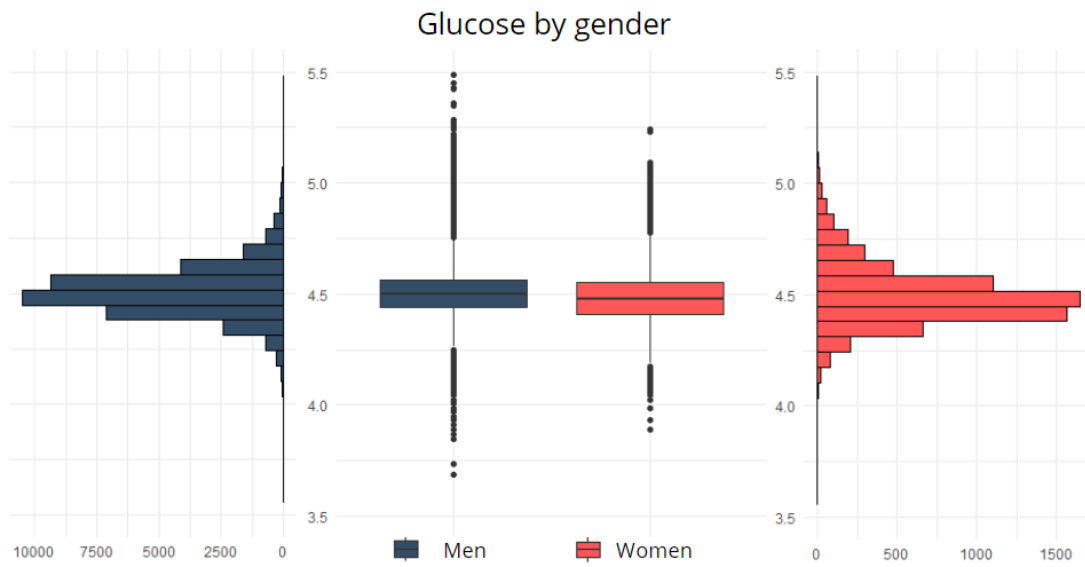


Figure 1.12: Boxplot and histograms of glucose levels, grouped by gender

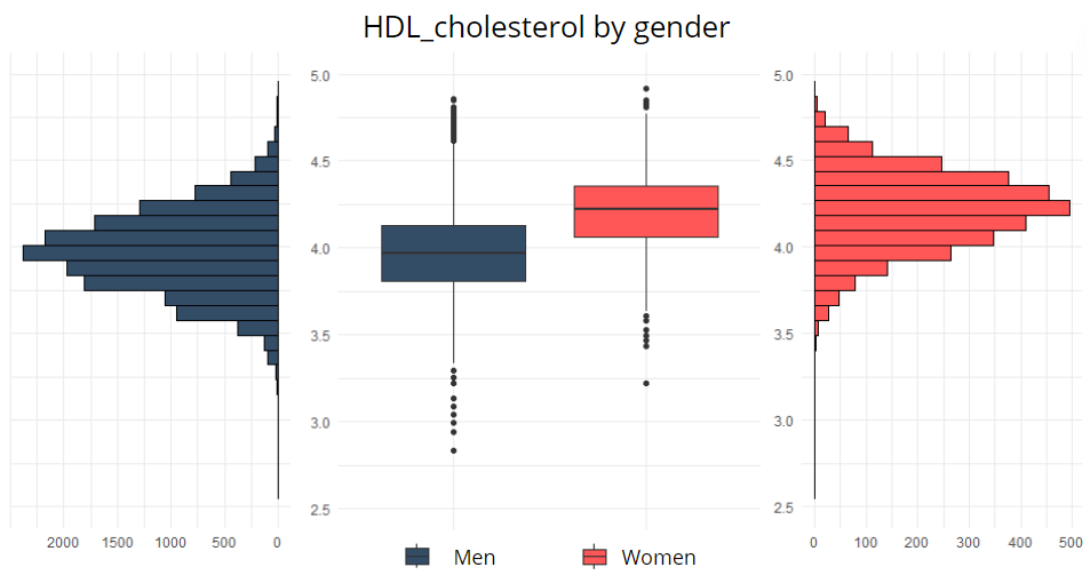


Figure 1.13: Boxplot and histograms of Hdl cholesterol levels, grouped by gender

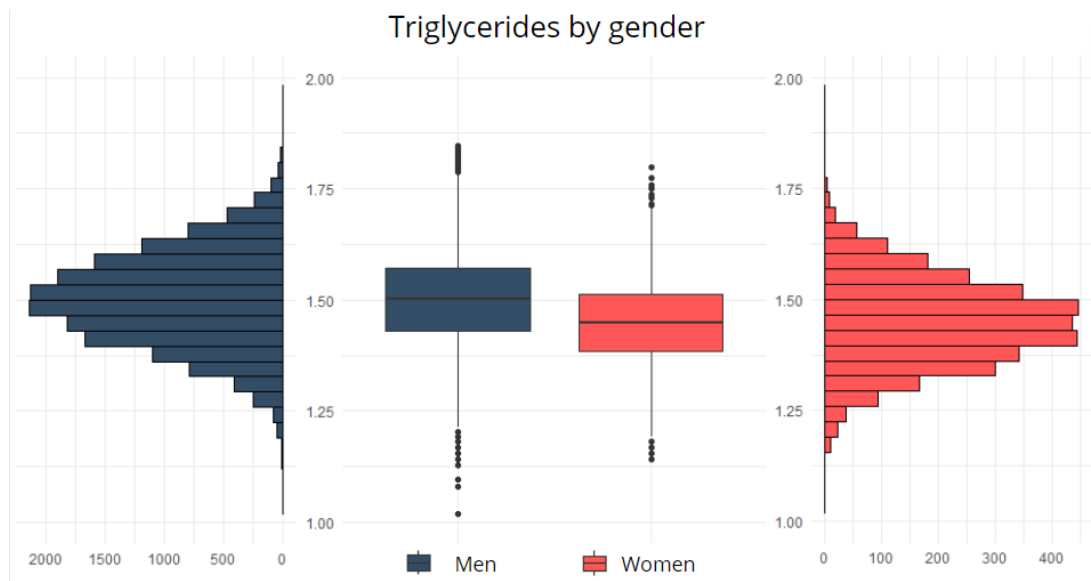


Figure 1.14: Boxplot and histograms of triglycerides levels, grouped by gender

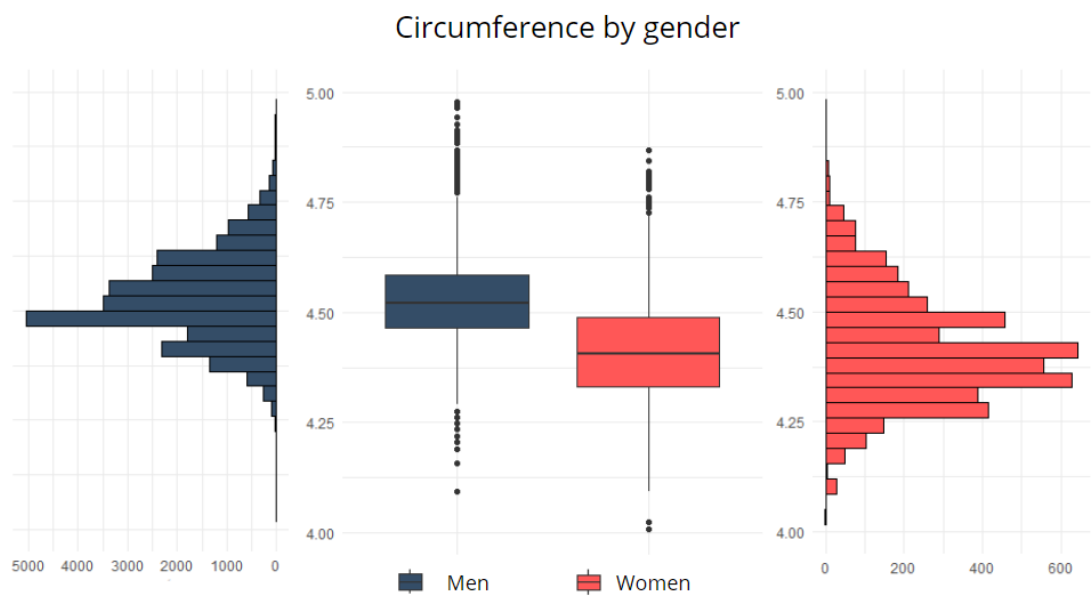


Figure 1.15: Boxplot and histograms of waist circumference, grouped by gender

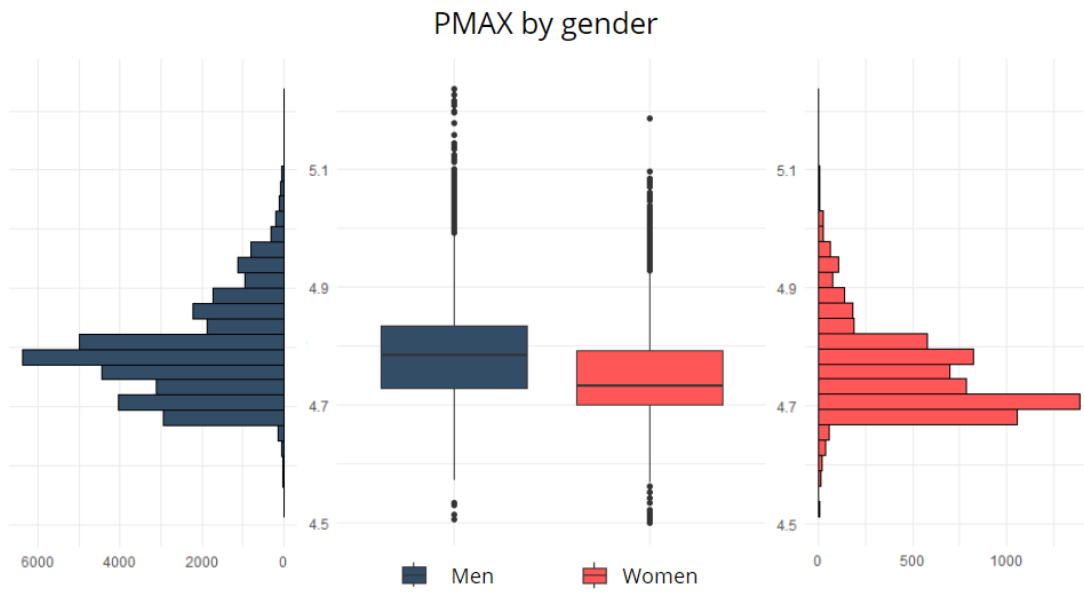


Figure 1.16: Boxplot and histograms of systolic pressure, grouped by gender



Figure 1.17: Boxplot and histograms of diastolic pressure, grouped by gender

### 1.4.3. Creation of the datasets

To predict the value of the target variable based on past examination results, we needed to adapt the data retrieved from the AVIS databases to our requirements.

The first step was to group the data by individual and exam. The resultant dataset had a number of rows equal to the number of donors, with each column representing the time series of one of the covariates. Next, we divided the dataset into two halves, ensuring that both subsets maintained equivalent distributions of male and female donors, as well as consistent age distributions. These datasets are referred to as Dataset 1 and Dataset 2.

As previously mentioned, different target variables are measured at varying frequencies. Our objective is to accurately predict the future values of these target variables before making inferences about the onset of metabolic syndrome. Therefore, we required a distinct dataset for each target variable to properly train the models.

In the second step, Dataset 1 and Dataset 2 were used to create datasets specific for each target variable. For each value of the target variable for a given donor, the previous measurement of each covariate was retrieved, and these values were then combined into a single row of the target variable's dataset.

All datasets obtained from Dataset 1 were employed for feature selection and the exploration of prior knowledge. Datasets from Dataset 2 were further divided to obtain Training sets and Test sets for models development. From each of the datasets in Dataset 2, the information regarding the most recent donation of each patient was isolated to construct a Test set, while the remaining portion constituted the Training set.

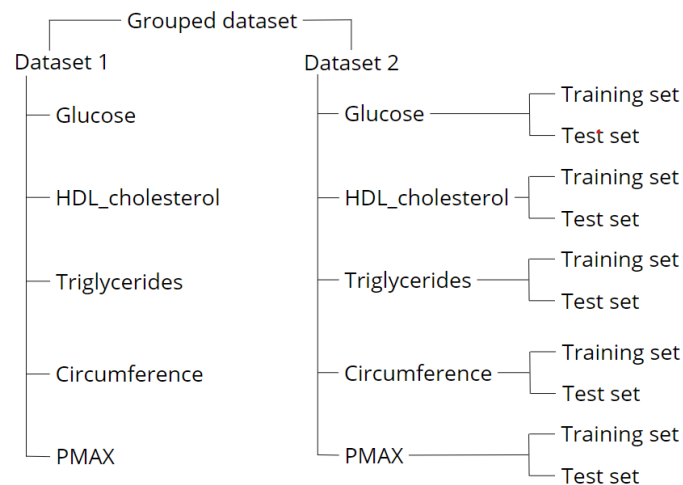


Figure 1.18: Division of the datasets



# 2 | Bayesian models for the metabolic syndrome

In this chapter, we describe the two-stage plug-in model. Which consists of three different Bayesian regression models, to predict the future values of the target variables, and the models to predict the overall risk of developing metabolic syndrome. For each stage, we describe the measures used for model comparison.

## 2.1. Mixed-effects models for longitudinal data

Longitudinal models are statistical methods used to analyze data collected over time from the same individuals or subjects. They allow to investigate within-subject changes over time, as well as between-subject variability in these changes. Longitudinal models provide a flexible framework and can account for various sources of variability.

We have first considered applying, ARIMA models but were discarded because the dataset presents irregular time intervals between measurements of the same variable. This violates the assumptions of equally-spaced observations of ARIMA models. Longitudinal models on the other hand do not have such limitations and are preferable for the kind of data provided by AVIS. Further details on ARIMA models can be found in Appendix B.2.

For this thesis, a random-effect model will be used in order to exploit the differences between individuals. Random-effect models are designed to model data at multiple levels by tuning different units of measurement nested within each other. For longitudinal models, the two levels consist of time which is nested within the individual, this allows us to differentiate between time-dependent variables and person-level variables that do not vary across time. In mixed-effects models we assume that the vector of repeated measurements for a specific subject can be represented by a linear regression model characterized by two types of parameters: fixed-effects parameters and random-effects parameters. Fixed-effects parameters are common to all individuals of the population and account for the within-subject variability. Random-effects parameters are unique to each individual subject and account for between-subject variability.

Laird and Ware (1982) introduce mixed-effects longitudinal models. Suppose we have  $N$  subjects, and let  $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, \dots, Y_{i,n_i})$  denote the  $n_i$ -dimensional vector of all repeated measurements for the  $i$ th subject, where  $n_i$  is the number of observations for the  $i$ th individual, for  $i = 1, 2, \dots, N$ . Let  $\beta$  be a  $p \times 1$  vector of unknown population parameters and  $X_i$  be a known  $n_i \times p$  design matrix linking  $\beta$  to  $\mathbf{Y}_i$ . Let  $\mathbf{b}_i$  be a  $k \times 1$  vector of unknown individual effects and  $Z_i$  be a known  $n_i \times k$  design matrix linking  $\mathbf{b}_i$  to  $\mathbf{Y}_i$ . We assume:

$$\begin{aligned} \mathbf{Y}_i &= \beta X_i + \mathbf{b}_i Z_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, R_i), \quad i = 1, \dots, N \\ \mathbf{b}_i &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, D), \end{aligned} \tag{2.1}$$

where  $R_i$  is a  $n_i \times n_i$  positive-definite covariance matrix,  $D$  is a  $k \times k$  positive-definite covariance matrix and  $\mathbf{b}_i$  are independent from  $\boldsymbol{\epsilon}_i$ . In this model  $\beta$  represents the fixed-effect parameters while  $\mathbf{b}_i$  are the random-effect parameters. We complete the model specification by introducing a prior for the regression parameters  $\beta$ .

## 2.2. Application of mixed-effects longitudinal models to the AVIS dataset

As explained before the metabolic syndrome can be diagnosed by observing five variables. The first step in assessing its onset risk will be to predict the future value of each of these target variables. Afterward, we will combine these results to predict the probability of the disease. To make predictions on the target variable, a mixed-effects longitudinal model will be applied to each of them separately.

The notation used is summarized here :

- Each target variable is associated with a number  $k = 1, \dots, 5$ 
  - Glucose corresponds to  $k = 1$
  - HDL\_cholesterol corresponds to  $k = 2$
  - Triglycerides corresponds to  $k = 3$
  - Circumference corresponds to  $k = 4$
  - PMAX corresponds to  $k = 5$
- Target variable  $k$  has a number of covariates equal to  $P_k$ ,  $k = 1, \dots, 5$ ;

- For any donor  $i = 1, \dots, N$ , the number of donations is expressed by  $j = 1, \dots, n_i^k$  where  $n_i^k$  is the total number of measurements of the variable  $k$  for donor  $i$ .
- The response  $Y_{i,j}^k$  denotes the target variable  $k$  and it is modeled by adapting Equation (2.1) as follows:

$$\begin{aligned} \mathbf{Y}_i^k &= \boldsymbol{\beta}^k X_i^k + \mathbf{b}_i^k Z_i^k + \boldsymbol{\epsilon}_i^k, & \boldsymbol{\epsilon}_i^k &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, R_i^k), & i = 1, \dots, N, \\ \mathbf{b}_i^k &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, D^k). & & & k = 1, \dots, 5 \end{aligned} \quad (2.2)$$

Where  $\boldsymbol{\beta}^k$  is the vector of fixed-effects for the target variable  $k$  and  $X_i^k$  is the design matrix linking  $\boldsymbol{\beta}^k$  to  $\mathbf{Y}_i^k$ .  $\mathbf{b}_i^k$  is the vector subject-specific random effects for the target variable  $k$  and  $Z_i^k$  is the design matrix linking  $\mathbf{b}_i^k$  to  $\mathbf{Y}_i^k$ .  $R_i$  and  $D$  are positive-definite covariance matrixes, and  $\mathbf{b}_i$  are independent from  $\boldsymbol{\epsilon}_i$ .

### 2.2.1. Covariates

In this analysis, we consider fixed-time and time-dependent donor-specific covariates. The maximum number of covariates is 36 (excluding interactions), and are reported in Tables 2.1, 2.2.

Feature selection is done by comparing the significant covariates obtained by fitting a model on Dataset 1 using horseshoes priors. Goodness-of-fit indicators were also used to compare the models obtained. The matrixes containing the data from the selected variables are the  $X_i^k$  matrix expressed in the model.

Further details on the selection model can be found in Appendix C.

Name	Type	Description
Gender	binary	Gender: 1 for males, 2 for female
AB0	factor	Blood type: A, B, AB, 0
Rh	binary	Rhesus factor: 1 for POS, 0 for NEG
Smoke	binary	Smoking habits: 1 for Smoker, 0 for Non-smoker
Alcohol	binary	Drinking habits: 1 for Habitual drinker, 0 for Non habitual drinker
Activity	factor	Physical activity habits: 1 for Active lifestyle, 2 for Moderate lifestyle, 3 for Sedentary lifestyle
Age	numeric	Age at the time of the donation
Age <sup>2</sup>	numeric	Age squared
Age <sup>3</sup>	numeric	Cubic Age
ALT	numeric	Alanine aminotransferase (ALT)
Albumin	numeric	Albumin
A1_globulins	numeric	Alpha-1-globulins
A2_globulins	numeric	Alpha-2-globulins
Basophils_perc	numeric	Percentage of basophils
B1_globulins	numeric	Beta-1-globulins
B2_globulins	numeric	Beta-2-globulins
BMI	numeric	Body Mass Index
Total_cholesterol	numeric	Total cholesterol
Creatinine	numeric	Creatinine
Volume_distribution	numeric	Red blood cell distribution width
Hematocrit	numeric	Hematocrit
Hemoglobin_mean_conc	numeric	Mean corpuscular hemoglobin concentration
Hemoglobin_mean_mass	numeric	Mean corpuscular hemoglobin
Hemoglobin	numeric	Hemoglobin

Table 2.1: Complete list of covariates

Name	Type	Description
Eosinophils_perc	numeric	Percentage of Eosinophils
Erythrocytes	numeric	Red blood cells
Ferritin	numeric	Ferritin
Total_iron	numeric	Total iron
G_globulins	numeric	Gamma globulins
Leukocytes	numeric	Leukocytes
Lymphocytes_perc	numeric	Percentage of lymphocytes
Monocytes_perc	numeric	Percentage of monocytes
Platelets	numeric	Platelets
Heart_rate	numeric	Heart rate
Total_proteins	numeric	Total proteins
Mean_Volume	numeric	Mean Cell Volume (MCV)

Table 2.2: Complete list of covariates (continued)

The sets of covariates and interactions selected for target variable are reported afterward.

- Glucose covariates:

- Gender
- AB0
- Rh
- Age
- Age<sup>2</sup>
- ALT
- Albumin
- A1\_globulins
- B1\_globulins
- BMI
- Volume\_distribution
- Ferritin
- Total\_iron
- G\_globulins
- Monocytes\_perc
- Heart\_rate
- Gender:Age
- Gender:A2\_globulins
- Gender:BMI
- Age:Ferritin
- Age:ALT
- Age:BMI

- HDL\_cholesterol covariates:

- Gender
- AB0
- Rh
- Smoke
- Alcohol
- Activity

- Age
- Age<sup>2</sup>
- Age<sup>3</sup>
- ALT
- Basophils\_perc
- B2\_globulins
- BMI
- Total\_cholesterol
- Volume\_distribution
- Hemoglobin
- G\_globulins
- Heart\_rate
- Mean\_Volume
- Gender:Total\_cholesterol
- Gender:Heart\_rate
- Gender:Platelets
- Gender:ALT
- Gender:G\_globulins
- Age:Hemoglobin\_mean\_conc
- Age:Platelets
- Age:A2\_globulins

- Triglycerides covariates:

- Gender
- AB0
- Rh
- Smoke
- Activity
- Age
- ALT
- A2\_globulins
- B1\_globulins
- B2\_globulins
- BMI
- Creatinine
- Total\_cholesterol
- Volume\_distribution
- Ferritin
- G\_globulins
- Leukocytes
- Lymphocytes\_perc
- Heart\_rate
- Gender:Mean\_volume
- Gender:Total\_cholesterol
- Gender:Hemoglobin
- Gender:B1\_globulins

- Circumference covariates:

- Gender
- AB0
- Activity
- Age
- ALT
- Albumin
- B1\_globulins
- B2\_globulins
- BMI
- Creatinine
- Hematocrit
- Erythrocytes
- Ferritin
- Total\_iron
- Leukocytes
- Heart\_rate
- Mean\_Volume
- Gender:Age
- Gender:Heart\_rate
- Gender:Platelets
- Gender:ALT
- Gender:B2\_globulins
- Gender:Mean\_volume
- Gender:BMI

- Age:Leukocytes
- Age:Albumin
- Age:ALT
- Age:BMI
- PMAX covariates:
  - Gender
  - ABO
  - Age
  - Age<sup>2</sup>
  - Age<sup>3</sup>
  - ALT
  - Albumin
  - A1\_globulins
  - Basophils\_perc
  - B1\_globulins
  - B2\_globulins
  - BMI
  - Total\_cholesterol
  - Creatinine
  - Volume\_distribution
  - Ferritin
  - Lymphocytes\_perc
  - Heart\_rate
  - Total\_proteins
  - Gender:Age
  - Gender:Lymphocytes\_perc
  - Gender:B2\_globulins
  - Gender:G\_globulins
  - Gender:BMI
  - Age:Mean\_volume
  - Age:Albumin
  - Age:B1\_globulins
  - Age:B2\_globulins
  - Age:G\_globulins
  - Age:ALT
  - Age:BMI
  - Age:Alcohol
  - Age:Volume\_distribution
  - Age:Lymphocytes
  - Age:Monocytes\_perc
  - Age:Albumin

### 2.2.2. Model 1

The first model, named *Model 1*, is a Bayesian regression model where time is taken into account only by time-varying covariates. This is a special case of (2.2), where the design matrix  $\mathbf{Z}_i^k$  is a null matrix. This model is the simplest among all other models and serves as a baseline for comparison with more complex models. Summing up, *Model 1* can be described as:

$$\begin{aligned}
 \mathbf{Y}_i^k &= \boldsymbol{\beta}^k X_i^k + \boldsymbol{\epsilon}_i^k, & \boldsymbol{\epsilon}_i^k &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, R_i^k), & i &= 1, \dots, N, \\
 \boldsymbol{\beta}^k &\sim \mathcal{N}(\mathbf{0}, \Sigma^k), & & & k &= 1, \dots, 5, \\
 R_i^k &\sim \mathcal{N}(\mu_e, \sigma_e).
 \end{aligned} \tag{2.3}$$

### 2.2.3. Model 2

The second model, *Model 2*, introduces a donor-specific random effect to consider the between-subject variability. In this model the design matrix  $\mathbf{Z}_i^k$  is a vector of ones to introduce the random effect only as a subject-specific intercept.

