Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Fride Nordstrand Nilsen

Prediction Models of Systolic Blood Pressure Based on HUNT Study Data

Master's thesis in Applied Physics and Mathematics
Supervisor: Ingelin Steinsland

May 2020

NTNU
Norwegian University of
Science and Technology

Fride Nordstrand Nilsen

Prediction Models of Systolic Blood Pressure Based on HUNT Study Data

Master's thesis in Applied Physics and Mathematics
Supervisor: Ingelin Steinsland
May 2020

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

In this thesis, prediction models of systolic blood pressure are proposed, implemented, evaluated, and compared to the Framingham model, based on data from The Troendelag Health Study, hereafter shortened to the HUNT Study. The ability of the models to classify the binary systolic hypertension status of the participants is also evaluated. In addition to this, we study the effect of physical activity, measured by PAI (Personal Activity Intelligence), on systolic blood pressure. The goal of the prediction models is to predict the systolic blood pressure at HUNT3 for people with initially healthy blood pressure at HUNT2, based on information from HUNT2.

Firstly, we examine the available data set from HUNT and select the relevant population and variables from the total available data set. Secondly, we correct the effect of blood pressure medication on the systolic blood pressure at HUNT3 for the people using this kind of medication at the time of HUNT3. The final data set includes the systolic blood pressure, and 15 relevant explanatory variables from HUNT2, as well as a few other variables with various information, for n=17 365 participants. We perform an exploratory data analysis on the final data set, where the main results are that the distribution of systolic blood pressure at HUNT3 is approximately normal with a somewhat heavier right tail, and the systolic blood pressure at HUNT3 is mainly correlated with the systolic and diastolic blood pressure at HUNT2, birth year and BMI at HUNT2. Before fitting the models we standardize the explanatory variables.

We consider four prediction models; a small and large version of a Gaussian generalized linear model, and a small and a large version of a gamma generalized linear model. In addition to this, we implement a modified version of the Framingham model, which is a well-known prediction model of hypertension risk from literature, on our data set. We immediately observe that the fitted prediction generalized linear models have very similar regression coefficients and residuals. Furthermore, we discover that the standard deviation of the residuals depends linearly on the predicted systolic blood pressure and on the explanatory variables. We also observe that the effect of physical activity, measured by PAI, on the predicted systolic blood pressure is surprisingly small. Finally, we evaluate the performance of the models with some common evaluation methods such as root mean squared error, Brier score, Continuous Rank Probability Score, PIT diagrams, sensitivity, specificity, and C-statistic.

We conclude that the prediction models we propose are able to identify some clear trends in the data, for instance the importance of birth year and previous systolic and diastolic blood pressure. Furthermore, they generally predict a higher probability of systolic hypertension for the participants who become systolic hypertensive, and have a C-statistic similar to C-statistic of the Framingham model by Parikh et al. (2008). However, the variances in the individual prediction distributions are large and the models are not able to accurately predict the systolic blood pressure at HUNT3. As possible future work we suggest including lifestyle explanatory variables from later time points, for instance HUNT3, and choosing a prediction model that models the variance.

Sammendrag

I denne oppgaven blir prediksjonsmodeller for systolisk blodtrykk foreslått, implementert, evaluert og sammenlignet med Framingham modellen, basert på data fra Helseundersøkelsen i Trøndelag, heretter forkortet til HUNT-studien. Modellens evne til å klassifisere den binære systoliske hypertensjon statusen til deltakerne vil også evalueres. I tillegg til dette så ser vi nærmere på effekten fysisk aktivitet, målt gjennom PAI (Personlig Aktivitets-Intelligens), har på det systoliske blodtrykket. Målet til prediksjonsmodellene er å predikere det systoliske blodtrykket ved HUNT3 for personer som i utgangspunktet har sunt blodtrykk ved HUNT2, basert på informasjon fra HUNT2.

Vi starter med å utforske det tilgjengelige datasettet fra HUNT og velge ut de relevante deltakerne og variablene. Deretter korrigerer vi effekten av blodtrykksmedisin på det systoliske blodtrykket ved HUNT3 for deltakerne som bruker slik medisin ved HUNT3. Det endelige datasettet inneholder det systoliske blodtrykket og 15 relevante forklaringsvariabler fra HUNT2, samt noen få ekstra variabler med diverse nyttig informasjon, for n= 17 365 deltakere. Vi utfører en utforskende dataanalyse av det endelige datasettet, der hovedresultatene er at distribusjonen til det systoliske blodtrykket ved HUNT3 er tilnærmet normalfordelt med en litt tyngre høyre hale, og at det systoliske blodtrykket ved HUNT3 hovedsakelig er korrelert med det systoliske og diastoliske blodtrykket ved HUNT2, fødselsår og BMI ved HUNT2. Vi standardiserer så forklaringsvariablene før vi tilpasser modellene.