Summing up, *Model 2* can be described as:

$$\begin{aligned}
\mathbf{Y}_i^k &= X_i^k \boldsymbol{\beta}^k + Z_i^k b_i^k + \boldsymbol{\epsilon}_i^k, & \boldsymbol{\epsilon}_i^k &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, R_i^k), & i = 1, \dots, N, \\
Z_i^k &= [1, \dots, 1]', & & & k = 1, \dots, 5, \\
R_i^k &\sim \mathcal{N}(\mu_e, \sigma_e), \\
\boldsymbol{\beta}^k &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^k), \\
b_i^k &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \eta^2), \\
\mu &\sim \mathcal{N}(\mu_0, \sigma_{\mu_0}^2), \\
\eta^2 &\sim \text{inv-gamma}(\alpha_0, \beta_0).
\end{aligned} \tag{2.4}$$

Where  $b_i^k$ ,  $R_i^k$ ,  $\mu$  and  $\eta^2$  are exchangeable priors. This allows us to assume that the order in which we observe the data is not important

### 2.2.4. Model 3

The third model, *Model 3*, adds to *Model 2* a donor-specific random slope for the variable BMI:

$$\begin{aligned}
\mathbf{Y}_i^k &= X_i^k \boldsymbol{\beta}^k + Z_i^k b_i^k + \boldsymbol{\epsilon}_i^k, & \boldsymbol{\epsilon}_i^k &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, R_i^k), & i = 1, \dots, N, \\
Z_i^k &= \begin{bmatrix} 1, & \dots, & 1 \\ BMI_1, & \dots, & BMI_{n_i} \end{bmatrix}', & & & k = 1, \dots, 5, \\
R_i^k &\sim \mathcal{N}(\mu_e, \sigma_e), \\
\boldsymbol{\beta}^k &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^k), \\
\mathbf{b}_i^k &\stackrel{\text{iid}}{\sim} \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \eta_1^2 & 0 \\ 0 & \eta_2^2 \end{bmatrix}\right), \\
\mu_l &\sim \mathcal{N}(\mu_0, \sigma_{\mu_0}^2) & l = 1, 2, \\
\eta_l^2 &\sim \text{inv-gamma}(\alpha_0, \beta_0) & l = 1, 2.
\end{aligned} \tag{2.5}$$

### 2.2.5. Model selection for Bayesian models

To select the best possible model for our work, we compare them via goodness-of-fit criteria. The goodness-of-fit indicators used are the Watanabe-Akaike information criterion (*WAIC*) and leave-one-out cross validation (*LOO*). *WAIC*, introduced by Watanabe (2010), is a predictive goodness-of-fit tool that approximates the log point-wise predictive density. It is the generalized version of the Akaike information criterion (*AIC*), and it is computed as:

$$WAIC = -2LPPD + 2p_{WAIC}, \quad (2.6)$$

where *LPPD* is the log pointwise predictive density, which is computed as:

$$LPPD = \sum_{i=1}^n \log \int p(y_i|\theta)p_{post}(\theta)d\theta, \quad (2.7)$$

and  $p_{WAIC}$  is the penalty term of *WAIC* and represents the variance of individual terms in the log predictive density summed over the  $n$  data points. This term is also fully Bayesian and can be expressed as:

$$p_{WAIC} = \sum_{i=1}^n \text{var}_{post}(\log p(y_i|\theta)). \quad (2.8)$$

*LOO* unlike *WAIC* does not require a penalty term as explained in Gelmane et al.. It is computed as:

$$LOO = -2LPPD_{loo} = -2 \sum_{i=1}^n \log \int p(y_i|\theta)p_{post(-i)}(\theta)d\theta, \quad (2.9)$$

where  $p_{post(-i)}(\theta)$  is the posterior distribution based on the data minus data point  $i$ .

Unlike *LPPD* which uses data point  $i$  for both the computation of posterior distribution and the prediction, here  $LPPD_{loo}$  only uses it for prediction, and hence there is no need for a penalty term to correct the potential bias introduced by using data twice.

Both *WAIC* and *LOO* indicate better models when their values are lower. It is important to note that these indicators provide information only about the relative quality of a model compared to other models, not about the absolute quality of a model.

## 2.3. Prediction model

After applying these three models and choosing the best one using goodness-of-fit indicators, the selected model was applied to each target variable to infer future values. Since we are using a Bayesian model, the outcome for each variable and donor is represented as a distribution, known as the posterior predictive distribution. This distribution describes the uncertainty on the value that the target variable will take, given all other available knowledge. It is calculated as the average of the probability distributions of the target variable conditional on all the unknown parameters, weighted with the posterior distribution of the parameters given all known observations. To predict the presence of metabolic syndrome in a patient, we examine the probability that this distribution exceeds, or falls below, as in the case of HDL\_cholesterol, the following target variable thresholds for at least three target variables:

- PMAX>130 mm Hg;
- Glucose>100 mg/dL;
- Circumference>102 cm in men or Circumference>89 cm in women;
- Triglycerides>150 mg/dL;
- HDL cholesterol<40 mg/dL in men or HDL cholesterol<50 mg/dL in women.

From each distribution  $Y_i^k$ , we define  $w_i^k$  as the probability of the target variable  $k$  of falling outside the specified threshold. We then construct a vector containing all the target variables probabilities:  $\mathbf{w}_i = (w_i^1, w_i^2, w_i^3, w_i^4, w_i^5)$ . Collectively, these vectors form the matrix  $W$ . These individual probabilities are then used to calculate the overall probability of developing metabolic syndrome, denoted as  $p_i(\mathbf{w}_i)$ .

Two different methods are used in this work to obtain the total probability  $p_i(\mathbf{w}_i)$ , a logistic regression and the mathematical computation that at least three target variables fall outside the normal range.

### 2.3.1. Logistic regression model

The logistic regression is a classification technique used to predict binary responses. It uses the logistic function to model the probability of the binary outcome:

$$p_i(\mathbf{w}_i) = \frac{1}{1 + e^{-v_i}}. \quad (2.10)$$

where  $v_i$  is the linear combination of the predictors,

$$v_i = c_0 + \sum_{k=1}^5 c_k w_{ik}. \quad (2.11)$$

$w_{ik}$  are the elements of the matrix  $W$  and  $c_k$  are the corresponding coefficients. Figure 2.1 shows a representation of the logistic function often also called sigmoid function.

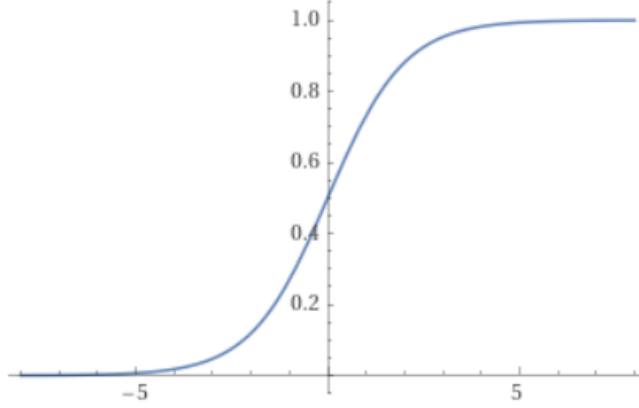


Figure 2.1: Logistic function

### 2.3.2. Combinatorial model

A combination  $C(n, k)$  is a selection of  $k$  elements from a set composed of  $n$  elements where the order of the elements does not matter.

We propose to model  $p_i(\mathbf{w}_i)$  as the probability that at least 3 out of the 5 variables are out of bounds. The vector  $\mathbf{w}_i$  contains the individual probabilities of the event happening.  $p_i(\mathbf{w}_i)$  will then be the sum of the probability of exactly 3 of the target variables being out of bound, exactly 4 and exactly five. Unlike the logistic regression model which introduces an ulterior error factor this model is the mathematically correct one. We know from combinatorial theory that the number of combinations will be given by:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

So we will have to compute 10 combinations to calculate the probability of three events happening out of five, 5 for four events, and 1 for all five events happening simultaneously. For each of these combinations, we multiply the probabilities  $w_i$  of the events happening and  $1 - w_i$  for the events not taken into consideration. The resulting product will be referred to as  $m_c$  with  $c = 1, \dots, 16$ . The final probability of developing the metabolic

syndrome  $p_i(\mathbf{w}_i)$  will be:

$$p_i(\mathbf{w}_i) = \sum_{c=1}^{16} m_c \quad (2.12)$$

### 2.3.3. Model selection for prediction models

To choose between the logistic regression model or the combinatorial model we take into consideration different performance measures. After using the prediction models we will have for each instance of our dataset an observed label and a predicted label. This allows us to identify four different outcomes:

- True positive (TP): correct positive prediction,
- False positive (FP): incorrect positive prediction,
- True negative (TN): correct negative prediction,
- False negative (FN): incorrect negative prediction.

These outcomes can be organized into a  $2 \times 2$  table called the *confusion matrix* for binary classification. The most common measures derived from a confusion matrix are the *accuracy* (ACC) and consequently the *error rate* (ERR). The ACC measures the proportion of correct classifications:

$$ACC = \frac{TP + TN}{TP + TN + FN + FP}, \quad (2.13)$$

whereas the ERR is the proportion of incorrect predictions:

$$ERR = \frac{FP + FN}{TP + TN + FN + FP} = 1 - ACC. \quad (2.14)$$

While both accuracy and error rate are very useful, they do not account for the different costs of errors in positive and negative predictions. Therefore, other measures are introduced.

*Sensitivity* (SN) also called *recall* or *true positive rate* is the proportion of correct positive predictions out of all the positive predictions:

$$SN = \frac{TP}{TP + FN}. \quad (2.15)$$

*Specificity* (SP) also called *true negative rate* is calculated as the number of correct neg-

ative predictions divided by the total number of negatives:

$$SP = \frac{TN}{TN + FP}. \quad (2.16)$$

*Precision* (PR) also called *positive predictive value* measures how good the model is at assigning positive events to the positive class:

$$PR = \frac{TP}{TP + FP}. \quad (2.17)$$

Finally, the *F1-score* ( $F_1$ ) is a harmonic mean of the precision and the recall:

$$F_1 = \frac{2 \cdot PR \cdot REC}{PR + REC}. \quad (2.18)$$

To assess whether the values obtained are good we have to compare them to those of other models, usually, they are compared to the models that assign to all predictions the positive or negative class. Another useful model to compare is the one that classifies the instances randomly.



# 3 | Posterior analysis

In this chapter we illustrate the posterior inference for the Bayesian models described in Chapter 2. They are applied to the AVIS data presented in Chapter 1. The focus is on *Model 2* which, as a result of the analysis carried out, turns out to be the most performing model in terms of WAIC and LOO.

## 3.1. Stan Software

Sampling from the posterior distribution is achieved via the software platform called Stan, which is a probabilistic programming language for statistical inference written in C++. In particular, Stan uses Hamiltonian Monte Carlo (HMC), a family of MCMC algorithms which promise improved efficiency and faster inference (Stan Development Team (2020)). Sampling in all the analysis in this chapter was done using 3 chains with 3000 iterations each and a 50% burn in consisting of 1500 iterations.

## 3.2. Comparison of the models

The convergence diagnostics of the different simulations have been checked, showing that all the MCMC chains reach stationarity as seen in Appendix D. Table 3.1 shows the WAIC values of each model and target variable while Table 3.2 shows the LOO values. It is immediately evident that Models 2 and 3 perform better than Model 1 for all target variables. Model 3 does not show a significant performance improvement against Model 2 even if it is more computationally complex. Therefore Model 2 was chosen as the best model, and all the following analyses will be based on its posterior results.

	Model 1	Model 2	Model 3
Glucose	-61384.2	-73588.1	-73778.9
HDL_cholesterol	-4728.3	-30772.7	-31217.9
Triglycerides	-38821.4	-52718.3	-52667.4
Circumference	-69990.1	-89117.3	-89042.5
PMAX	-103482.0	-114739.7	-117721.5

Table 3.1: WAIC value for each model

	Model 1	Model 2	Model 3
Glucose	-61384.1	-73612.5	-73692.8
HDL_cholesterol	-4728.2	-30292.1	-31053.6
Triglycerides	-38821.4	-52454.8	-52584.3
Circumference	-70277.2	-91665.3	-91247.9
PMAX	-103481.6	-114690.9	-117468.2

Table 3.2: LOO value for each model

### 3.3. Posterior inference for Model 2

In this section, we will analyze the posterior distribution for both the fixed effect and random effect coefficients under *Model 2*. This analysis will help us better understand the relationship between the target variable and the coefficients. Particular emphasis has been put on the covariates related to the donor's lifestyle, such as BMI, smoking habits, and activity levels. This focus is due to the fact that these aspects of a donor's life are the ones a doctor will recommend improving upon when diagnosing metabolic syndrome.

#### 3.3.1. Fixed effects: $\beta$ regression coefficients

A summary of the  $\beta$ 's posterior densities is reported and analyzed for each target variable. Furthermore, a graph containing the 95% credible intervals of all coefficients is shown, aiding the understanding of significant covariates and their effect on the target variable considered.

## Glucose

In Table 3.3 a summary of the posterior densities for the components of  $\beta^{(1)}$  are shown.

Parameter	Mean	SD	Q0.025-	Q0.5	Q0.975
$\beta_{Age}$	$2.4 \cdot 10^{-2}$	$1.4 \cdot 10^{-3}$	$2.4 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$	$2.7 \cdot 10^{-2}$
$\beta_{B1\_globulins}$	$5.6 \cdot 10^{-3}$	$9.6 \cdot 10^{-4}$	$3.8 \cdot 10^{-3}$	$5.6 \cdot 10^{-3}$	$7.5 \cdot 10^{-3}$
$\beta_{BMI}$	$5.4 \cdot 10^{-3}$	$9.6 \cdot 10^{-4}$	$3.5 \cdot 10^{-3}$	$5.4 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$
$\beta_{ALT}$	$2.8 \cdot 10^{-3}$	$7.6 \cdot 10^{-4}$	$1.4 \cdot 10^{-3}$	$2.8 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$
$\beta_{Total\_proteins}$	$2.0 \cdot 10^{-3}$	$7.7 \cdot 10^{-4}$	$5.7 \cdot 10^{-4}$	$2.0 \cdot 10^{-3}$	$3.5 \cdot 10^{-3}$
$\beta_{A2\_globulins}$	$1.8 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	$-1.9 \cdot 10^{-4}$	$1.8 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$
$\beta_{A1\_globulins}$	$1.5 \cdot 10^{-3}$	$8.6 \cdot 10^{-4}$	$-1.7 \cdot 10^{-4}$	$1.5 \cdot 10^{-3}$	$3.2 \cdot 10^{-3}$
$\beta_{Age:ALT}$	$1.3 \cdot 10^{-3}$	$7.6 \cdot 10^{-4}$	$2.8 \cdot 10^{-5}$	$1.3 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$
$\beta_{Ferritin}$	$1.3 \cdot 10^{-3}$	$8.0 \cdot 10^{-4}$	$-3.1 \cdot 10^{-4}$	$1.2 \cdot 10^{-3}$	$2.8 \cdot 10^{-3}$
$\beta_{Age^2}$	$9.7 \cdot 10^{-4}$	$6.8 \cdot 10^{-4}$	$-4.0 \cdot 10^{-4}$	$9.8 \cdot 10^{-4}$	$2.3 \cdot 10^{-3}$
$\beta_{Heart\_rate}$	$8.9 \cdot 10^{-4}$	$5.5 \cdot 10^{-4}$	$-1.5 \cdot 10^{-4}$	$8.7 \cdot 10^{-4}$	$2.0 \cdot 10^{-3}$
$\beta_{Age:BMI}$	$7.6 \cdot 10^{-4}$	$7.6 \cdot 10^{-4}$	$-7.1 \cdot 10^{-4}$	$7.7 \cdot 10^{-4}$	$2.3 \cdot 10^{-3}$
$\beta_{Age:Ferritin}$	$4.6 \cdot 10^{-4}$	$6.4 \cdot 10^{-4}$	$-7.9 \cdot 10^{-4}$	$4.5 \cdot 10^{-4}$	$1.7 \cdot 10^{-3}$
$\beta_{Gender2:BMI}$	$-7.1 \cdot 10^{-4}$	$1.8 \cdot 10^{-3}$	$-4.3 \cdot 10^{-3}$	$-7.2 \cdot 10^{-4}$	$2.9 \cdot 10^{-3}$
$\beta_{Total\_iron}$	$-1.2 \cdot 10^{-3}$	$6.1 \cdot 10^{-4}$	$-2.4 \cdot 10^{-3}$	$-1.2 \cdot 10^{-3}$	$-2.7 \cdot 10^{-5}$
$\beta_{Monocytes\_perc}$	$-1.7 \cdot 10^{-3}$	$7.1 \cdot 10^{-4}$	$-3.0 \cdot 10^{-3}$	$-1.7 \cdot 10^{-3}$	$-2.6 \cdot 10^{-4}$
$\beta_{Volume\_distribution}$	$-1.7 \cdot 10^{-3}$	$8.2 \cdot 10^{-4}$	$-3.2 \cdot 10^{-3}$	$-1.7 \cdot 10^{-3}$	$-1.4 \cdot 10^{-5}$
$\beta_G\_globulins$	$-3.0 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$	$-5.4 \cdot 10^{-3}$	$-3.0 \cdot 10^{-3}$	$-7.6 \cdot 10^{-4}$
$\beta_{Gender2:A2\_globulins}$	$-4.0 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$	$-8.5 \cdot 10^{-3}$	$-4.1 \cdot 10^{-3}$	$4.4 \cdot 10^{-4}$
$\beta_{Gender2}$	$-5.1 \cdot 10^{-3}$	$3.7 \cdot 10^{-3}$	$-1.2 \cdot 10^{-2}$	$-5.0 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$
$\beta_{RhPOS}$	$-6.6 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$-1.4 \cdot 10^{-2}$	$-6.6 \cdot 10^{-3}$	$3.9 \cdot 10^{-4}$
$\beta_{Gender2:Age}$	$-9.5 \cdot 10^{-3}$	$2.4 \cdot 10^{-3}$	$-1.4 \cdot 10^{-2}$	$-9.5 \cdot 10^{-3}$	$-4.8 \cdot 10^{-3}$

Table 3.3: Posterior summaries of the  $\beta$ s parameters for the log Glucose

Figure 3.1 shows  $\beta^{(1)}$ 's credible intervals. In particular the covariates in red increase the Glucose value and consequentially the risk of metabolic syndrome; those in green decrease the Glucose value and consequentially the risk of metabolic syndrome. Finally, the covariates in grey are not significant as their 95% credibility intervals contain 0. This color legend also applies to the credible intervals of the other target variables apart from HDL\_cholesterol.

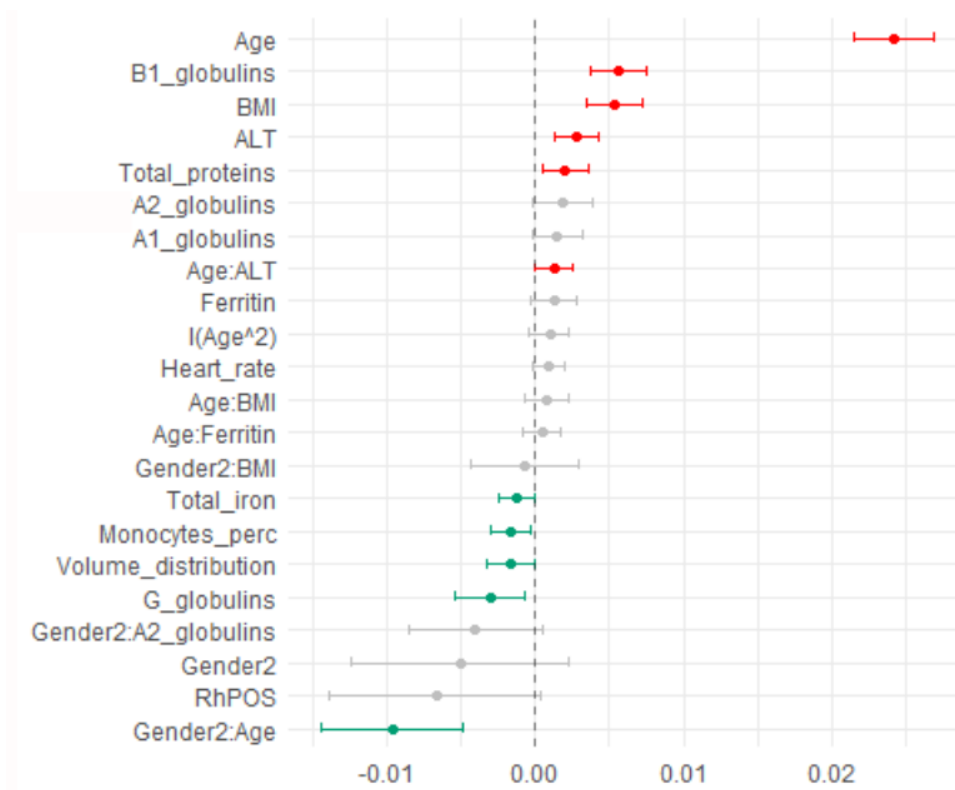


Figure 3.1:  $\beta^{(1)}$ 's posterior credible intervals, under *Model 2*

First of all, from the analysis of Figure 3.1, we notice that as expected the age of the donor has a positive effect meaning that the older a person is, the higher their glucose levels. We can also see that the age of the donor affects men more heavily than women, and BMI also has a positive effect, while its relationship with age does not impact glucose levels. Moreover, all the serum proteins levels effects are consistent with the literature. Indeed we can see that B1\_globulins increase the glucose levels, while G\_globulins decrease it. A1 and A2 globulins are not significant but they still have a positive mean. See Gul and Rahman (2006). In general, the Total\_protein variable has a positive effect on glucose levels.

## HDL\_cholesterol

In Table 3.4 a summary of the posterior densities for the components of  $\beta^{(2)}$  are shown.

Parameter	Mean	SD	Q <sub>0.025</sub>	Q <sub>0.5</sub>	Q <sub>0.975</sub>
$\beta_{Gender2}$	$1.9 \cdot 10^{-1}$	$9.6 \cdot 10^{-3}$	$1.7 \cdot 10^{-1}$	$1.9 \cdot 10^{-1}$	$2.0 \cdot 10^{-1}$
$\beta_{Mean\_Volume}$	$2.0 \cdot 10^{-2}$	$1.9 \cdot 10^{-3}$	$1.7 \cdot 10^{-2}$	$2.0 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$
$\beta_{Smoke}$	$1.7 \cdot 10^{-2}$	$3.8 \cdot 10^{-3}$	$9.4 \cdot 10^{-3}$	$1.7 \cdot 10^{-2}$	$2.4 \cdot 10^{-2}$
$\beta_{Gender2:Total\_cholesterol}$	$1.4 \cdot 10^{-2}$	$4.0 \cdot 10^{-3}$	$6.1 \cdot 10^{-3}$	$1.4 \cdot 10^{-2}$	$2.2 \cdot 10^{-2}$
$\beta_{Gender2:ALT}$	$8.9 \cdot 10^{-3}$	$3.5 \cdot 10^{-3}$	$2.0 \cdot 10^{-3}$	$8.9 \cdot 10^{-3}$	$1.6 \cdot 10^{-2}$
$\beta_{Volume\_distribution}$	$6.5 \cdot 10^{-3}$	$1.5 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$6.5 \cdot 10^{-3}$	$9.4 \cdot 10^{-3}$
$\beta_{A1\_globulins}$	$6.4 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	$3.1 \cdot 10^{-3}$	$6.4 \cdot 10^{-3}$	$9.6 \cdot 10^{-3}$
$\beta_{A2\_globulins}$	$5.1 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$	$5.1 \cdot 10^{-3}$	$8.4 \cdot 10^{-3}$
$\beta_{Age:Platelets}$	$1.9 \cdot 10^{-3}$	$1.5 \cdot 10^{-3}$	$-9.0 \cdot 10^{-4}$	$1.9 \cdot 10^{-3}$	$4.8 \cdot 10^{-3}$
$\beta_{Age^3}$	$1.0 \cdot 10^{-3}$	$7.7 \cdot 10^{-4}$	$-5.6 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$
$\beta_{Gender2:Heart\_rate}$	$-7.8 \cdot 10^{-5}$	$2.6 \cdot 10^{-3}$	$-5.1 \cdot 10^{-3}$	$-8.5 \cdot 10^{-5}$	$5.0 \cdot 10^{-3}$
$\beta_{Age:A2\_globulins}$	$-1.4 \cdot 10^{-4}$	$1.3 \cdot 10^{-3}$	$-2.7 \cdot 10^{-3}$	$-1.6 \cdot 10^{-4}$	$2.4 \cdot 10^{-3}$
$\beta_{Age:Hemoglobin\_mean\_conc}$	$-5.6 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$	$-2.7 \cdot 10^{-3}$	$-5.6 \cdot 10^{-4}$	$1.6 \cdot 10^{-3}$
$\beta_{Activity\_Moderate\_lifestyle}$	$-6.9 \cdot 10^{-4}$	$2.8 \cdot 10^{-3}$	$-6.2 \cdot 10^{-3}$	$-6.9 \cdot 10^{-4}$	$5.0 \cdot 10^{-3}$
$\beta_{Heart\_rate}$	$-2.7 \cdot 10^{-3}$	$9.5 \cdot 10^{-4}$	$-4.5 \cdot 10^{-3}$	$-2.7 \cdot 10^{-3}$	$-8.6 \cdot 10^{-4}$
$\beta_{Platelets}$	$-3.5 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	$-7.3 \cdot 10^{-3}$	$-3.5 \cdot 10^{-3}$	$1.1 \cdot 10^{-4}$
$\beta_{Gender2:G\_globulins}$	$-3.6 \cdot 10^{-3}$	$4.8 \cdot 10^{-3}$	$-1.3 \cdot 10^{-2}$	$-3.6 \cdot 10^{-3}$	$5.9 \cdot 10^{-3}$
$\beta_{Age^3}$	$-3.6 \cdot 10^{-3}$	$2.2 \cdot 10^{-3}$	$-8.0 \cdot 10^{-3}$	$-3.6 \cdot 10^{-3}$	$-6.0 \cdot 10^{-4}$
$\beta_{G\_globulins}$	$-4.7 \cdot 10^{-3}$	$2.6 \cdot 10^{-3}$	$-9.9 \cdot 10^{-3}$	$-4.7 \cdot 10^{-3}$	$3.6 \cdot 10^{-4}$
$\beta_{Total\_cholesterol}$	$-7.0 \cdot 10^{-3}$	$1.6 \cdot 10^{-3}$	$-1.0 \cdot 10^{-2}$	$-7.0 \cdot 10^{-3}$	$-3.9 \cdot 10^{-3}$
$\beta_{Alcohol}$	$-7.2 \cdot 10^{-3}$	$2.9 \cdot 10^{-3}$	$-1.3 \cdot 10^{-2}$	$-7.1 \cdot 10^{-3}$	$-1.6 \cdot 10^{-3}$
$\beta_{B2\_globulins}$	$-8.8 \cdot 10^{-3}$	$2.0 \cdot 10^{-3}$	$-1.3 \cdot 10^{-2}$	$-8.8 \cdot 10^{-3}$	$-4.9 \cdot 10^{-3}$
$\beta_{Gender2:Platelets}$	$-1.2 \cdot 10^{-2}$	$4.0 \cdot 10^{-3}$	$-2.0 \cdot 10^{-2}$	$-1.2 \cdot 10^{-2}$	$-4.0 \cdot 10^{-3}$
$\beta_{Activity\_Sedentary\_lifestyle}$	$-1.2 \cdot 10^{-2}$	$3.4 \cdot 10^{-3}$	$-1.8 \cdot 10^{-2}$	$-1.2 \cdot 10^{-2}$	$-5.4 \cdot 10^{-3}$
$\beta_{ALT}$	$-1.4 \cdot 10^{-2}$	$1.4 \cdot 10^{-3}$	$-1.6 \cdot 10^{-2}$	$-1.4 \cdot 10^{-2}$	$-1.1 \cdot 10^{-2}$
$\beta_{Hemoglobin}$	$-1.4 \cdot 10^{-2}$	$1.8 \cdot 10^{-3}$	$-1.8 \cdot 10^{-2}$	$-1.4 \cdot 10^{-2}$	$-1.1 \cdot 10^{-2}$
$\beta_{Age}$	$-1.6 \cdot 10^{-2}$	$3.2 \cdot 10^{-3}$	$-2.2 \cdot 10^{-2}$	$-1.6 \cdot 10^{-2}$	$-9.0 \cdot 10^{-3}$
$\beta_{BMI}$	$-1.8 \cdot 10^{-2}$	$1.3 \cdot 10^{-3}$	$-2.1 \cdot 10^{-2}$	$-1.8 \cdot 10^{-2}$	$-1.6 \cdot 10^{-2}$

Table 3.4: Posterior summaries of the  $\beta$ s parameters for the log HDL\_cholesterol

Figure 3.2 shows  $\beta^{(2)}$ 's credible intervals. In particular, the covariates in red decrease the

logarithm of the HDL\_cholesterol value and consequentially increase the risk of metabolic syndrome, those in green increase the HDL\_cholesterol value and consequentially decrease the risk of metabolic syndrome. Finally, the covariates in grey are not significant as their 95% credibility intervals contain 0.

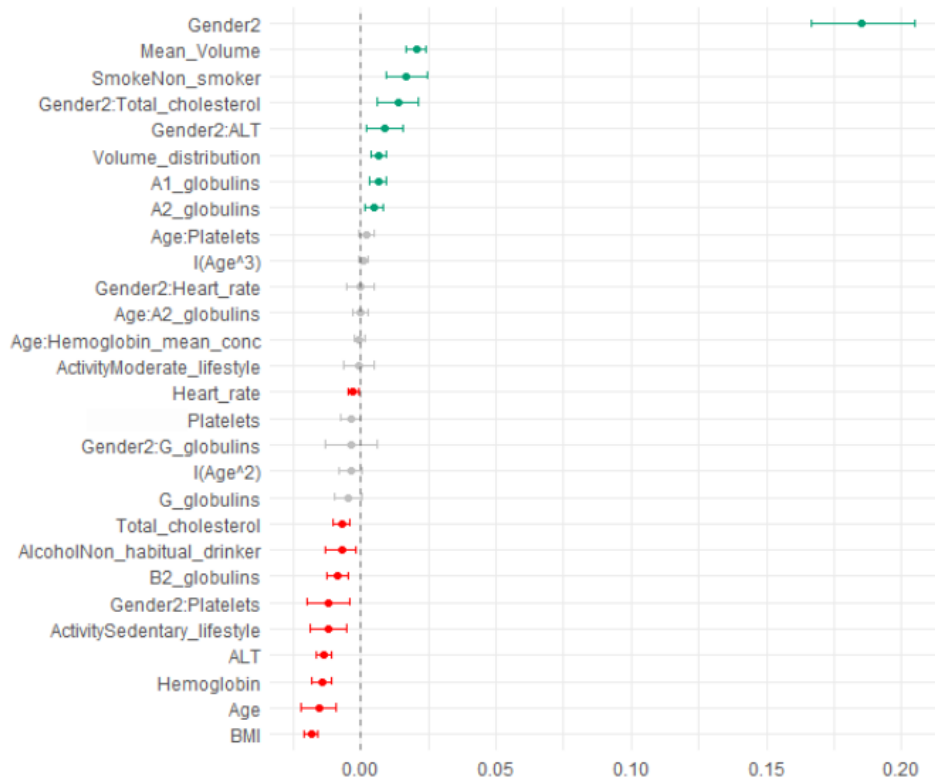


Figure 3.2:  $\beta^{(2)}$ 's posterior credible intervals, under *Model 2*

The analysis of Figure 3.2 leads to the following conclusion. As seen in Chapter 1.4.2 Figure 3.2 shows that women have a higher level of HDL\_cholesterol which is also considered when diagnosing metabolic syndrome. As expected Age and BMI have a negative effect as having a sedentary lifestyle and smoking. Variables such as Mean\_Volume, Hemoglobin and Platelets have effects consistent with literature about the topic. (See Fessler et al. (2013)).

## Triglycerides

Table 3.5 contains a summary of the posterior densities for the components of  $\beta^{(3)}$ .

Parameter	Mean	SD	Q0.025-	Q0.5	Q0.975
$\beta_{Age}$	$2.1 \cdot 10^{-2}$	$1.1 \cdot 10^{-3}$	$1.8 \cdot 10^{-2}$	$2.1 \cdot 10^{-2}$	$2.3 \cdot 10^{-2}$
$\beta_{Total\_cholesterol}$	$1.0 \cdot 10^{-2}$	$8.8 \cdot 10^{-4}$	$8.3 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$
$\beta_{BMI}$	$8.6 \cdot 10^{-3}$	$8.1 \cdot 10^{-4}$	$7.1 \cdot 10^{-3}$	$8.6 \cdot 10^{-3}$	$1.0 \cdot 10^{-2}$
$\beta_{Activity\_Sedentary\_lifestyle}$	$7.2 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	$3.8 \cdot 10^{-3}$	$7.3 \cdot 10^{-3}$	$1.1 \cdot 10^{-2}$
$\beta_{Gender2:Hemoglobin}$	$6.7 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	$3.2 \cdot 10^{-3}$	$6.7 \cdot 10^{-3}$	$1.0 \cdot 10^{-2}$
$\beta_{Leukocytes}$	$6.6 \cdot 10^{-3}$	$8.6 \cdot 10^{-4}$	$5.0 \cdot 10^{-3}$	$6.6 \cdot 10^{-3}$	$8.3 \cdot 10^{-3}$
$\beta_{Hemoglobin}$	$5.4 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	$3.3 \cdot 10^{-3}$	$5.4 \cdot 10^{-3}$	$7.4 \cdot 10^{-3}$
$\beta_{ALT}$	$5.4 \cdot 10^{-3}$	$7.1 \cdot 10^{-4}$	$4.0 \cdot 10^{-3}$	$5.4 \cdot 10^{-3}$	$6.8 \cdot 10^{-3}$
$\beta_{B2\_globulins}$	$5.0 \cdot 10^{-3}$	$9.4 \cdot 10^{-4}$	$3.1 \cdot 10^{-3}$	$5.0 \cdot 10^{-3}$	$6.9 \cdot 10^{-3}$
$\beta_{B1\_globulins}$	$3.5 \cdot 10^{-3}$	$8.9 \cdot 10^{-4}$	$1.8 \cdot 10^{-3}$	$3.5 \cdot 10^{-3}$	$5.2 \cdot 10^{-3}$
$\beta_{Lymphocytes\_perc}$	$2.7 \cdot 10^{-3}$	$7.8 \cdot 10^{-4}$	$1.2 \cdot 10^{-3}$	$2.7 \cdot 10^{-3}$	$4.2 \cdot 10^{-3}$
$\beta_{Creatinine}$	$2.0 \cdot 10^{-3}$	$8.6 \cdot 10^{-4}$	$3.8 \cdot 10^{-4}$	$2.0 \cdot 10^{-3}$	$3.7 \cdot 10^{-3}$
$\beta_{Ferritin}$	$1.5 \cdot 10^{-3}$	$7.2 \cdot 10^{-4}$	$5.9 \cdot 10^{-5}$	$1.5 \cdot 10^{-3}$	$2.9 \cdot 10^{-3}$
$\beta_{Activity\_Moderate\_lifestyle}$	$9.8 \cdot 10^{-4}$	$1.6 \cdot 10^{-3}$	$-2.1 \cdot 10^{-3}$	$9.5 \cdot 10^{-4}$	$4.1 \cdot 10^{-3}$
$\beta_{Gender2:B1\_globulins}$	$1.3 \cdot 10^{-4}$	$2.0 \cdot 10^{-3}$	$-3.8 \cdot 10^{-3}$	$1.2 \cdot 10^{-4}$	$4.0 \cdot 10^{-3}$
$\beta_{Heart\_rate}$	$8.3 \cdot 10^{-5}$	$5.1 \cdot 10^{-4}$	$-9.2 \cdot 10^{-4}$	$8.2 \cdot 10^{-5}$	$1.1 \cdot 10^{-3}$
$\beta_{RhPOS}$	$-1.1 \cdot 10^{-4}$	$3.7 \cdot 10^{-3}$	$-7.2 \cdot 10^{-3}$	$-7.1 \cdot 10^{-5}$	$7.3 \cdot 10^{-3}$
$\beta_{Gender2:Total\_cholesterol}$	$-8.8 \cdot 10^{-4}$	$2.1 \cdot 10^{-3}$	$-5.0 \cdot 10^{-3}$	$-8.9 \cdot 10^{-4}$	$3.2 \cdot 10^{-3}$
$\beta_{Volume\_distribution}$	$-1.9 \cdot 10^{-3}$	$8.1 \cdot 10^{-4}$	$-3.5 \cdot 10^{-3}$	$-1.9 \cdot 10^{-3}$	$-2.5 \cdot 10^{-4}$
$\beta_{Mean\_Volume}$	$-2.3 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$	$-4.4 \cdot 10^{-3}$	$-2.3 \cdot 10^{-3}$	$-1.5 \cdot 10^{-4}$
$\beta_{A2\_globulins}$	$-3.2 \cdot 10^{-3}$	$9.0 \cdot 10^{-4}$	$-4.9 \cdot 10^{-3}$	$-3.2 \cdot 10^{-3}$	$-1.3 \cdot 10^{-3}$
$\beta_{Smoke}$	$-3.8 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	$-7.5 \cdot 10^{-3}$	$-3.9 \cdot 10^{-3}$	$-6.9 \cdot 10^{-5}$
$\beta_{G\_globulins}$	$-4.2 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$	$-6.4 \cdot 10^{-3}$	$-4.2 \cdot 10^{-3}$	$-2.1 \cdot 10^{-3}$
$\beta_{Gender2:Mean\_Volume}$	$-8.2 \cdot 10^{-3}$	$2.1 \cdot 10^{-3}$	$-1.2 \cdot 10^{-2}$	$-8.3 \cdot 10^{-3}$	$-4.1 \cdot 10^{-3}$

Table 3.5: Posterior summaries of the  $\beta$ s parameters for the log Triglycerides

Figure 3.3 shows  $\beta^{(3)}$  credible intervals.

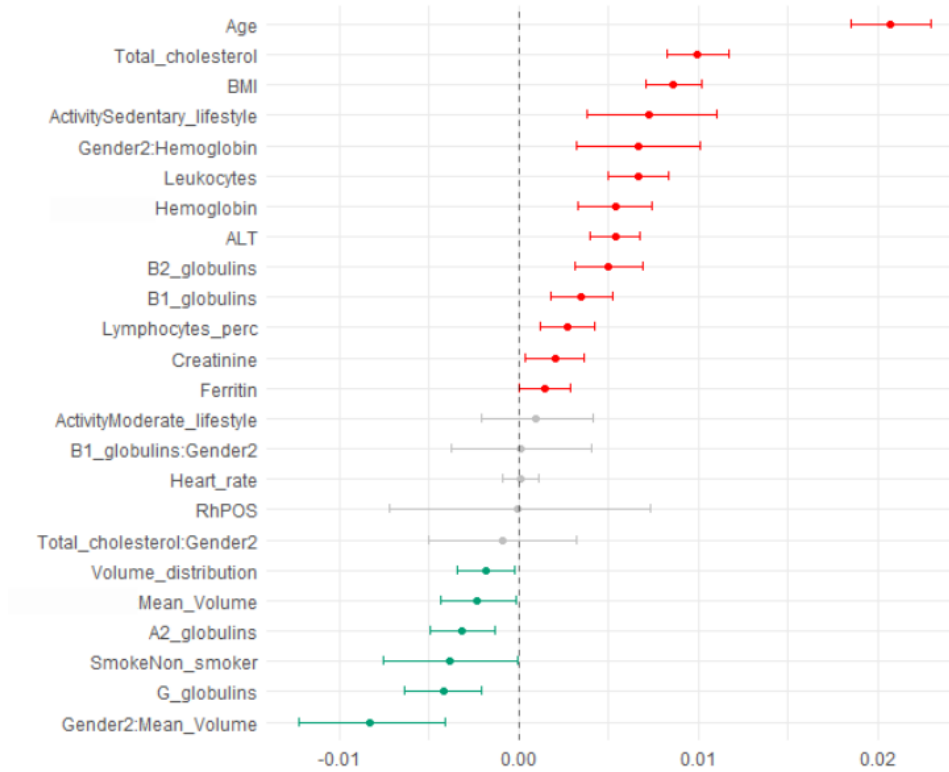


Figure 3.3:  $\beta^{(3)}$ 's posterior credible intervals, under *Model 2*

From Figure 3.3 we can see that age has a positive effect on triglyceride levels and increases the risk of metabolic syndrome. Variables related to the donor lifestyle influence the triglycerides levels as expected, BMI and a sedentary lifestyle have positive effects while not smoking has a negative effect. Variables such as Hemoglobin and Leukocytes have effects consistent with literature on the topic. See Naqvi et al. (2017) and Blé et al. (2001).

## Circumference

Table 3.6 contains a summary of the posterior densities for the components of  $\beta^{(4)}$ .

Parameter	Mean	SD	Q <sub>0.025</sub>	Q <sub>0.5</sub>	Q <sub>0.975</sub>
$\beta_{Erythrocytes}$	$2.1 \cdot 10^{-1}$	$6.2 \cdot 10^{-2}$	$9.0 \cdot 10^{-2}$	$2.1 \cdot 10^{-1}$	$3.3 \cdot 10^{-1}$
$\beta_{Mean\_Volume}$	$1.2 \cdot 10^{-1}$	$3.7 \cdot 10^{-2}$	$4.5 \cdot 10^{-2}$	$1.2 \cdot 10^{-1}$	$1.9 \cdot 10^{-1}$
$\beta_{Age}$	$3.3 \cdot 10^{-2}$	$1.2 \cdot 10^{-3}$	$3.1 \cdot 10^{-2}$	$3.3 \cdot 10^{-2}$	$3.6 \cdot 10^{-2}$
$\beta_{BMI}$	$1.6 \cdot 10^{-2}$	$7.3 \cdot 10^{-4}$	$1.4 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$	$1.7 \cdot 10^{-2}$
$\beta_{Activity\_Sedentary\_lifestyle}$	$1.4 \cdot 10^{-2}$	$1.4 \cdot 10^{-3}$	$1.1 \cdot 10^{-2}$	$1.3 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$
$\beta_{Gender2:B2\_globulins}$	$1.0 \cdot 10^{-2}$	$1.7 \cdot 10^{-3}$	$6.9 \cdot 10^{-3}$	$1.0 \cdot 10^{-2}$	$1.4 \cdot 10^{-2}$
$\beta_{Activity\_Moderate\_lifestyle}$	$7.1 \cdot 10^{-3}$	$1.2 \cdot 10^{-3}$	$4.9 \cdot 10^{-3}$	$7.1 \cdot 10^{-3}$	$9.4 \cdot 10^{-3}$
$\beta_{Gender2:BMI}$	$6.7 \cdot 10^{-3}$	$1.5 \cdot 10^{-3}$	$3.7 \cdot 10^{-3}$	$6.6 \cdot 10^{-3}$	$9.7 \cdot 10^{-3}$
$\beta_{ALT}$	$5.2 \cdot 10^{-3}$	$5.4 \cdot 10^{-4}$	$4.1 \cdot 10^{-3}$	$5.2 \cdot 10^{-3}$	$6.2 \cdot 10^{-3}$
$\beta_{Leukocytes}$	$4.3 \cdot 10^{-3}$	$5.6 \cdot 10^{-4}$	$3.2 \cdot 10^{-3}$	$4.3 \cdot 10^{-3}$	$5.4 \cdot 10^{-3}$
$\beta_{B2\_globulins}$	$2.8 \cdot 10^{-3}$	$6.5 \cdot 10^{-4}$	$1.5 \cdot 10^{-3}$	$2.8 \cdot 10^{-3}$	$4.0 \cdot 10^{-3}$
$\beta_{B1\_globulins}$	$2.5 \cdot 10^{-3}$	$6.4 \cdot 10^{-4}$	$1.2 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	$3.7 \cdot 10^{-3}$
$\beta_{Age:BMI}$	$2.2 \cdot 10^{-3}$	$6.1 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$	$2.2 \cdot 10^{-3}$	$3.4 \cdot 10^{-3}$
$\beta_{Ferritin}$	$1.3 \cdot 10^{-3}$	$5.1 \cdot 10^{-4}$	$3.1 \cdot 10^{-4}$	$1.3 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$
$\beta_{Creatine}$	$9.1 \cdot 10^{-4}$	$6.2 \cdot 10^{-4}$	$-3.3 \cdot 10^{-4}$	$9.0 \cdot 10^{-4}$	$2.1 \cdot 10^{-3}$
$\beta_{Heart\_rate}$	$5.9 \cdot 10^{-4}$	$3.9 \cdot 10^{-4}$	$-1.9 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$	$1.4 \cdot 10^{-3}$
$\beta_{Age:ALT}$	$5.5 \cdot 10^{-4}$	$4.4 \cdot 10^{-4}$	$-3.2 \cdot 10^{-4}$	$5.4 \cdot 10^{-4}$	$1.4 \cdot 10^{-3}$
$\beta_{Age:Leukocytes}$	$4.5 \cdot 10^{-4}$	$4.6 \cdot 10^{-4}$	$-4.3 \cdot 10^{-4}$	$4.7 \cdot 10^{-4}$	$1.4 \cdot 10^{-3}$
$\beta_{Total\_iron}$	$-4.3 \cdot 10^{-4}$	$3.9 \cdot 10^{-4}$	$-1.2 \cdot 10^{-3}$	$-4.4 \cdot 10^{-4}$	$3.3 \cdot 10^{-4}$
$\beta_{Gender2:Heart\_rate}$	$-4.6 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$	$-2.5 \cdot 10^{-3}$	$-4.6 \cdot 10^{-4}$	$-1.6 \cdot 10^{-3}$
$\beta_{Gender2:Mean\_Volume}$	$-2.2 \cdot 10^{-3}$	$1.9 \cdot 10^{-3}$	$-5.9 \cdot 10^{-3}$	$-2.2 \cdot 10^{-3}$	$1.5 \cdot 10^{-3}$
$\beta_{Age:Albumin}$	$-3.1 \cdot 10^{-3}$	$4.1 \cdot 10^{-4}$	$-4.0 \cdot 10^{-3}$	$-3.1 \cdot 10^{-3}$	$-2.3 \cdot 10^{-3}$
$\beta_{Albumin}$	$-4.3 \cdot 10^{-3}$	$5.9 \cdot 10^{-4}$	$-5.4 \cdot 10^{-3}$	$-4.3 \cdot 10^{-3}$	$-3.1 \cdot 10^{-3}$
$\beta_{Gender2:Age}$	$-1.4 \cdot 10^{-2}$	$2.5 \cdot 10^{-3}$	$-1.9 \cdot 10^{-2}$	$-1.4 \cdot 10^{-2}$	$-9.5 \cdot 10^{-3}$
$\beta_{Gender2}$	$-8.2 \cdot 10^{-2}$	$4.0 \cdot 10^{-3}$	$-9.0 \cdot 10^{-2}$	$-8.2 \cdot 10^{-2}$	$-7.5 \cdot 10^{-2}$
$\beta_{Hematocrit}$	$-1.7 \cdot 10^{-1}$	$5.1 \cdot 10^{-2}$	$-2.7 \cdot 10^{-1}$	$-1.7 \cdot 10^{-1}$	$-7.0 \cdot 10^{-2}$

Table 3.6: Posterior summaries of the  $\beta$ s parameters for the log Circumference

Figure 3.4 shows  $\beta^{(4)}$  credible intervals.

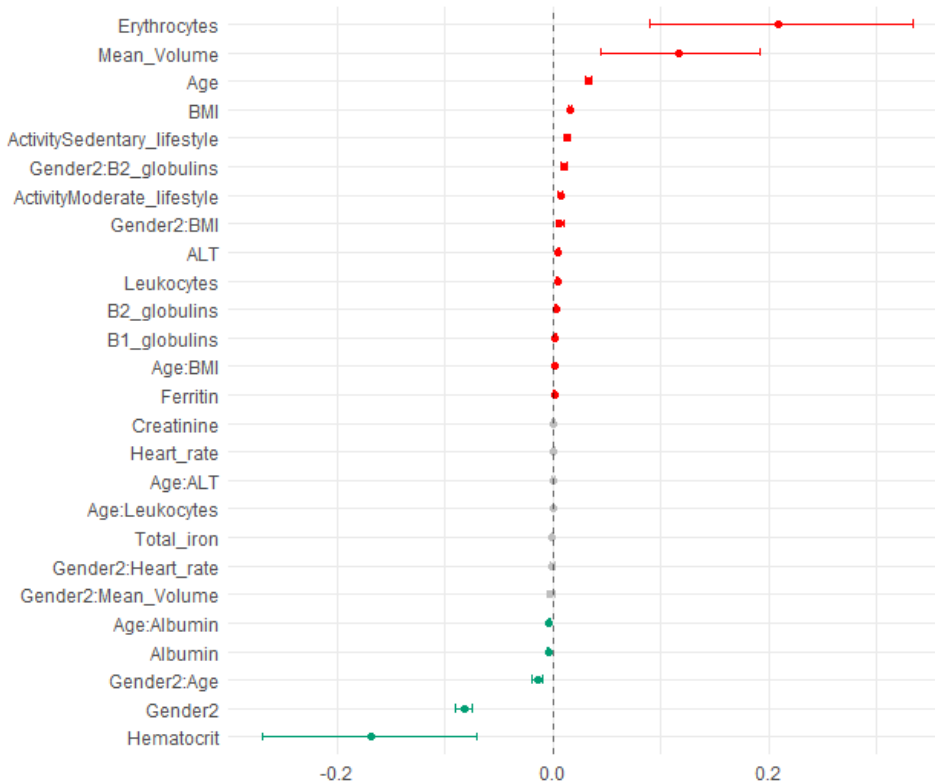


Figure 3.4:  $\beta^{(4)}$ s posterior credible intervals, under *Model 2*

Figure 3.4 shows that there is an increase in the logarithmic waist circumference measurements as the age increases. Variables related to the donor lifestyle influence the waist circumference as expected, indeed both BMI and a sedentary lifestyle have a positive effect. Women have a smaller waist circumference than men, this is also taken into consideration when diagnosing metabolic syndrome as men and women have two different threshold values. Variables such as Hematocrit, Erythrocytes, and Leukocytes have effects consistent with literature about the topic. See Vuong et al. (2014).

## PMAX

In Tables 3.7 and 3.8 a summary of the posterior densities for the components of  $\beta^{(5)}$  are shown.

Parameter	Mean	SD	Q0.025	Q0.5	Q0.975
$\beta_{Age}$	$2.9 \cdot 10^{-2}$	$1.1 \cdot 10^{-3}$	$2.6 \cdot 10^{-2}$	$2.9 \cdot 10^{-2}$	$3.1 \cdot 10^{-2}$
$\beta_{Age^2}$	$1.6 \cdot 10^{-2}$	$7.2 \cdot 10^{-4}$	$1.5 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$	$1.7 \cdot 10^{-2}$
$\beta_{BMI}$	$7.0 \cdot 10^{-3}$	$5.9 \cdot 10^{-4}$	$5.8 \cdot 10^{-3}$	$7.0 \cdot 10^{-3}$	$8.1 \cdot 10^{-3}$
$\beta_{Gender2:B2\_globulins}$	$3.1 \cdot 10^{-3}$	$1.6 \cdot 10^{-3}$	$-6.7 \cdot 10^{-5}$	$3.1 \cdot 10^{-3}$	$6.3 \cdot 10^{-3}$
$\beta_{Age^3}$	$3.0 \cdot 10^{-3}$	$2.7 \cdot 10^{-4}$	$2.4 \cdot 10^{-3}$	$3.0 \cdot 10^{-3}$	$3.5 \cdot 10^{-3}$
$\beta_{ALT}$	$2.2 \cdot 10^{-3}$	$4.8 \cdot 10^{-4}$	$1.2 \cdot 10^{-3}$	$2.2 \cdot 10^{-3}$	$3.1 \cdot 10^{-3}$
$\beta_{Age: BMI}$	$2.0 \cdot 10^{-3}$	$4.5 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$	$2.0 \cdot 10^{-3}$	$2.9 \cdot 10^{-3}$
$\beta_{B2\_globulins}$	$1.9 \cdot 10^{-3}$	$7.4 \cdot 10^{-4}$	$4.9 \cdot 10^{-4}$	$1.9 \cdot 10^{-3}$	$3.4 \cdot 10^{-3}$
$\beta_{Total\_proteins}$	$1.4 \cdot 10^{-3}$	$4.7 \cdot 10^{-4}$	$5.2 \cdot 10^{-4}$	$1.4 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$
$\beta_{ABO\_AB}$	$1.3 \cdot 10^{-3}$	$4.0 \cdot 10^{-3}$	$-6.6 \cdot 10^{-3}$	$1.3 \cdot 10^{-3}$	$9.0 \cdot 10^{-3}$
$\beta_{A1\_globulins}$	$1.0 \cdot 10^{-3}$	$5.8 \cdot 10^{-4}$	$-1.3 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$	$2.1 \cdot 10^{-3}$
$\beta_{Total\_cholesterol}$	$9.9 \cdot 10^{-4}$	$5.5 \cdot 10^{-4}$	$-9.1 \cdot 10^{-5}$	$9.8 \cdot 10^{-4}$	$2.1 \cdot 10^{-3}$
$\beta_{Heart\_rate}$	$9.1 \cdot 10^{-4}$	$3.3 \cdot 10^{-4}$	$2.5 \cdot 10^{-4}$	$9.1 \cdot 10^{-4}$	$1.6 \cdot 10^{-3}$
$\beta_{Age:B2\_globulins}$	$7.2 \cdot 10^{-4}$	$5.7 \cdot 10^{-4}$	$-3.8 \cdot 10^{-4}$	$7.2 \cdot 10^{-4}$	$1.8 \cdot 10^{-3}$
$\beta_{Age:Alcohol}$	$7.2 \cdot 10^{-4}$	$9.1 \cdot 10^{-4}$	$-1.1 \cdot 10^{-3}$	$7.1 \cdot 10^{-4}$	$2.5 \cdot 10^{-3}$
$\beta_{Ferritin}$	$6.6 \cdot 10^{-4}$	$4.7 \cdot 10^{-4}$	$-2.3 \cdot 10^{-4}$	$6.5 \cdot 10^{-4}$	$1.6 \cdot 10^{-3}$
$\beta_{Age:B1\_globulins}$	$5.2 \cdot 10^{-4}$	$4.8 \cdot 10^{-4}$	$-4.0 \cdot 10^{-4}$	$5.2 \cdot 10^{-4}$	$1.5 \cdot 10^{-3}$
$\beta_{B\_globulins}$	$3.4 \cdot 10^{-4}$	$5.5 \cdot 10^{-4}$	$-7.4 \cdot 10^{-4}$	$3.5 \cdot 10^{-4}$	$1.4 \cdot 10^{-3}$
$\beta_{Age:ALT}$	$2.6 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$-5.1 \cdot 10^{-4}$	$2.6 \cdot 10^{-4}$	$1.0 \cdot 10^{-3}$
$\beta_{Gender2:Lymphocytes\_perc}$	$1.9 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$	$-1.9 \cdot 10^{-3}$	$1.9 \cdot 10^{-4}$	$2.4 \cdot 10^{-3}$
$\beta_{Age:Monocytes\_perc}$	$1.8 \cdot 10^{-4}$	$3.7 \cdot 10^{-4}$	$-5.3 \cdot 10^{-4}$	$1.8 \cdot 10^{-4}$	$9.1 \cdot 10^{-4}$
$\beta_{Creatine}$	$-1.9 \cdot 10^{-4}$	$5.4 \cdot 10^{-4}$	$-1.3 \cdot 10^{-3}$	$-1.8 \cdot 10^{-4}$	$8.6 \cdot 10^{-4}$
$\beta_{Age:Albumin}$	$-2.0 \cdot 10^{-4}$	$4.5 \cdot 10^{-4}$	$-1.1 \cdot 10^{-3}$	$-2.0 \cdot 10^{-4}$	$6.8 \cdot 10^{-4}$
$\beta_{Age:Lymphocytes\_perc}$	$-2.7 \cdot 10^{-4}$	$3.8 \cdot 10^{-4}$	$-9.9 \cdot 10^{-4}$	$-2.7 \cdot 10^{-4}$	$4.8 \cdot 10^{-4}$
$\beta_{Age:Volume\_distribution}$	$-2.8 \cdot 10^{-4}$	$4.4 \cdot 10^{-4}$	$-1.1 \cdot 10^{-3}$	$-2.8 \cdot 10^{-4}$	$-6.0 \cdot 10^{-4}$
$\beta_{ABO\_A}$	$-4.1 \cdot 10^{-4}$	$1.7 \cdot 10^{-3}$	$-3.7 \cdot 10^{-3}$	$-4.1 \cdot 10^{-4}$	$2.8 \cdot 10^{-3}$

Table 3.7: Posterior summaries of the  $\beta$ s parameters for the log PMAX

Parameter	Mean	SD	Q0.025	Q0.5	Q0.975
$\beta_{Volume\_distribution}$	$-4.4 \cdot 10^{-4}$	$5.2 \cdot 10^{-4}$	$-1.4 \cdot 10^{-3}$	$-4.5 \cdot 10^{-4}$	$6.0 \cdot 10^{-4}$
$\beta_{Age:Mean\_Volume}$	$-7.4 \cdot 10^{-4}$	$5.3 \cdot 10^{-4}$	$-1.8 \cdot 10^{-3}$	$-7.4 \cdot 10^{-4}$	$3.0 \cdot 10^{-4}$
$\beta_{Age:G\_globulins}$	$-8.0 \cdot 10^{-4}$	$5.6 \cdot 10^{-4}$	$-1.9 \cdot 10^{-3}$	$-8.0 \cdot 10^{-4}$	$2.7 \cdot 10^{-4}$
$\beta_{G\_globulins}$	$-1.3 \cdot 10^{-3}$	$8.4 \cdot 10^{-4}$	$-2.9 \cdot 10^{-3}$	$-1.3 \cdot 10^{-3}$	$3.8 \cdot 10^{-4}$
$\beta_{Lymphocytes\_perc}$	$-1.5 \cdot 10^{-3}$	$4.9 \cdot 10^{-4}$	$-2.5 \cdot 10^{-3}$	$-1.5 \cdot 10^{-3}$	$-5.6 \cdot 10^{-4}$
$\beta_{Gender2:BMI}$	$-3.1 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$	$-5.2 \cdot 10^{-3}$	$-3.1 \cdot 10^{-3}$	$-9.8 \cdot 10^{-4}$
$\beta_{Gender2:Age}$	$-3.6 \cdot 10^{-3}$	$1.6 \cdot 10^{-3}$	$-6.6 \cdot 10^{-3}$	$-3.6 \cdot 10^{-3}$	$-5.6 \cdot 10^{-4}$
$\beta_{Mean\_Volume}$	$-4.4 \cdot 10^{-3}$	$6.4 \cdot 10^{-4}$	$-5.7 \cdot 10^{-3}$	$-4.4 \cdot 10^{-3}$	$-3.2 \cdot 10^{-3}$
$\beta_{ABO\_B}$	$-4.6 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	$-9.5 \cdot 10^{-3}$	$-4.6 \cdot 10^{-3}$	$4.4 \cdot 10^{-4}$
$\beta_{Gender2:G\_globulins}$	$-5.3 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	$-8.4 \cdot 10^{-3}$	$-5.3 \cdot 10^{-3}$	$-2.0 \cdot 10^{-3}$
$\beta_{Gender2}$	$-3.2 \cdot 10^{-2}$	$2.3 \cdot 10^{-3}$	$-3.6 \cdot 10^{-2}$	$-3.2 \cdot 10^{-2}$	$-2.7 \cdot 10^{-2}$

Table 3.8: Posterior summaries of the  $\beta$ s parameters for the log PMAX (continued)

Figure 3.5 shows  $\beta^{(5)}$  credible intervals of order 95%.



Figure 3.5:  $\beta^{(5)}$ s posterior credible intervals, under *Model 2*

As expected from the literature the systolic pressure increases as the age increases with a relationship that has a linear, quadratic, and cubic component. Variables related to the donor lifestyle such as BMI and heart rate have a positive effect on the systolic pressure as expected. Women present lower systolic pressure than men but this is not taken into consideration in the threshold used to diagnose metabolic syndrome. Variables such as Total proteins and ALT have effects consistent with literature about the topic. See Jia et al. (2021) and Koenig et al. (1991).

In conclusion, the posterior estimates support the findings that, as expected, lifestyle variables heavily influence our target variables. Other highly significant variables are ALT, globulins variables, Hemoglobin and Mean\_Volume.

### 3.3.2. Random effects $b_i$

The  $i$ th-subject specific random effects of the models, represented by the coefficient  $b_i$ , are here analyzed for each target variable model. They represent the intercept of the model and are specific for each donor. We analyze the mean  $\mu_i$  and standard deviation  $\sigma_i$  of the distributions of  $b_i$ s. Table 3.9 contains a summary of the means of  $\mu_i$ s and  $\sigma_i$ s divided by gender.

Target variable	Gender	Mean	Mean standard deviation
Glucose	M	4.513449	0.02874095
	F	4.513481	0.03528949
HDL_cholesterol	M	3.985294	0.05085526
	F	3.974896	0.05328285
Triglycerides	M	1.493982	0.02703157
	F	1.483368	0.03213596
Circumference	M	4.52047	0.02052385
	F	4.520233	0.02602579
PMAX	M	4.780523	0.01814564
	F	4.780142	0.01884282

Table 3.9: Summary of the means of  $b_i$ 's parameters distribution

We can notice that the behavior of the means of the distributions do not seem to vary between genders. The mean standard deviation, on the other hand, seems to differ between

genders slightly. From the following Figures 3.6-3.15, we can see that the distributions of  $b_i$  present a higher degree of connection with the age of the donor but it is still not enough to identify specific clusters. This suggests that a clustering based on gender would not help the prediction process. A sample of 50 donors, 25 male and 25 female, was selected randomly from the population to facilitate the representation of the coefficients. The posterior distributions of the coefficients have been analyzed to find if a grouping by gender or age is possible.

## Glucose

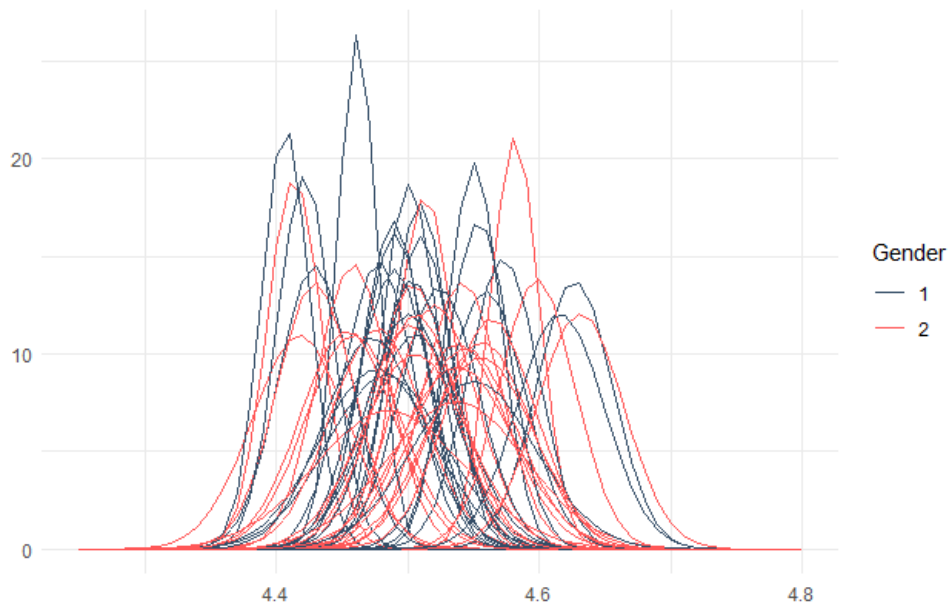


Figure 3.6: Posterior densities of  $b_i^1$  under *Model 2*, blu curves represent males and red represent female

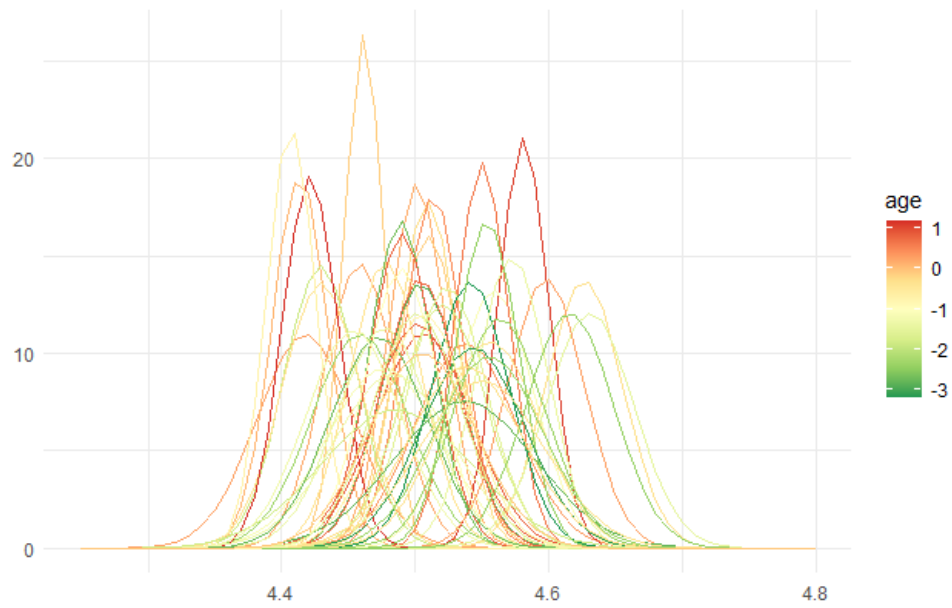


Figure 3.7: Posterior densities of  $b_i^1$  under *Model 2*, colored by age of the donor

### HDL\_cholesterol

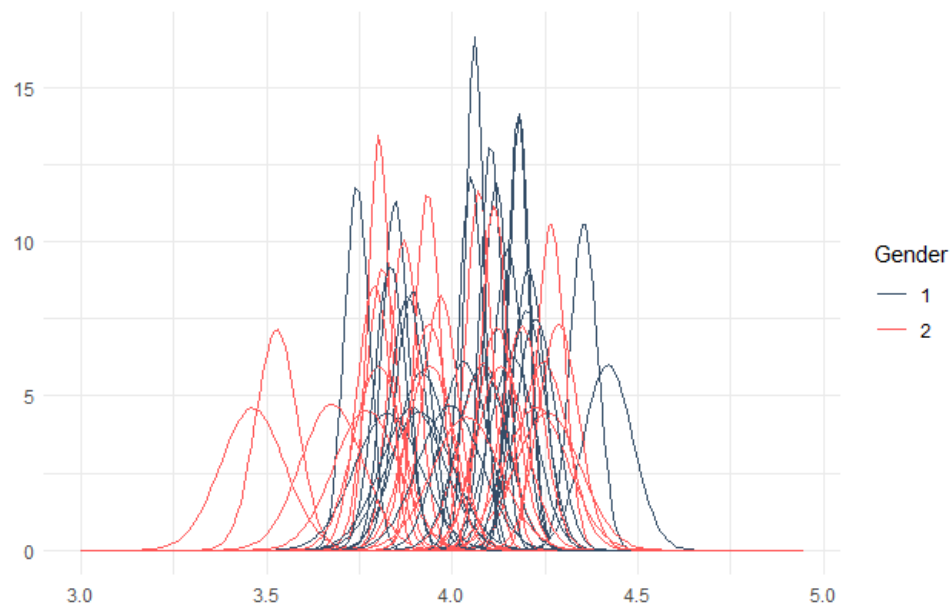


Figure 3.8: Posterior densities of  $b_i^2$  under *Model 2*, blu curves represent males and red represent female

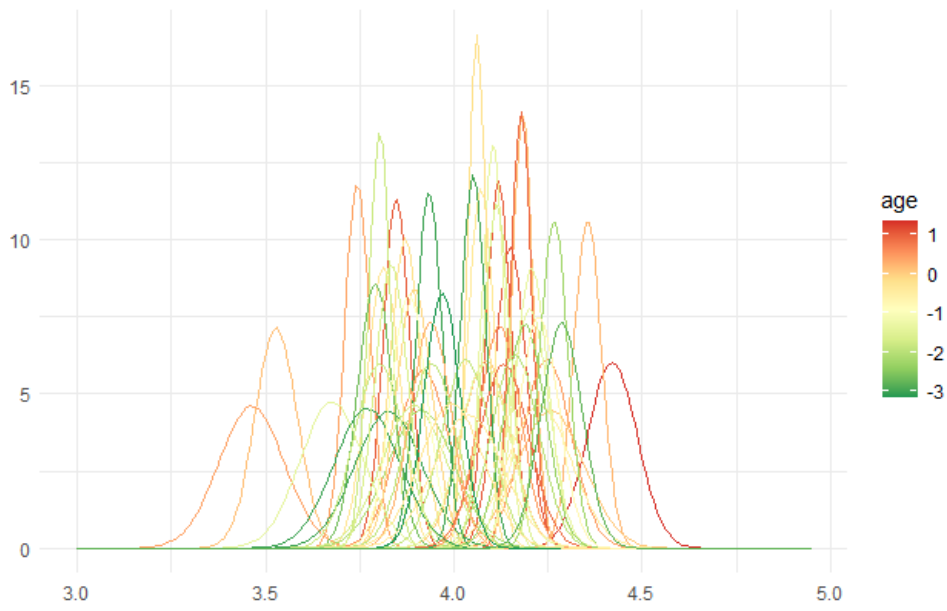


Figure 3.9: Posterior densities of  $b_i^2$  under *Model 2* colored by age of the donor

## Triglycerdides

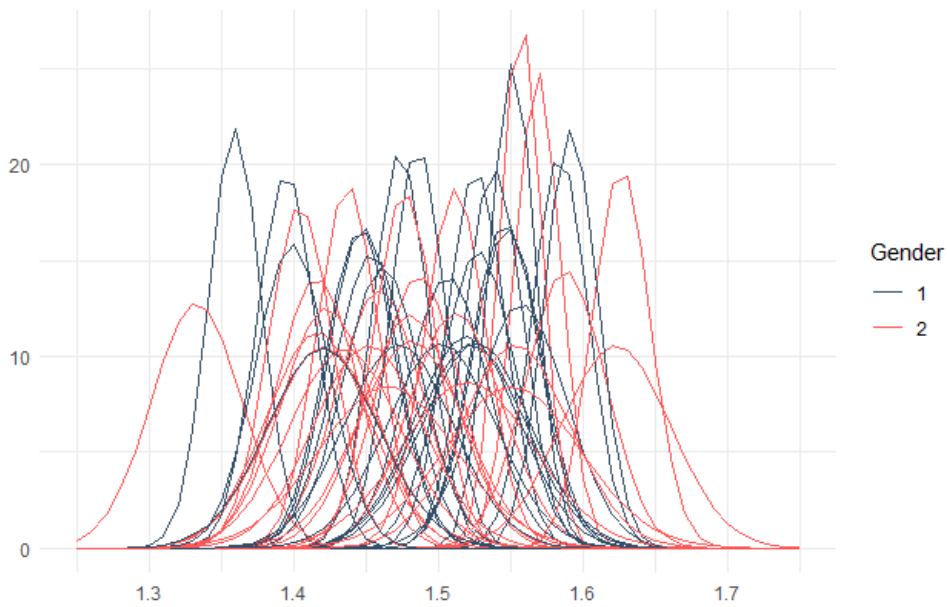


Figure 3.10: Posterior densities of  $b_i^3$  under *Model 2*, blu curves represent males and red represent female

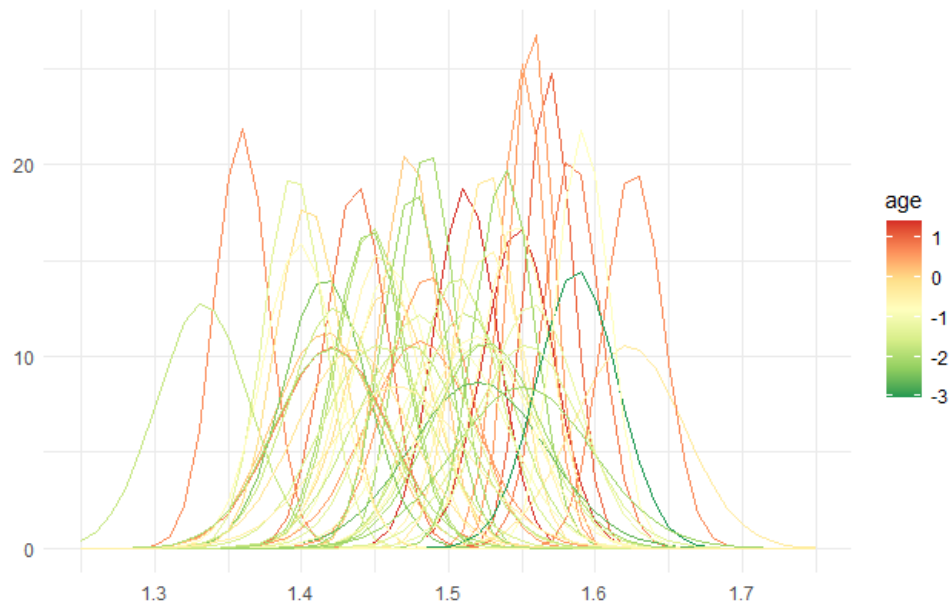


Figure 3.11: Posterior densities of  $b_i^3$  under *Model 2*, colored by age of the donor

### Circumference

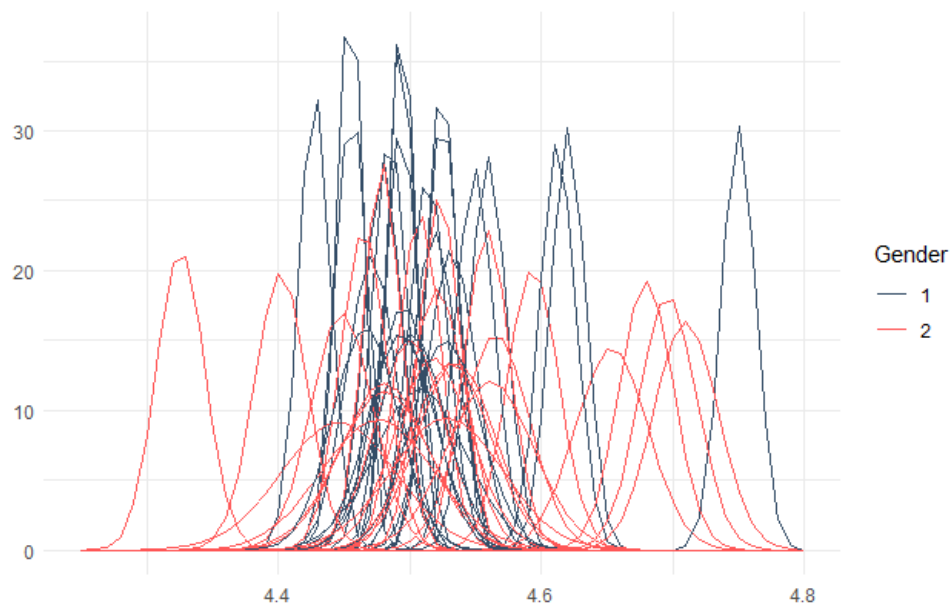


Figure 3.12: Posterior densities of  $b_i^4$  under *Model 2*, blu curves represent males and red represent female

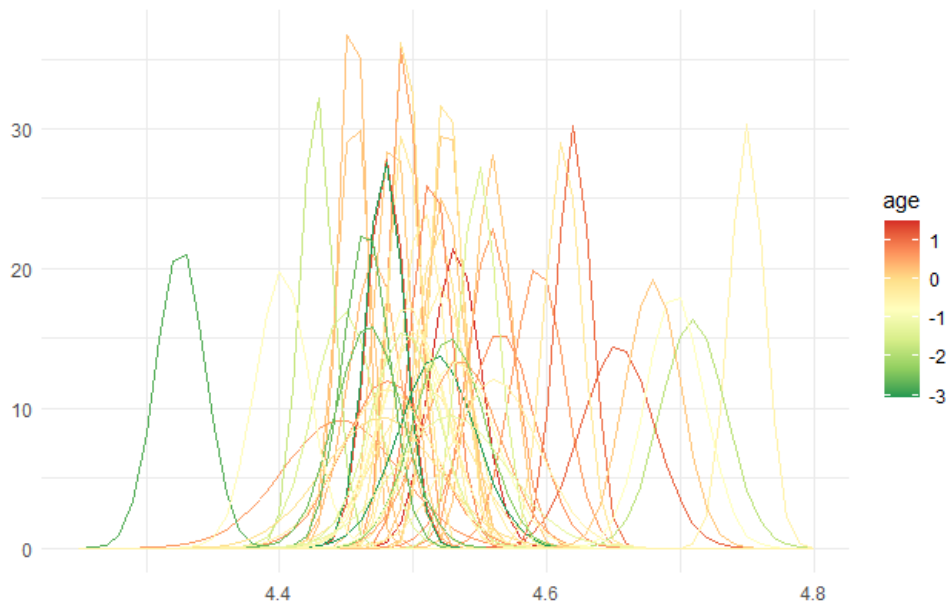


Figure 3.13: Posterior densities of  $b_i^4$  under *Model 2*, colored by age of the donor

## PMAX

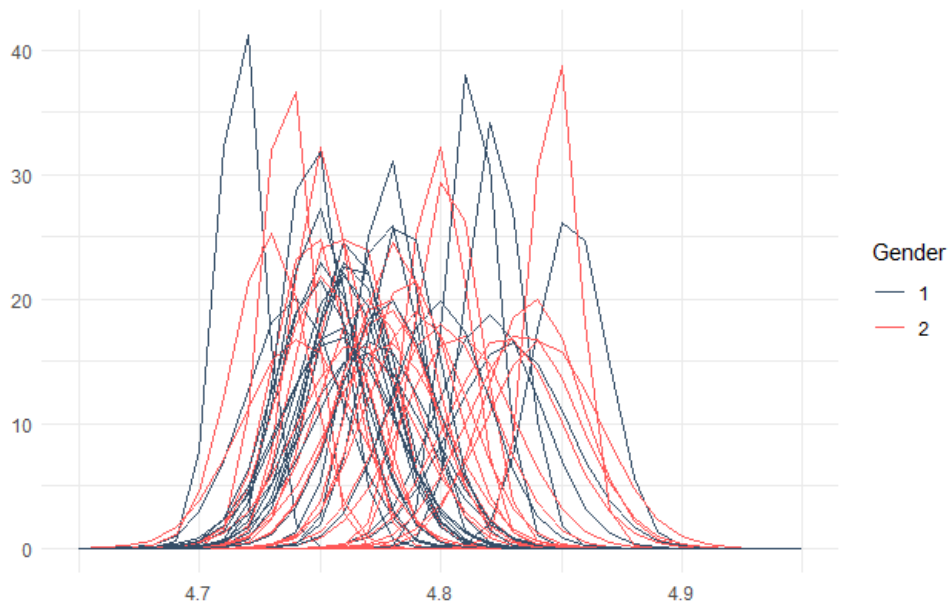


Figure 3.14: Posterior densities of  $b_i^5$  under *Model 2*, blu curves represent males and red represent female

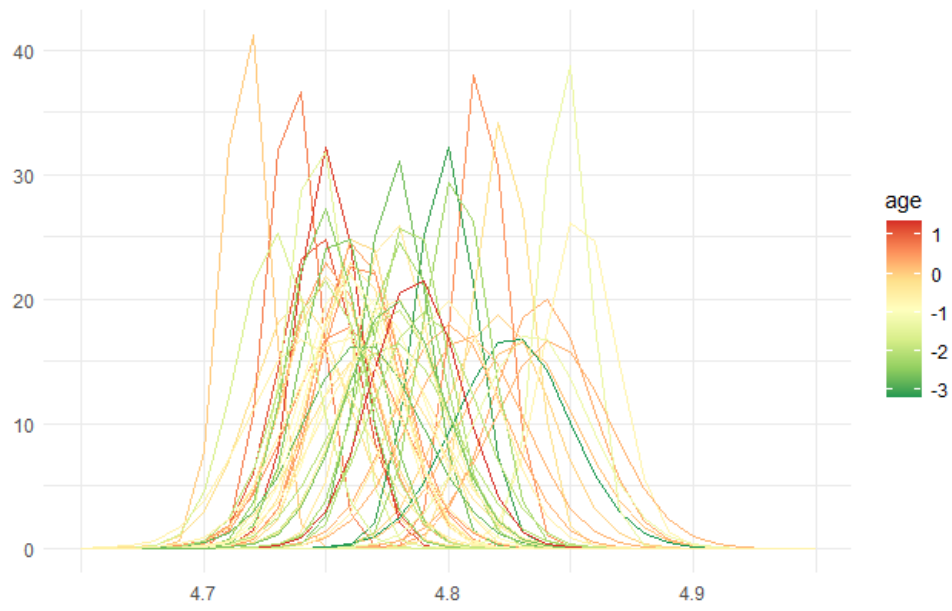


Figure 3.15: Posterior densities of  $b_i^5$  under *Model 2*, colored by age of the donor



# 4 | Predictions

In this chapter, we focus on the results of the two-stage plug-in predictive model and provide an example of the tool given to AVIS. The Test set used for the models consists of the most recent observations of all donors in Dataset 2, as described in Section 1.4.3. The prediction was carried out using two models detailed in Section 2.3. The first is a logistic regression model, while the second uses combinatorial formulas to calculate the exact probability of developing metabolic syndrome from the predicted values of the target variables.

## 4.1. Prediction

### 4.1.1. Dataset composition

To predict the probability of the onset of metabolic syndrome, we need to use *Model 2* on the Test set to forecast the next value of each target variable. As explained in Section 2.3, the result of this step will be a distribution. We can create a matrix  $W$  consisting of as many rows as the donors in the Test set, with a column for each target variable. Each element  $w_i^k$  of this matrix is the probability that the  $k$ -th target variable of the  $i$ -th individual is outside the normal thresholds used to diagnose the metabolic syndrome. For each of these rows, we can associate the true presence of the disease with a binary indicator where value 1 means that donor  $i$  has metabolic syndrome and 0 that the donor  $i$  is healthy.

Around 55% of this dataset is used to train the logistic regression shown in Section 2.3.1 and 45% to test it. The combinatorial model in Section 2.3.2, on the other hand, can be applied to the entire dataset. However, to facilitate a comparison between the two models, it will also be applied to the same test set as the logistic regression.

### 4.1.2. Comparison of the predictions

To make an accurate comparison of the two models, we first analyze the distribution of the total probability of the disease  $p_i(\mathbf{w}_i)$ , grouping the results by the true presence of

the disease. The results are shown in Figure 4.1.

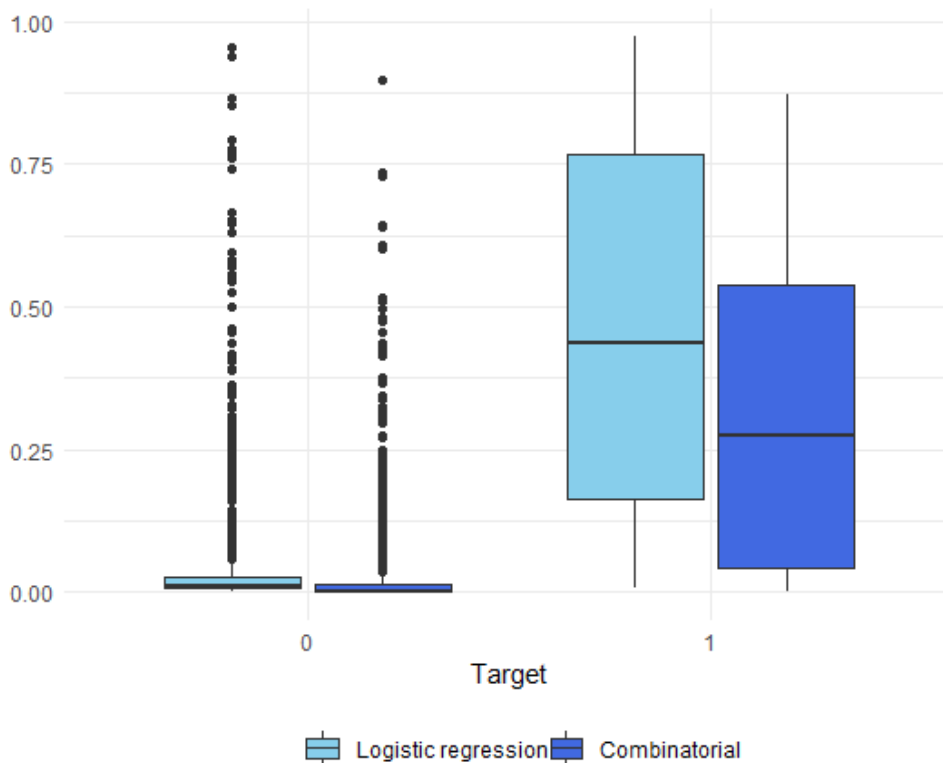


Figure 4.1: Comparison of the distribution of the total probability  $p_i(\mathbf{w}_i)$  between the logistic regression model and the combinatorial model under the target

We observe that the probabilities obtained from both methods are quite low, even for the positive group. This will influence the selection of the threshold used to classify donors as having metabolic syndrome, thus assigning them a value of 1.

The  $p_i(\mathbf{w}_i)$  obtained with the combinatorial model has a lower mean compared to that from the logistic regression, suggesting a higher number of false negatives for this method. Additionally, we know that male and female donors exhibit different donation rates due to AVIS rules. For this reason, we hypothesize that the accuracy for men will be higher than that for women. Figure 4.2 illustrates their difference in the total probability distribution.

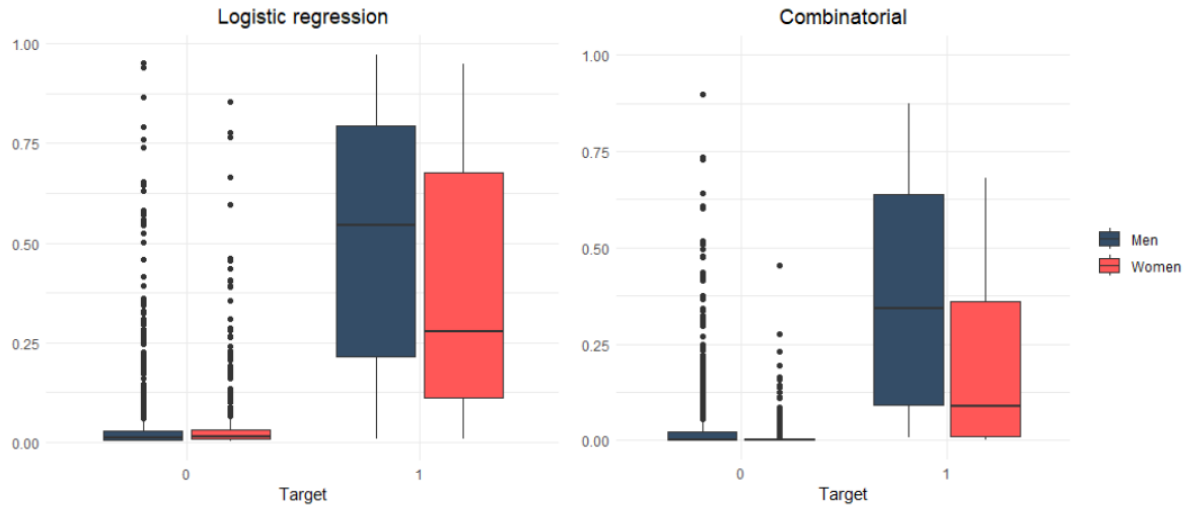


Figure 4.2: Comparison of the distribution of the total probability between the male and female population for the logistic regression results and the combinatorial method result

As expected, the posterior mean of  $p_i(\mathbf{w}_i)$  for women is lower than for men, making it more challenging to accurately classify female donors. This is especially evident with the combinatorial model.

In order to compute the parameters used to assess the goodness of the classification, we need to determine a threshold for the value of  $p_i(\mathbf{w}_i)$ . When  $p_i(\mathbf{w}_i)$  is above this threshold, donor  $i$  will be classified as having the disease. On the other hand when  $p_i(\mathbf{w}_i)$  is below the threshold the donor will be considered healthy. From Figure 4.1, we understand that the optimal threshold will be very low, so we consider 10 threshold values between 0 and 0.2 to compare the resulting parameters. Given that our dataset is highly skewed (with less than 7% of the donor population having metabolic syndrome), the classification parameters that will be most important in choosing a threshold are the F1-score and the sensitivity. However, the sensitivity plays a more critical role than the F1-score in our classification because having a high number of false negatives is more detrimental than having a high number of false positives. This is because the advice given to a donor predicted as positive is aimed at improving lifestyle choices and will not harm a healthy individual.

Figure 4.3 reports the values of the classification tools (detailed in Section 2.3.3) for the logistic regression method, while Figure 4.4, reports the values of the classification tools for the combinatorial method.

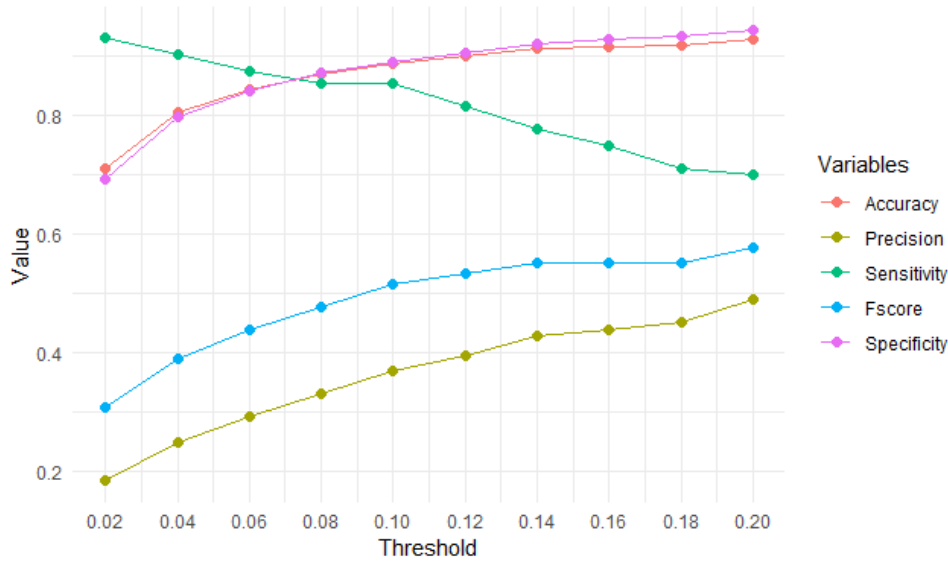


Figure 4.3: Classification tools for different threshold values for the logistic regression method

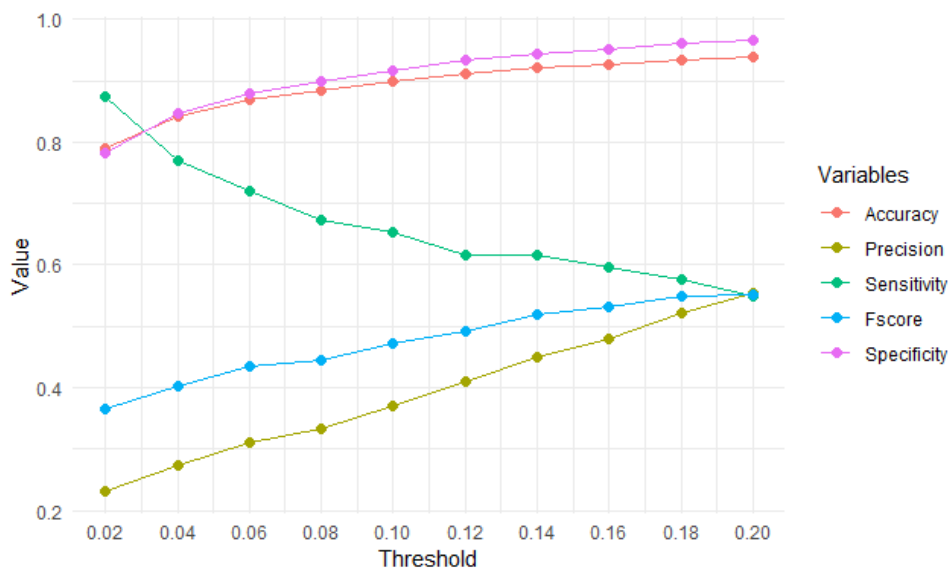


Figure 4.4: Classification parameters for different threshold values for the combinatorial method

As the threshold increases, all measures except the sensitivity increase as the threshold increases. The sensitivity is highest at lower threshold values as almost all the donors with metabolic syndrome are correctly classified, despite a high number of false positive predictions.

As expected the logistic model generally presents higher values for both the sensitivity

and the F1-score. Therefore the logistic method is the best choice between the two and will be used in the final tool for AVIS.

### 4.1.3. Threshold choice

As mentioned earlier we need to use a threshold value to classify our donors properly. Our objective is not only to perform a binary classification to predict whether donors will develop metabolic syndrome but also to identify risk zones to better prevent the onset of the disease. We will classify these zones, where  $p_i(\mathbf{w}_i)$  can fall, as follows:

- "Green" if there is little to no risk of developing metabolic syndrome;
- "Yellow" if there is a moderate risk of developing metabolic syndrome;
- "Red" if there is a high risk of developing metabolic syndrome.

To define these zones, we need to identify two thresholds. In Figure 4.5 a zoomed version of Figure 4.1 is shown where the focus is on the overlap of the distributions of  $p_i(\mathbf{w}_i)$  between healthy donors and those with metabolic syndrome. The proposed thresholds are marked in red at  $p_i(\mathbf{w}_i) = 0.04$  and  $p_i(\mathbf{w}_i) = 0.16$ . Choosing these values for the thresholds ensures that the main body of the boxplot of  $p_i(\mathbf{w}_i)$  for healthy donors falls within the "Green" zone and the main body of the boxplot of  $p_i(\mathbf{w}_i)$  for donors with metabolic syndrome falls within the "Red" zone. The remaining values of  $p_i(\mathbf{w}_i)$ , which could belong to either healthy donors or those with metabolic syndrome, fall within the "Yellow" zone.

The classification parameters for these values are reported in Table 4.1, and the corresponding confusion matrixes can be found in Figure 4.6.

The zones are defined as:

- "Green":  $0 < p_i(\mathbf{w}_i) < 0.04$ ;
- "Yellow" :  $0.04 < p_i(\mathbf{w}_i) < 0.16$ ;
- "Red":  $0.16 < p_i(\mathbf{w}_i) < 1$ .

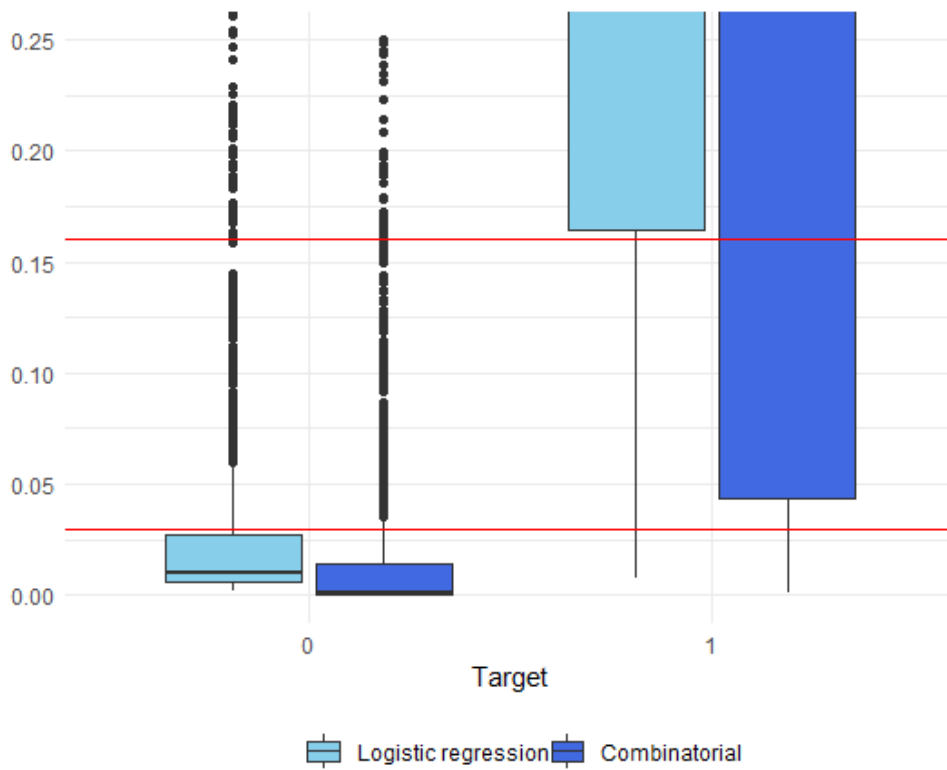


Figure 4.5: Zoomed portion of Figure 4.1, the red lines at  $p_i(\mathbf{w}_i) = 0.04$  and  $p_i(\mathbf{w}_i) = 0.16$  represent the proposed thresholds

	<b>0.04</b>	<b>0.16</b>
<b>Accuracy</b>	0.805	0.916
<b>Precision</b>	0.250	0.438
<b>Sensitivity</b>	0.904	0.750
<b>F1-score</b>	0.392	0.553
<b>Specificity</b>	0.798	0.928

Table 4.1: Classification parameters for thresholds 0.04 and 0.16

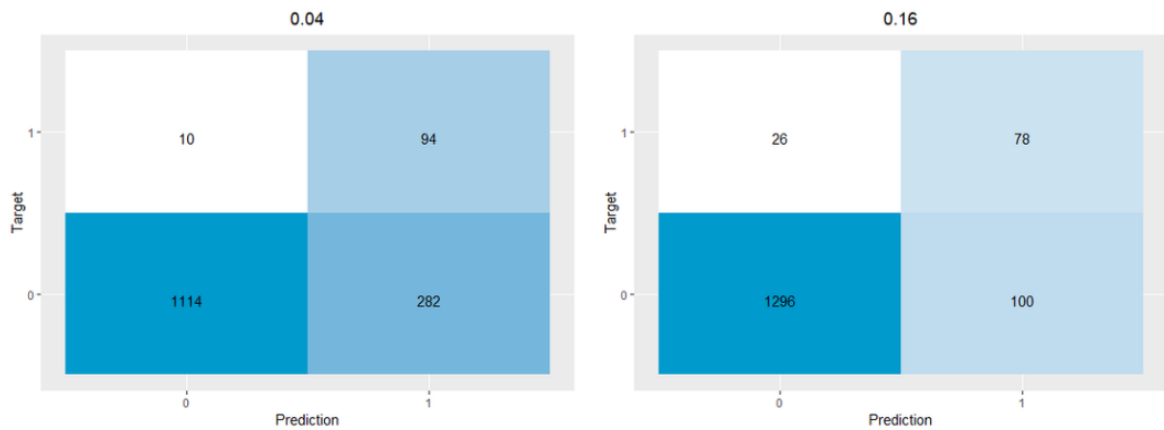


Figure 4.6: Confusion matrixes for threshold 0.04 (left) and 0.16 (right)

In Table 4.7 a summary of the classification is shown. The column named 0 represents healthy individuals while the column named 1 represents donors with metabolic syndrome. The percentage frequencies are obtained using the total number of healthy individuals for column 0 and sick for column 1.

	0	1
<b>Green</b>	1114	10
<b>Yellow</b>	182	16
<b>Red</b>	100	78

(a) Classification summary with absolute frequencies

	0	1
<b>Green</b>	79.8%	9.6%
<b>Yellow</b>	13%	15.4%
<b>Red</b>	7.2%	75%

(b) Classification summary with percentage frequencies

Figure 4.7: Classification summaries

We observe that 80% of healthy individuals are accurately classified as having no risk of developing metabolic syndrome, and 75% of individuals with the condition are correctly identified. Including the "Yellow" zone, the correct identification of future metabolic syndrome cases increases to 90%. Tables 4.8 and 4.9 provide a gender-specific breakdown of these results. Based on the findings in Figure 4.2, we hypothesized that the overall accuracy would be higher for men compared to women. These tables offer a quantitative assessment of this difference: 91% of true positives are identified in males, whereas this number drops to 65

	0	1		0	1
<b>Green</b>	818	6	<b>Green</b>	337	12
<b>Yellow</b>	141	8	<b>Yellow</b>	24	8
<b>Red</b>	66	56	<b>Red</b>	10	14

(a) Classification summary with absolute frequencies for male donors

(b) Classification summary with absolute frequencies for female donors

Figure 4.8: Classification summary with absolute frequencies divided by gender

	0	1		0	1
<b>Green</b>	79.8%	8.6%	<b>Green</b>	90.8%	35.3%
<b>Yellow</b>	13.8%	11.4%	<b>Yellow</b>	6.5%	23.5%
<b>Red</b>	6.4%	80%	<b>Red</b>	2.7%	41.2%

(a) Classification summary with percentage frequencies for male donors

(b) Classification summary with percentage frequencies for female donors

Figure 4.9: Classification summary with percentage frequencies divided by gender

In conclusion, this stoplight system proves particularly valuable for doctors conducting an initial screening, enabling them to more easily identify at-risk donors. By focusing their attention on the 25% of the population classified as "Yellow" and "Red," or especially on the 12% classified as "Red," doctors can efficiently prioritize their time. This approach increases the likelihood of visiting at-risk donors from 7% when sampling randomly from the population, to 25% when focusing on the "Yellow" and "Red" zones, and up to 44% when focusing only on the "Red" zone. The decision to include only the "Red" zone or to extend to the "Yellow" zone should consider available time for thorough visits and the proportion of at-risk donors identified. Specifically, visiting only donors from the "Red" zone ensures 75% of at-risk donors are seen. However, expanding visits to include the "Yellow" zone allows for 90% coverage of at-risk donors.

## 4.2. Tool provided to AVIS

This section introduces the tool developed for AVIS, designed to aid doctors during clinical visits by assessing the risk of donors developing metabolic syndrome. Initially conceived to facilitate easy identification of high-risk donors using the stoplight classification system discussed earlier, the tool has evolved to include a secondary feature. This feature evaluates whether the risk classification would change if the donor were to lose or gain 10 kilograms in weight. Weight was chosen due to its significant correlation with all target variables and its practicality for donors to monitor, without requiring blood tests.

The tool provided consists of an R batch program that takes the most recent measurements of all covariates as input, and returns the predicted risk zone for the donor. The interface for the tool is an Excel file where the covariate values for a donor can be entered, as shown in Figure 4.10. After delivering the tool to AVIS, the batch program will be integrated with the existing systems so that the data can be automatically retrieved from the databases and seamlessly incorporated into the existing workflows.

	A	B	C
1	<b>Insert your last blood test measurement to find your risk for metabolic syndrome:</b>		
2			
3	<b>Exam:</b>	<b>Value:</b>	<b>Explanation:</b>
4	Gender	2	Write 1 if you are male and 2 if you are female
5	AB0	A	Valid blood types are: A, B, AB, 0
6	Rh	1	Write 1 if your Rhesus factor is positive and 0 if it is negative
7	Smoke	0	Write 1 if you smoke and 0 if you do not
8	Alcohol	0	Write 1 if you drink habitually and 0 if you do not
9	Activity	2	Write 1 if you have an active lifestyle, 2 if your lifestyle is moderate and 3 if it is sedentary
10	Age	13/08/1999	Write your date of birth, the age will be automatically calculated
11	ALT	12	Value of the Alanine aminotranferase measurement
12	Albumin	4	Value of the Creatine measurement
13	A1_globulins	0,2	Value of the Alpha-1-globulins measurement
14	A2_globulins	0,7	Value of the Alpha-2-globulins measurement
15	Basophils_perc	0,3	Value of the percentage of basophils measurement
16	B1_globulins	0,9	Value of the Beta-1-globulins measurement
17	B2_globulins	1,4	Value of the Beta-2-globulins measurement

Figure 4.10: Interface of the tool prototype for AVIS

The values inputted are accordingly transformed as mentioned in Chapter 1 and are given to the models to predict the risk of metabolic syndrome. The output of the program depends on the Zone the donor falls in. Three different messages, shown in Figure 4.11, are visualized as an output depending on the zone.

The donor's risk of developing metabolic syndrome is

**LOW**

(a) Output for donors classified in the "Green" zone

The donor's risk of developing metabolic syndrome is

**MODERATE**

(b) Output for donors classified in the "Yellow" zone

The donor's risk of developing metabolic syndrome is

**HIGH**

(c) Output for donors classified in the "Red" zone

Figure 4.11: Possible outputs

The tool also automatically performs two simulations to assess the effects of an increase or decrease in the donor's weight of 10 kilograms. If these simulations lead to a change in the donor's risk classification, the tool displays one of the messages shown in Figures 4.12 and 4.13 as an output.

If the donor loses 10 kg, their new risk of developing metabolic syndrome would decrease to

**LOW**

(a)

If the donor loses 10 kg, their new risk of developing metabolic syndrome would decrease to

**MODERATE**

(b)

Figure 4.12: Additional output in the case of decreased weight

If the donor gains 10 kg, their new risk of developing metabolic syndrome would increase to

**MODERATE**

(a)

If the donor gains 10 kg, their new risk of developing metabolic syndrome would increase to

**HIGH**

(b)

Figure 4.13: Additional output in the case of increased weight

## 5 | Conclusions and future developments

Metabolic syndrome, a cluster of conditions that elevate the risk of cardiovascular disease, is an increasingly prevalent problem worldwide. This thesis leverages data from AVIS Milano databases to model the risk of metabolic syndrome among blood donors. Such models are crucial for organizations like AVIS, as they enhance the ability to monitor and intervene in the health of their volunteers. Early identification and prevention of metabolic syndrome can alleviate the burden on the healthcare system, both by reducing the need for extensive medical examinations and by minimizing economic costs.

The initial data from AVIS was cleansed and appropriately pre-processed. A two-stage plug-in model was developed. In the first stage, three different Bayesian mixed-effect longitudinal models were analyzed to explore the relationship between each of the five target variables used to diagnose metabolic syndrome and the covariates used in this study. Autoregressive models have been considered but were discarded early on, the time is considered in the Bayesian models through the variable age and the time-varying covariates. The selection of the best model between the three was done by comparing their *WAIC* and *LOO* values and also considering their computational cost. The selected model incorporates an individual-specific random-effect intercept to account for between-subject variability.

Posterior analysis of this model reveals that lifestyle-related variables, such as BMI, physical activity levels, and smoking habits, were particularly significant, aligning with existing literature. Additionally, significant variables were identified in serum proteins, including albumin and various globulins.

Posterior findings have been plugged in a logistic regression model to predict the overall risk of developing metabolic syndrome. Two thresholds were selected to categorize donors into three distinct risk zones, enabling our model to correctly identify up to 90% of donors with metabolic syndrome, and up to 80% of healthy donors. Additionally, these risk zones allow for an initial screening of donors, enabling doctors to focus on a smaller segment of the population for more in-depth evaluations. From these models, we have developed a

tool capable of taking the latest blood test measurements of a donor and returning the risk of developing metabolic syndrome. This tool will need to be integrated into the AVIS systems to access their databases and provide results to the doctors visiting the donors.

This work sets the stage for several extensions, beginning with broadening the scope of diseases the tool can predict. A potential future direction could involve targeting liver diseases, a frequent cause for donors to discontinue donations. Another potential development is creating a joint probability distribution for the five target variables, allowing for more accurate predictions. One approach could involve sequentially predicting each variable, using each preceding prediction as a covariate for the current one.

Improving the individual models with new and more complex techniques could yield better predictions. For instance, Bayesian models could be used to compute the total probability of metabolic syndrome, or neural networks could be employed for classification. Currently, when using Model 2 to predict the target variables, our predictions are limited to individuals who have donated at least once. Extending the model to predict the risk of developing metabolic syndrome for new donors or individuals external to AVIS is another possible enhancement.

## Bibliography

- Avis. Avis website. <https://www.avis.it/diventa-donatore/>. Accessed: 2024-06-21.
- A. Blé, E. Palmieri, S. Volpato, F. Costantini, R. Fellin, and G. Zuliani. White blood cell count is associated with some features of metabolic syndrome in a very old population. *Nutr Metab Cardiovasc Dis*, (11), 8 2001.
- C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80. PMLR, 16–18 Apr 2009.
- M. B. Fessler, K. Rose, Y. Zhang, R. Jaramillo, and D. C. Zeldin. Relationship between serum cholesterol and indices of erythrocytes and platelets in the us population. *Journal of Lipid Research*, (54):3177–3188, 11 2013.
- B. Gavish, I. Ben-Dov, and M. Bursztyn. Linear relationship between systolic and diastolic blood pressure monitored over 24 h: assessment and correlates. *Hypertens.*, 26(2):199–209, 2 2008.
- A. Gelmane, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. 08 2013.
- A. Gul and M. Rahman. Comparison of blood protein levels between diabetic and non-diabetic patients with retinopathy. *Coll Physicians Surg Pak.*, 16:408–411, 6 2006.
- J. Jia, Y. Yang, F. Liu, M. Zhang, Q. Xu, T. Guo, L. Wang, Z. Peng, Y. He, Y. Wang, Y. Zhang, H. Zhang, H. Shen, Y. Zhang, D. Yan, X. Ma, and P. Zhang. The association between serum alanine aminotransferase and hypertension: A national based cross-sectional analysis among over 21 million chinese adults. *BMC cardiovascular disorders*, 21, 3 2021.
- W. Koenig, M. Sund, E. Ernst, U. Keil, J. Rosenthal, and V. Hombach. Association between plasma viscosity and blood pressure. results from the monica-project augsburg. *Am J Hypertens.*, pages 529–36, 6 1991.

- N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, (38):963–974, 1982.
- C. Liu, J. Vehí, P. Avari, M. Reddy, N. Oliver, P. Georgiou, and P. Herrero. Long-term glucose forecasting using a physiological model and deconvolution of the continuous glucose monitoring signal. *Sensors*, 19, 2019.
- S. M. Mohamed, M. A. Shalaby, R. A. El-Shiekh, H. A. El-Banna, and A. F. Emam, S. R. and Bakr. Metabolic syndrome: Risk factors, diagnosis, pathogenesis, and management with natural approaches. *Food Chemistry Advances*, 3, 2023.
- S. Naqvi, S. Naveed, Z. Ali, S. Ahmad, K. R. Asadullah, H. Raj, S. Shariff, C. Rupareliya, F. Zahra, and S. Khan. Correlation between glycated hemoglobin and triglyceride level in type 2 diabetes mellitus. *Cureus*, (9), 6 2017.
- A. Scuteri, S. Laurent, F. Cucca, J. Cockcroft, P. Cunha, L. Mañas, F. Mattace Raso, M. Muiesan, L. Rylis̄kytė, E. Rietzschel, J. Strait, C. Vlachopoulos, H. Völzke, E. Lakatta, and P. Nilsson. Metabolic syndrome across europe: different clusters of risk factors. *European journal of preventive cardiology vol. 22,4*, pages 486–91, 2015.
- Stan Developing Team. Stan modeling language users guide and reference manual. URL <https://mc-stan.org/users/documentation/>.
- J. Vuong, Y. Qiu, M. La, G. Clarke, D. Swinkels, and G. Cembrowski. Reference intervals of complete blood count constituents are highly correlated to waist circumference: should obese patients have their own "normal values?". *R Am J Hematol*, (89):671–677, 7 2014.
- S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 04 2010.

# A | STAN Code

In this chapter, we report the Stan code used for the computation and sampling of *Model 2*.

The package `rstan` is the R interface to Stan (See Stan Developing Team) and its source code is hosted on GitHub. The `stan` function does all of the work of fitting a Stan model and returning the results as an instance of `stanfit`. The main steps are the following:

- translate the Stan model to C++ code;
- compile the C++ code into a binary shared object, which is loaded into the current R session (an object `stanmodel` is created);
- draw samples and wrap them in a `stanfit` object.

The returned object can be used with methods such as `print`, `summary`, and `plot` to inspect and retrieve the results of the fitted model.

```
data {
  int<lower=1> N; //number of subjects
  int<lower=1> P; //number of covariates
  int<lower=1> T; //number of total observations
  int<lower=1> subj[T]; //subject id vector
  int y[N]; //outcome
  matrix[N,P] X; //predictors
}

parameters {
  vector[P] beta; //fixed intercept and slope
  vector[N] b; //subject intercepts
  real<lower=0> eta; //sd for subject intercepts
  real<lower=0> sigma_e; //error sd
}
```

```
model {
  vector[T] mu;

  //priors
  mub ~ normal(0, 2); //subj random effects
  eta~ inv_gamma(3,2); //variance of intercepts
  b ~ normal(mub, eta); //subj random effects
  w ~ normal(0, sigma_w); //item random effects
  beta ~ normal(0,5);
  sigma_e ~ normal(0,5);

  // likelihood
  for (i in 1:T){
    mu = x[i]*beta + b[subj[i]];
  }

  y ~ normal(mu, sigma_e);
}

generated quantities {
  vector[T] log_lik; //log-likelihood

  for (i in 1:T) {
    // generate predicted value
    real y_hat = x[i]*beta + b[subj[i]];

    // calculate log-likelihood
    log_lik[i] = normal_lpdf(y[i] | y_hat, e);
    // normal_lpdf is the log of the normal probability density function

  }
}
```

# B | Variations of Model 2

In this appendix, different variations of *Model 2* were explored to determine if they could enhance the tool's predictive capabilities.

## B.1. Model 2\_A

It was shown in Chapter 1 that the target variables are not correlated. As we decided to not pursue the computation of a joint marginal distribution over all the target variables we hypothesized that adding the previous measurement of the remaining target variables may improve our model. The target variables are then added as covariates to each of the models but the one predicting their own value. This was done by adding the component  $T_i^k$  and the corresponding fixed-effect coefficients  $\boldsymbol{\tau}^k$  to *Model 2*.

For clarity, as an example, *Model 2\_A* for the variable Glucose ( $k = 1$ ) is reported:

$$\begin{aligned}
 \mathbf{Y}_i^k &= X_i^k \boldsymbol{\beta}^k + Z_i^k b_i^k + T_i \boldsymbol{\tau} + \boldsymbol{\epsilon}_i^k, & \boldsymbol{\epsilon}_i^k &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, R_i^k), & i &= 1, \dots, N, \\
 Z_i^k &= [1, \dots, 1]', & & & k &= 1, \dots, 5, \\
 \boldsymbol{\beta}^k &\sim \mathcal{N}(\mathbf{0}, \Sigma^k), \\
 b_i^k &\stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \eta^2), \\
 \mu &\sim \mathcal{N}(\mu_0, \sigma_{\mu_0}^2), \\
 \eta^2 &\sim \text{inv-gamma}(\alpha_0, \beta_0), \\
 \boldsymbol{\tau} &\sim \mathcal{N}(\mathbf{0}, \Omega).
 \end{aligned} \tag{B.1}$$

Where  $T_i$  is a matrix with  $n_i$  rows and four columns containing the measurements for HDL\_cholesterol, Triglycerides, Circumference, and PMAX.

## B.2. Model 2\_B

Another approach was to add an autoregressive component to the model since most of the target variables are usually modeled with ARIMA models.

ARIMA (AutoRegressive Integrated Moving Average) models are a class of time series models widely used for analyzing and forecasting time-dependent data. The standard notation for this model is  $ARIMA(p, d, q)$ , which extends the more widely known  $ARMA(p, q)$  model by applying it to a time series that has been differenced  $d$  times.

An  $ARMA(p, q)$  model describes the behavior of a time series through an AutoRegressive component (a polynomial of order  $p$ ), which models observations at time  $t$  as a linear function of previous observations up to time  $t - p$ , and a Moving Average component (a polynomial of order  $q$ ), which incorporates a linear combination of previous and current unobserved uncorrelated errors up to time  $t - q$ . A complete description of an  $ARMA(p, q)$  model involves assigning coefficients (which may potentially be null) to every degree of these polynomials, thereby achieving a comprehensive and time-invariant representation of the phenomenon under investigation.

An  $ARIMA(p, d, q)$  model extends the concept of an  $ARMA(p, q)$  model by incorporating a preprocessing step to ensure that the time series under examination exhibits time-invariant statistical properties, such as mean and temporal covariance. Specifically,  $ARIMA(p, d, q)$  models analyze the  $d$ -differenced, stationary series derived from the original by applying  $d$  successive 1-step differences ( $y_t - y_{t-1}$ ). For instance,  $ARIMA(p, 1, q)$  models are used to eliminate linear trends in data, such as increasing or decreasing observations mean over time.

ARIMA models rely on the assumption of stationarity, which breaks when the difference between sequential instances of the time series are not consistent throughout. As shown before, the donation times are not equally spaced so we cannot fit an ARIMA model directly over time. To determine if the target variables exhibit autocorrelation, we considered the time series of the number of the donation. This approach allows us to fit an ARIMA model to the time series even if the times are not equally spaced.

For each target variable we analyzed the frequency of the value of each parameter of the ARIMA model fitted on the singular donor. In Figures B.1 to B.5 the result are shown.

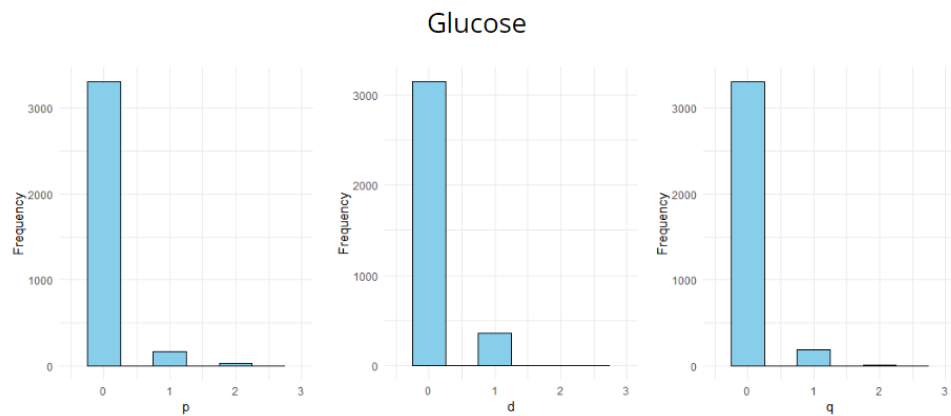


Figure B.1: Frequencies of parameters  $p, d$  and  $q$  of the ARIMA model for the Glucose variable

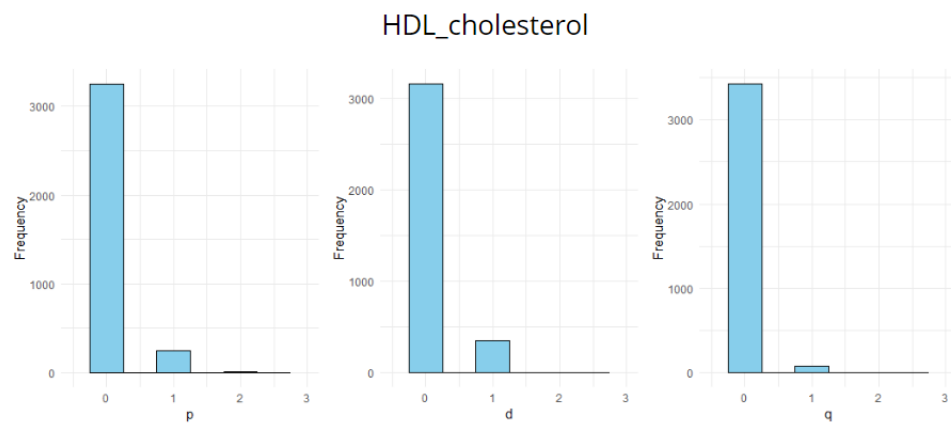


Figure B.2: Frequencies of parameters  $p, d$  and  $q$  of the ARIMA model for the HDL\_cholesterol variable

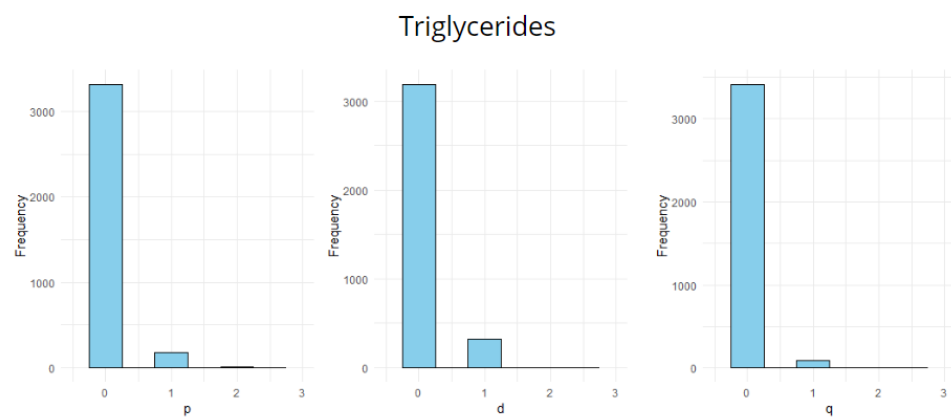


Figure B.3: Frequencies of parameters  $p, d$  and  $q$  of the ARIMA model for the Triglycerides variable

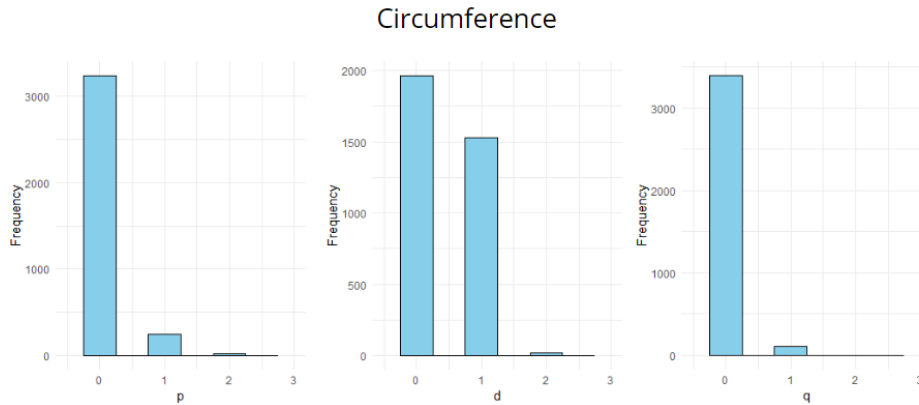


Figure B.4: Frequencies of parameters  $p, d$  and  $q$  of the ARIMA model for the Circumference variable

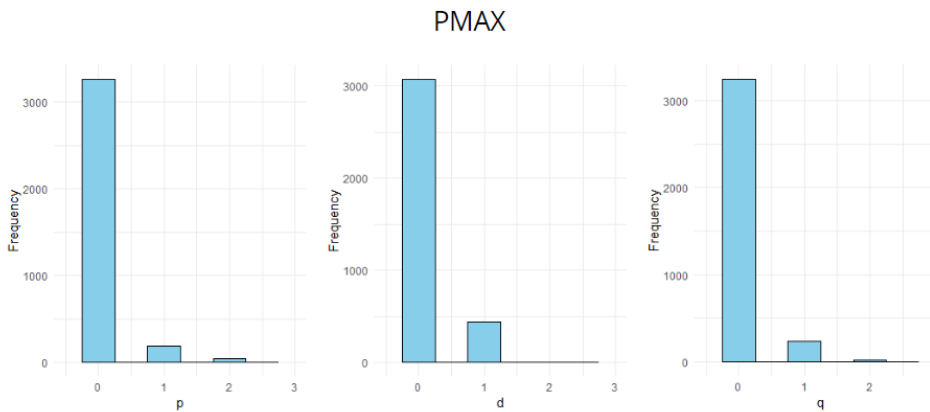


Figure B.5: Frequencies of parameters  $p, d$  and  $q$  of the ARIMA model for the PMAX variable

We observe that for all the target variables the most commonly fitted model is the model ARIMA(0,0,0). This can be explained by the long intertimes between donations. It is known that variables such as Glucose or HDL cholesterol are highly autocorrelated but most studies analyze this autocorrelation over a short time window, usually between 60 and 120 minutes (see Liu et al. (2019)). On the other hand, the repeated measurements for our target variable have a mean gap interval of 4 to 6 months. For this reason, an autocorrelation term is rarely present.

To potentially improve predictions, we attempted to include an autocorrelation term by using *Model 2* as a base.

### B.3. Comparison

In Table B.1 the WAIC values are shown for each of these variations and are compared to those of *Model 2*. It is evident that the fitness of the model does not improve using these variations.

	Model 2	Model 2A	Model 2B
Glucose	-73588.1	-33386.2	-71827.4
HDL_cholesterol	-30772.7	-30720.3	-28567.1
Triglycerides	-52718.3	-51229.3	-50671.3
Circumference	-89117.3	-86449.9	-83381.1
PMAX	-114739.7	-111769.8	-111557.4

Table B.1: WAIC value for each variation of *Model 2*

In Table B.2 the F-score values for the prediction are shown. Different threshold levels have been considered when assuming the probability of the onset of metabolic syndrome to be 1. We can notice that the variations give a slightly better prediction but the increase in the F1-score is not significant enough to prefer one of these models to *Model 2*.

Threshold	Model 2	Model 2A	Model 2B
0.02	0.2229730	0.2200957	0.2452107
0.06	0.2988506	0.3333333	0.3558282
0.1	0.3823529	0.4197531	0.3937008
0.14	0.3826087	0.4918033	0.4259259
0.18	0.4166667	0.4912281	0.4329897

Table B.2: F1\_score value for each model



# C | Horse-shoe prior and feature selection

In this appendix, we briefly describe the theoretical background of the horseshoe prior and provide an example of its application to Dataset 1 for the Glucose variable. This example demonstrates how variable selection was performed. Furthermore, some plots showing the convergence of significant variables in *Model 2* are reported here. For the sake of simplicity, the subscript  $k$  is omitted.

## C.1. Horseshoe Prior

Each response variable  $Y$  for each habitual donor  $i$  at each donation has distribution  $(\mathbf{Y}_i | \boldsymbol{\theta} \mathcal{N}(\boldsymbol{\beta}X_i, \mathbf{R}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'))$ . The horseshoes prior, assumes that each component  $\beta_p$  of  $\boldsymbol{\beta}$ , is conditionally independent given  $\lambda_p$  and  $\tau$ , with density  $\pi_{HS}(\beta_p | \tau)$ , where  $\pi_{HS}$  can be represented as a scale mixture of normals:

$$\begin{aligned} (\beta_p | \lambda_p, \tau) &\sim \mathcal{N}(0, \lambda_p^2 \tau^2), \\ \lambda_p &\sim C^+(0, 1), \\ \tau &\sim C^+(0, 1). \end{aligned} \tag{C.1}$$

$C^+(0, 1)$  is a half-Cauchy distribution for the standard deviation  $\lambda_p$  and  $\tau$ . We refer to the  $\lambda_p$ 's as the local shrinkage parameters and to  $\tau$  as the global shrinkage parameter. Further details on the horseshoe prior can be found in Carvalho et al. (2009). The horseshoe prior was chosen for feature selection because, as shown in Figure C.1, it has flat, Cauchy-like tails that maintain large signals in the posterior distribution. On the other hand, it has an infinitely tall spike at the origin, leading to severe shrinkage for the zero elements of  $\boldsymbol{\beta}$ .

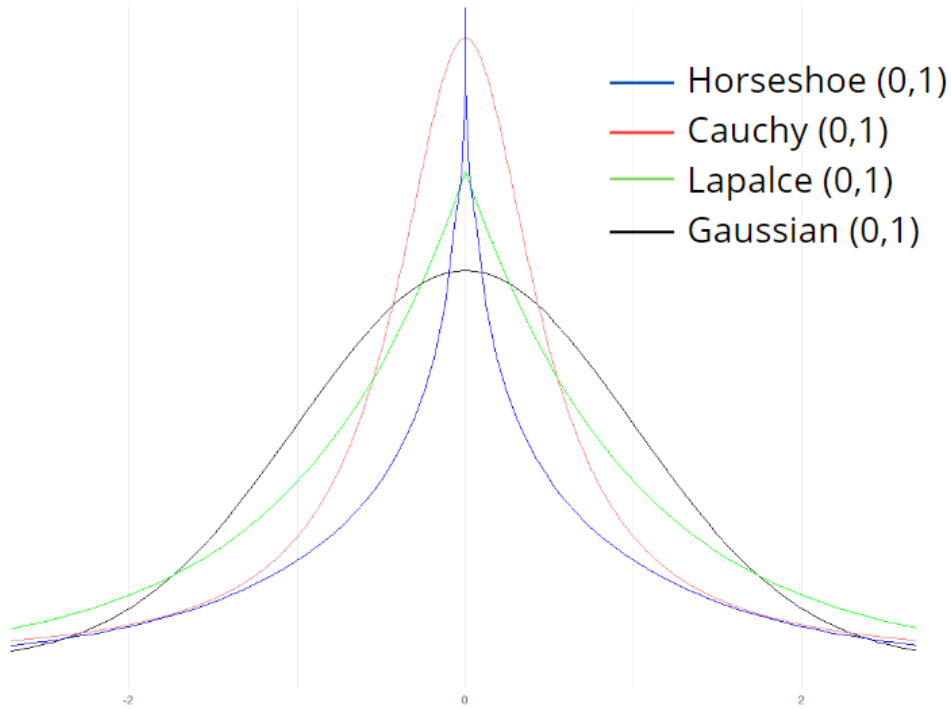


Figure C.1: Comparison of the horseshoe prior density versus Cauchy and Laplace densities

## C.2. Feature selection

To select the covariates from Tables 2.1 and 2.2 that compose the  $X_i^k$  matrix, multiple selection criteria have been taken into consideration.

Firstly all significant covariates, i.e. whose credible intervals do not contain 0, have been kept for the analysis. Secondly, for the remaining variables, the probability of direction has been considered. This value represents the probability that the effect goes to the positive or to the negative direction, and it is considered as the best equivalent for the p-value. In our analysis, we kept only the covariates with a probability of direction value over 80%.

# D | Convergence diagnostic

In this appendix, a convergence diagnostic is performed using the trace plot method. Trace plots depict the evolution of parameters across 1500 sampling iterations per chain, excluding warm-up iterations. They are crucial for assessing the chain mixing. In these plots, we aim to identify and minimize flat bits (where the chain remains in the same state for too long) and excessive consecutive steps in one direction. Figures D.1 to D.3 display the trace plots for all beta covariates of *Model 2* related to the target variable HDL cholesterol.

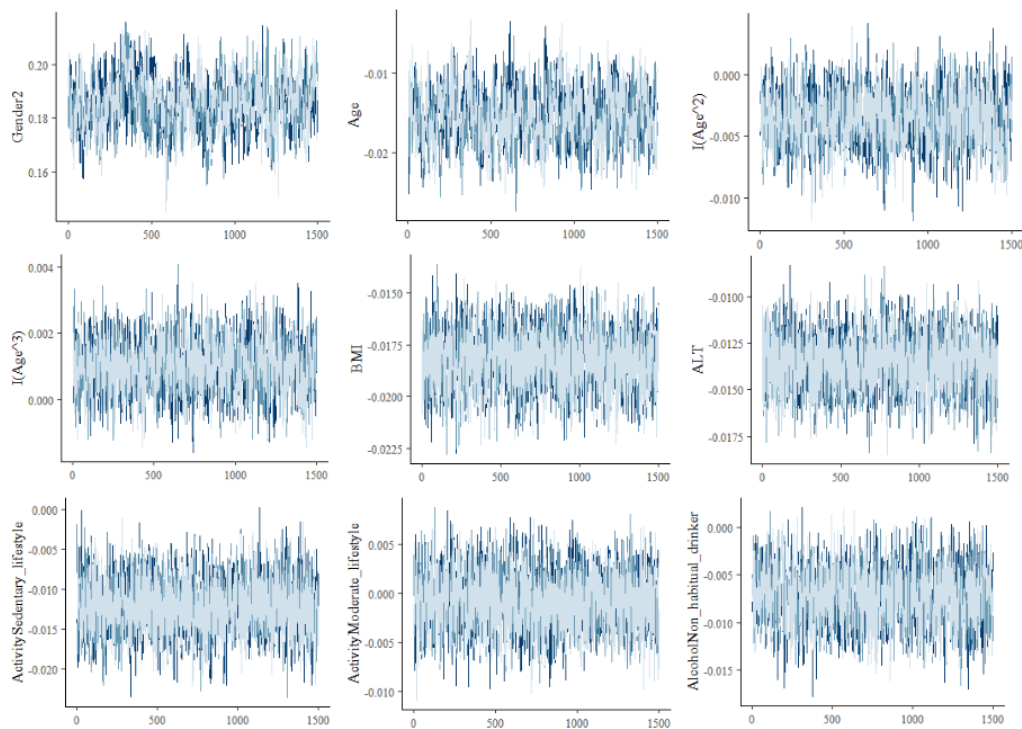


Figure D.1: Trace plots of *Model 2* for the HDL\_cholesterol variable

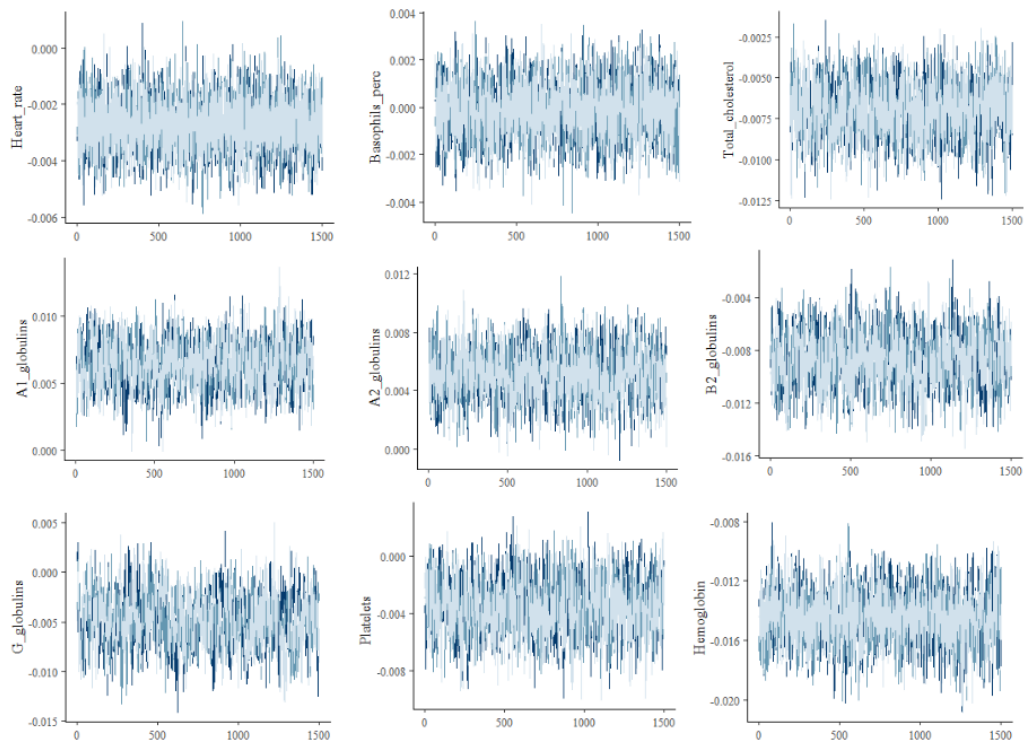


Figure D.2: Trace plots of *Model 2* for the HDL\_cholesterol variable (continued)

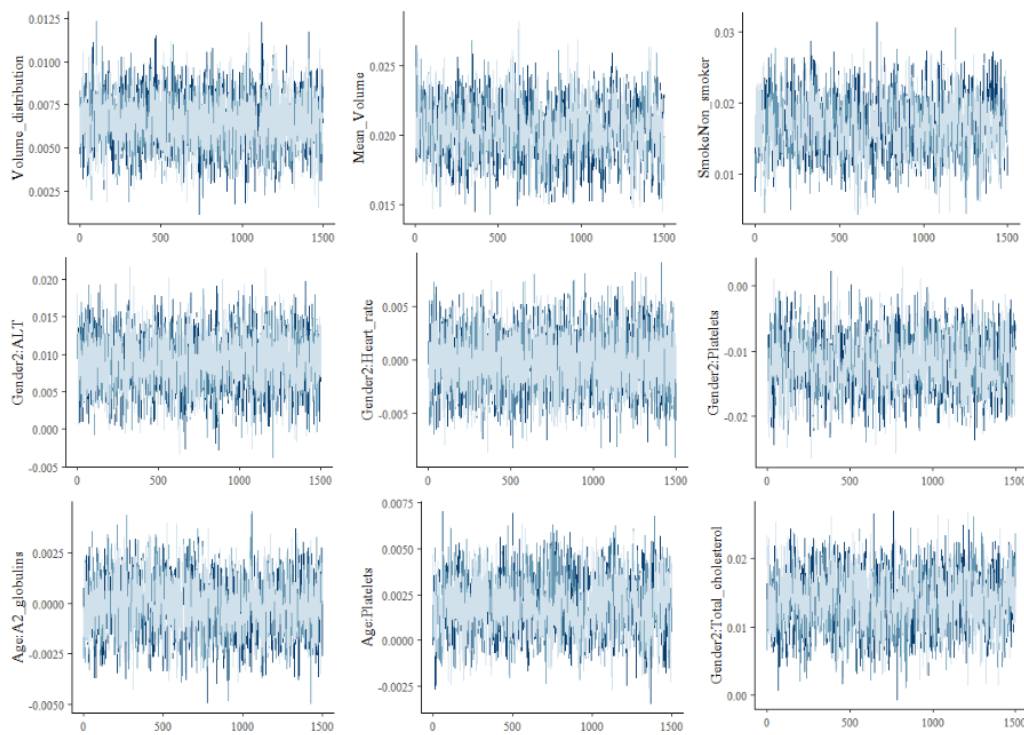


Figure D.3: Trace plots of *Model 2* for the HDL\_cholesterol variable (continued)

For the most significant covariates of the HDL\_cholesterol model, the autocorrelation plots are provided in Figures D.4 to D.7. These graphs show the autocorrelation for each Markov chain separately up to a user-specified number of lags. Positive autocorrelation means the chain tends to stay in the same area between iterations, a good graph has the autocorrelation drop quickly to zero as the lag increases. Negative autocorrelation, while less common, suggests rapid convergence of the sample mean towards the true mean.

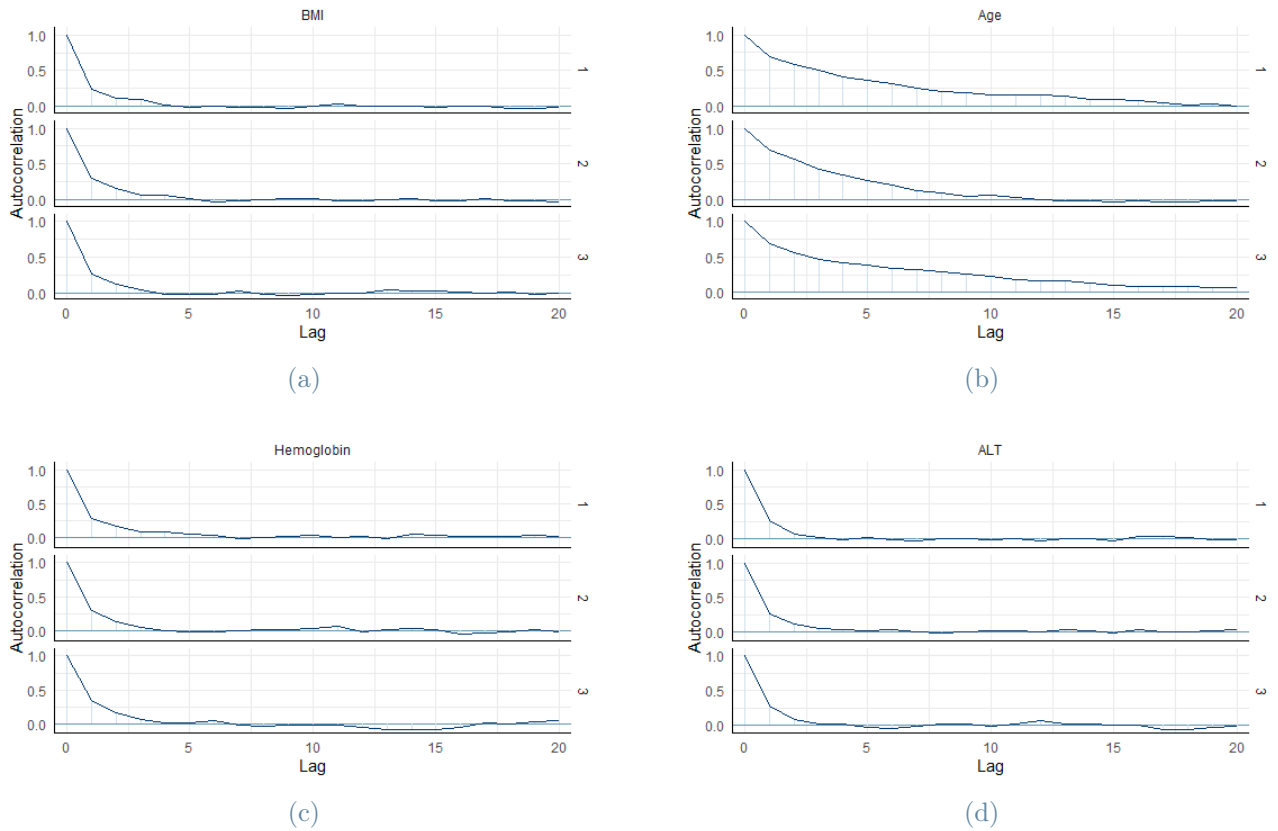
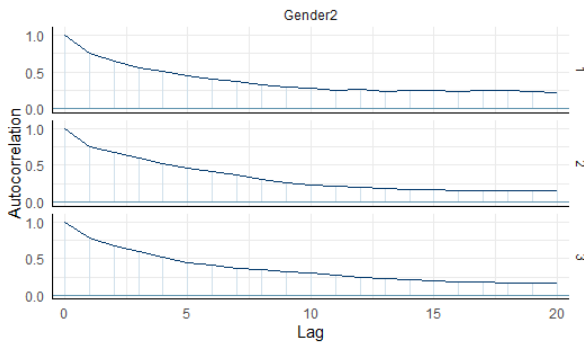
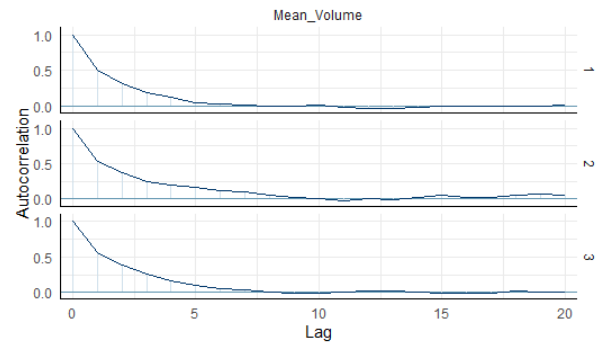


Figure D.4: Autocorrelation plots for variables with a negative effect on HDL\_cholesterol, under *Model 2*

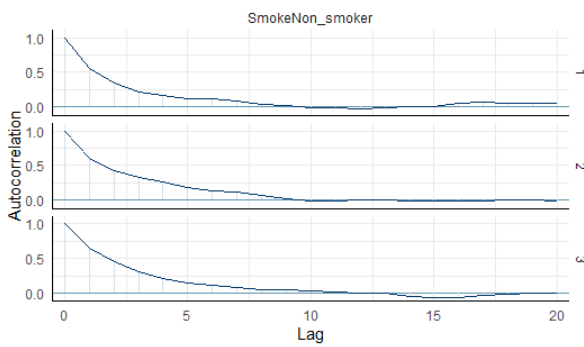
## D| Convergence diagnostic



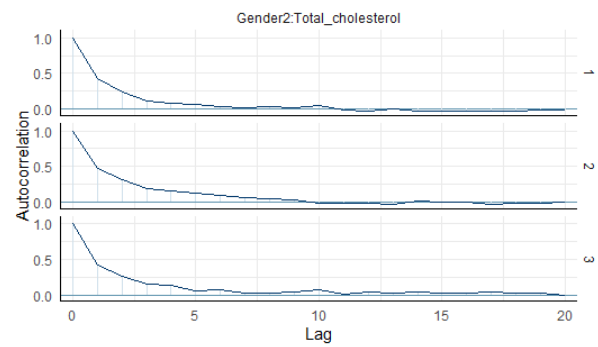
(a)



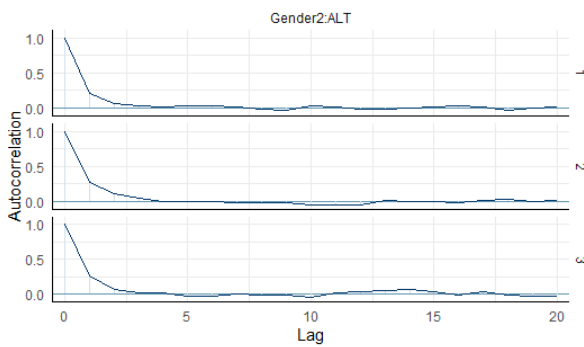
(b)



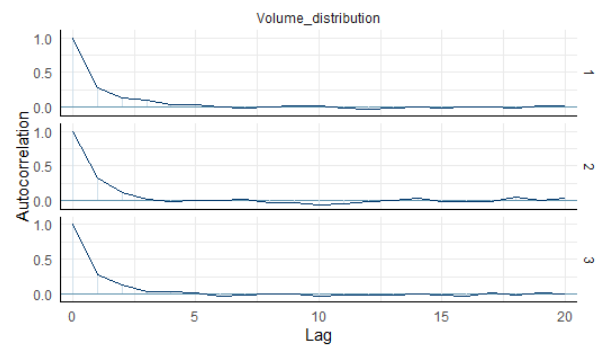
(c)



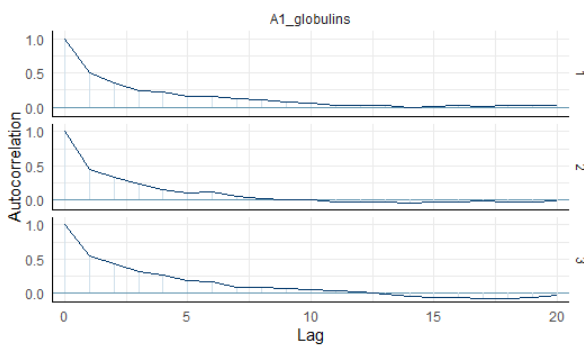
(d)



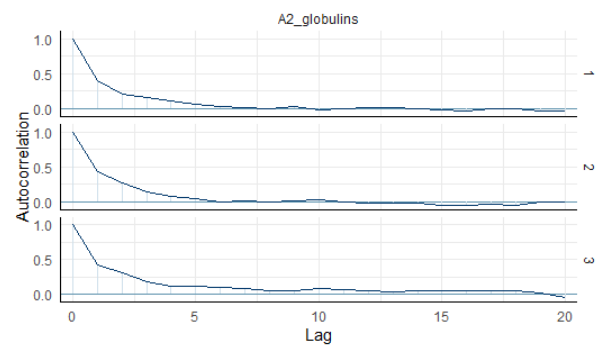
(e)



(f)



(g)



(h)

Figure D.5: Autocorrelation plots for variables with a positive effect on HDL\_cholesterol, under *Model 2*

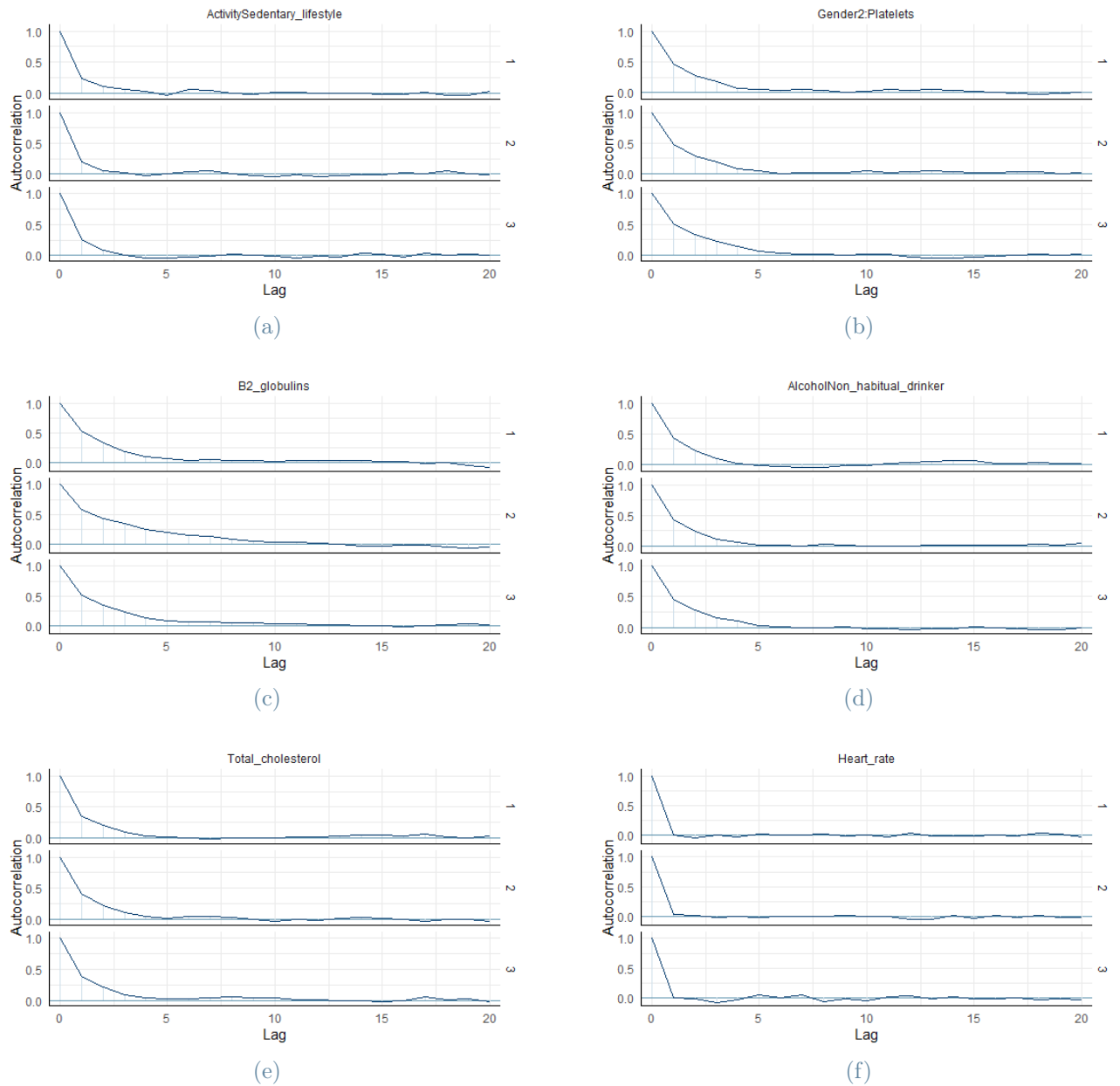
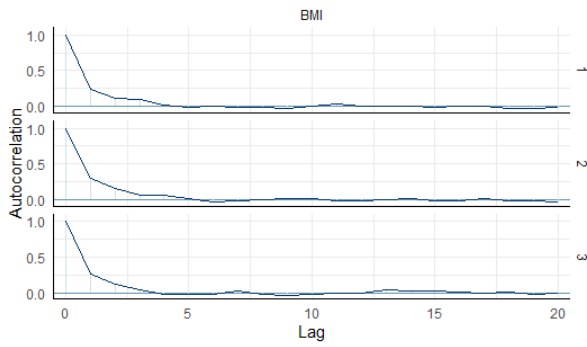
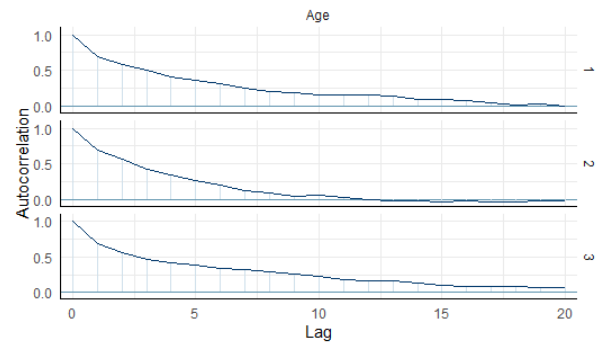


Figure D.6: Autocorrelation plots for variables with a negative effect on HDL\_cholesterol, under *Model 2*

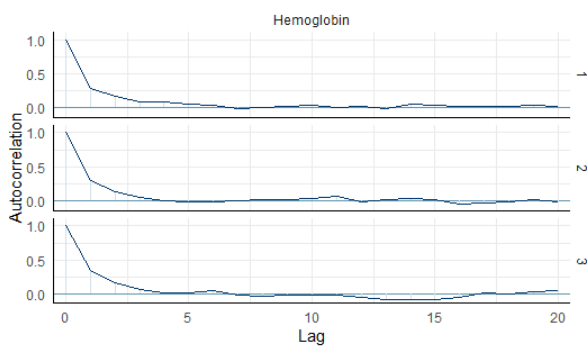
## D| Convergence diagnostic



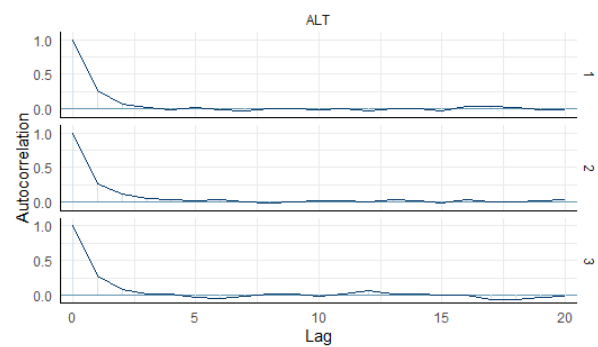
(a)



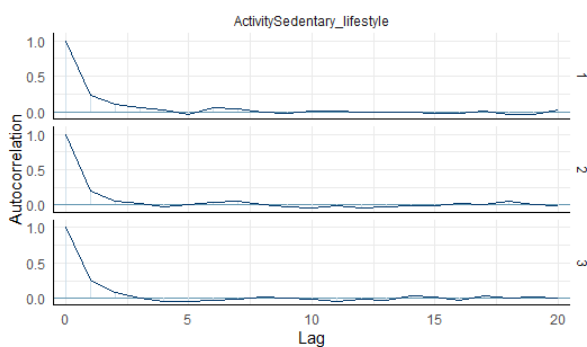
(b)



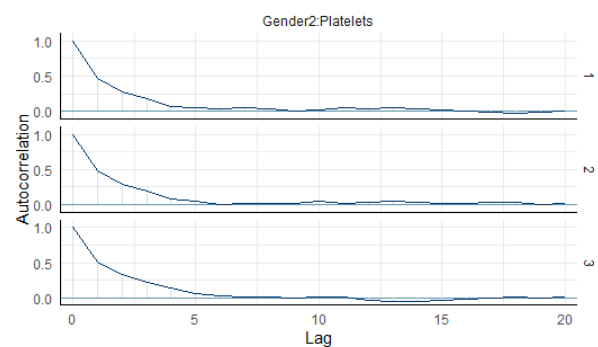
(c)



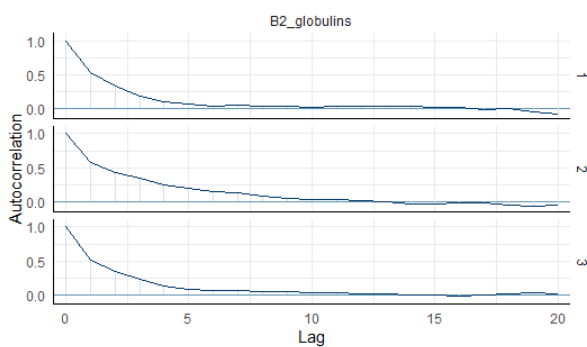
(d)



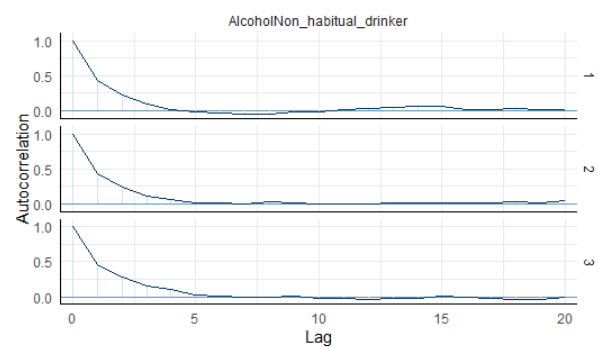
(e)



(f)



(g)



(h)

Figure D.7: Autocorrelation plots for variables with a negative effect on HDL\_cholesterol, under *Model 2* (continued)

## List of Figures

1.1	Proportion of missing exams . . . . .	10
1.2	Proportion of missing exams on the remaining variables . . . . .	11
1.3	Correlation between numeric covariates . . . . .	16
1.4	Correlation between target variables . . . . .	17
1.5	Gaptimes on log scale between successive measurements of Glucose. . . . .	18
1.6	Gaptimes on log scale between successive measurements of HDL_cholesterol. . . . . .	19
1.7	Gaptimes on log scale between successive measurements of Triglycerides. . . . .	19
1.8	Gaptimes on log scale between successive measurements of Circumference. . . . .	20
1.9	Gaptimes on log scale between successive measurements of PMAX. . . . .	20
1.10	Histograms of the target variables . . . . .	22
1.11	Histograms of the target variables on log scale . . . . .	23
1.12	Boxplot and histograms of glucose levels, grouped by gender . . . . .	24
1.13	Boxplot and histograms of Hdl cholesterol levels, grouped by gender . . . . .	24
1.14	Boxplot and histograms of triglycerides levels, grouped by gender . . . . .	25
1.15	Boxplot and histograms of waist circumference, grouped by gender . . . . .	25
1.16	Boxplot and histograms of systolic pressure, grouped by gender . . . . .	26
1.17	Boxplot and histograms of diastolic pressure, grouped by gender . . . . .	26
1.18	Division of the datasets . . . . .	27
2.1	Logistic function . . . . .	39
3.1	$\beta^{(1)}$ s posterior credible intervals, under <i>Model 2</i> . . . . .	46
3.2	$\beta^{(2)}$ s posterior credible intervals, under <i>Model 2</i> . . . . .	48
3.3	$\beta^{(3)}$ s posterior credible intervals, under <i>Model 2</i> . . . . .	50
3.4	$\beta^{(4)}$ s posterior credible intervals, under <i>Model 2</i> . . . . .	52
3.5	$\beta^{(5)}$ s posterior credible intervals, under <i>Model 2</i> . . . . .	54
3.6	Posterior densities of $b_i^1$ under <i>Model 2</i> , blu curves represent males and red represent female . . . . .	56
3.7	Posterior densities of $b_i^1$ under <i>Model 2</i> , colored by age of the donor . . . . .	57

3.8	Posterior densities of $b_i^2$ under <i>Model 2</i> , blu curves represent males and red represent female . . . . .	57
3.9	Posterior densities of $b_i^2$ under <i>Model 2</i> colored by age of the donor . . . . .	58
3.10	Posterior densities of $b_i^3$ under <i>Model 2</i> , blu curves represent males and red represent female . . . . .	58
3.11	Posterior densities of $b_i^3$ under <i>Model 2</i> , colored by age of the donor . . . . .	59
3.12	Posterior densities of $b_i^4$ under <i>Model 2</i> , blu curves represent males and red represent female . . . . .	59
3.13	Posterior densities of $b_i^4$ under <i>Model 2</i> , colored by age of the donor . . . . .	60
3.14	Posterior densities of $b_i^5$ under <i>Model 2</i> , blu curves represent males and red represent female . . . . .	60
3.15	Posterior densities of $b_i^5$ under <i>Model 2</i> , colored by age of the donor . . . . .	61
4.1	Comparison of the distribution of the total probability $p_i(\mathbf{w}_i)$ between the logistic regression model and the combinatorial model under the target . . . . .	64
4.2	Comparison of the distribution of the total probability between the male and female population for the logistic regression results and the combinatorial method result . . . . .	65
4.3	Classification tools for different threshold values for the logistic regression method . . . . .	66
4.4	Classification parameters for different threshold values for the combinatorial method . . . . .	66
4.5	Zoomed portion of Figure 4.1, the red lines at $p_i(\mathbf{w}_i) = 0.04$ and $p_i(\mathbf{w}_i) = 0.16$ represent the proposed thresholds . . . . .	68
4.6	Confusion matrixes for threshold 0.04 (left) and 0.16 (right) . . . . .	69
4.7	Classification summaries . . . . .	69
4.8	Classification summary with absolute frequencies divided by gender . . . . .	70
4.9	Classification summary with percentage frequencies divided by gender . . . . .	70
4.10	Interface of the tool prototype for AVIS . . . . .	71
4.11	Possible outputs . . . . .	72
4.12	Additional output in the case of decreased weight . . . . .	72
4.13	Additional output in the case of increased weight . . . . .	72
B.1	Frequencies of parameters $p, d$ and $q$ of the ARIMA model for the Glucose variable . . . . .	81
B.2	Frequencies of parameters $p, d$ and $q$ of the ARIMA model for the HDL_cholesterol variable . . . . .	81

B.3	Frequencies of parameters $p, d$ and $q$ of the ARIMA model for the Triglycerides variable . . . . .	81
B.4	Frequencies of parameters $p, d$ and $q$ of the ARIMA model for the Circumference variable . . . . .	82
B.5	Frequencies of parameters $p, d$ and $q$ of the ARIMA model for the PMAX variable . . . . .	82
C.1	Comparison of the horseshoe prior density versus Cauchy and Laplace densities . . . . .	86
D.1	Trace plots of <i>Model 2</i> for the HDL_cholesterol variable . . . . .	87
D.2	Trace plots of <i>Model 2</i> for the HDL_cholesterol variable (continued) . . .	88
D.3	Trace plots of <i>Model 2</i> for the HDL_cholesterol variable (continued) . . .	88
D.4	Autocorrelation plots for variables with a negative effect on HDL_cholesterol, under <i>Model 2</i> . . . . .	89
D.5	Autocorrelation plots for variables with a positive effect on HDL_cholesterol, under <i>Model 2</i> . . . . .	90
D.6	Autocorrelation plots for variables with a negative effect on HDL_cholesterol, under <i>Model 2</i> . . . . .	91
D.7	Autocorrelation plots for variables with a negative effect on HDL_cholesterol, under <i>Model 2</i> (continued) . . . . .	92



## List of Tables

1.1	Time-fixed covariates from the EMONET dataset . . . . .	7
1.2	Time-dependent categorical covariates from the EMONET dataset . . . . .	7
1.3	Time-dependent numerical covariates from the EMONET dataset part 1 . . . . .	8
1.4	Time-dependent numerical covariates from the EMONET dataset part 2 . . . . .	9
1.5	Time-dependent covariates from the AVIS dataset . . . . .	9
1.6	Prescription period associated with variables having a high proportion of missing data . . . . .	11
1.7	Percentage of levels . . . . .	12
1.8	Aggregated responses for the exams' categorical variables . . . . .	13
1.9	Aggregated responses for the variable Smoke . . . . .	13
1.10	Aggregated responses for the variable Alcohol . . . . .	14
1.11	Aggregated responses for the variable Activity . . . . .	14
1.12	Percentage of levels . . . . .	15
1.13	Summary of gap times between donations . . . . .	18
1.14	Number of implausible values . . . . .	21
1.15	Summary statistics after absurd values' removal . . . . .	22
2.1	Complete list of covariates . . . . .	32
2.2	Complete list of covariates (continued) . . . . .	33
3.1	WAIC value for each model . . . . .	44
3.2	LOO value for each model . . . . .	44
3.3	Posterior summaries of the $\beta$ s parameters for the log Glucose . . . . .	45
3.4	Posterior summaries of the $\beta$ s parameters for the log HDL_cholesterol . . . . .	47
3.5	Posterior summaries of the $\beta$ s parameters for the log Triglycerides . . . . .	49
3.6	Posterior summaries of the $\beta$ s parameters for the log Circumference . . . . .	51
3.7	Posterior summaries of the $\beta$ s parameters for the log PMAX . . . . .	53
3.8	Posterior summaries of the $\beta$ s parameters for the log PMAX (continued) . . . . .	54
3.9	Summary of the means of $b_i$ 's parameters distribution . . . . .	55
4.1	Classification parameters for thresholds 0.04 and 0.16 . . . . .	68

B.1	WAIC value for each variation of <i>Model 2</i> . . . . .	83
B.2	F1_score value for each model . . . . .	83

## Acknowledgements

Desidero esprimere la mia più profonda gratitudine a tutte le persone che mi hanno supportato e guidato durante il percorso di redazione di questa tesi.

In primo luogo, vorrei ringraziare le Professoressse Alessandra Guglielmi e Ilenia Epifani, per la loro costante disponibilità, pazienza e preziosi consigli.

Un altro sentito ringraziamento va al Dr. Sergio Casartelli da cui è originata l'idea per questa tesi e ad AVIS Milano che ha fornito i dati su cui basarsi.

Un ringraziamento speciale va alla mia famiglia, che mi ha sostenuto in ogni momento, offrendo incoraggiamento e amore incondizionato. Grazie ai miei genitori per aver sempre creduto in me e per avermi fornito tutte le opportunità necessarie per raggiungere i miei obiettivi. Un grazie speciale va a Chiara la migliore sorella che si possa avere e che in questo percorso di tesi ha risposto a tutti i miei dubbi medici e biologici.

Grazie anche a tutti i parenti per la loro vicinanza e il loro affetto.

Un grazie di cuore a tutti i miei amici che in diversi modi mi hanno sempre spinto a non arrendermi durante questo percorso. Ringrazio quindi gli amici dell'università con cui studiare per gli esami diventava meno pesante, gli amici del treno che hanno saputo riempire momenti di silenzio e di noia, gli amici visti di rado ma sempre nei miei pensieri e che ho tormentato con infiniti audio sulle gioie e le lamentele della vita. Un grazie anche agli amici degli svaghi, coloro che settimana dopo settimana hanno condiviso la mia passione per i giochi, il canto e i libri.

Grazie a tutti voi, senza il vostro sostegno questo traguardo non sarebbe stato possibile.