Vi foreslår fire prediksjonsmodeller; en liten og en stor versjon av en Gaussisk generalisert lineær modell, og en liten og en stor versjon av en gamma generalisert modell. I tillegg til dette så implementerer vi en modifisert versjon av Framingham-modellen, som er en velkjent prediksjonsmodell for risk av hypertensjon fra litteraturen, på vårt datasett. Vi oppdager umiddelbart at de tilpassede prediksjonsmodellene har veldig like regresjonskoeffisienter og residualer. Videre ser vi at standardavviket til residualene avhenger lineært av det predikerte systoliske blodtrykket og forklaringsvariablene. Vi observerer også en overraskende liten effekt av fysisk aktivitet, målt gjennom PAI, på det predikerte systoliske blodtrykket. Til slutt, bruker vi noen kjente evalueringsmetoder som rot-middel-kvadrat-avvik, Brier score, Continuous Rank Probability Score, PIT diagram, sensitivitet, spesifisitet og C-statistikken til å evaluere modellenes prediksjoner.

Vi konkluderer med at prediksjonsmodellene vi foreslår er i stand til å identifisere noen klare trender i datasettet, for eksempel viktigheten av fødselsår og tidligere systolisk og diastolisk blodtrykk. Modellene predikerer stort sett høyere sannsynlighet av systolisk hypertensjon for de som blir systolisk hypertensive, og har en C-statistikk som er lik C-statistikken til Framingham modellen av Parikh et al. (2008). På den andre siden så er variansen i de individuelle prediksjonsfordelingene stor og modellene klarer ikke å gi nøyaktige prediksjoner av det systoliske blodtrykket ved HUNT3. Som mulig videre arbeid foreslår vi å inkludere livsstilsvariabler fra senere tidspunkter, for eksempel ved HUNT3, og å velge en prediksjonsmodell som modellerer variansen.

Preface

This thesis represents the last semester of my Master of Science degree in Applied Physics and Mathematics with a specialization in Industrial Mathematics from the Norwegian University of Science and Technology in Trondheim.

Firstly, I would like to thank the HUNT Cloud team for help with accessing and storing the HUNT data. Secondly, I would like to thank Emma Ingström for her helpfulness, and answers to my many questions regarding the HUNT data set and the medical details of hypertension. Thirdly, I would like to thank Filip Schjerven for helpful discussions about the statistical literature on blood pressure prediction. Lastly, I would like to thank my supervisor, Ingelin Steinsland, who has offered helpful advice and feedback. In addition to this, I appreciate that she has been so understanding of my need to make adjustments to be able to finish the thesis.

The last year of my degree has been one of the most challenging years of my life. I have been very demotivated, and the COVID-19 situation with social distancing in the spring has not helped. However, through the support of my family and friends, and the guidance of my supervisor, I have managed to complete my degree. I am very proud of myself and grateful to all the people who have helped me reach this goal.

Fride Nordstrand Nilsen
May 2020

Table of Contents

Chapter 1

Introduction

Essential hypertension is a medical condition that affects more than a billion people globally and is one of the leading causes of premature death according to the World Health Organization. The symptoms of hypertension are often vague, if there at all, which causes many people to have undetected hypertension. Since untreated hypertension increases the risk of heart attack, heart failure, irregular heartbeat, and kidney failure, this is a big problem (WHO, 2019). A way of predicting hypertension would make patients and doctors able to start early preventive measures and treatments, and thus decrease the human suffering and economic consequences caused by hypertension.

WHO (2019) defines hypertension as persistently elevated blood pressure. The blood pressure varies as the heart contracts and relaxes, and is often measured by the systolic and the diastolic blood pressure. The systolic blood pressure is defined as the maximum blood pressure when the heart contracts and the diastolic blood pressure is defined as the minimum blood pressure while the heart rests. The criteria for a hypertension diagnosis is if the systolic blood pressure is measured as greater than or equal to 140 mmHg, and/or the diastolic blood pressure is measured as greater than or equal to 90 mmHg, for both measurements taken on two separate days (WHO, 2019).

Throughout the course of a life, the systolic and diastolic blood pressure will naturally change. Usually the systolic and diastolic blood pressure increase with age until approximately the age of 50. However, while the systolic blood pressure tends to continue to increase, the diastolic blood pressure tends to flatten out, or even lower somewhat, after the age of 50. This explains why it is increasingly common to get hypertension as you age and why systolic hypertension is the most common form of hypertension for people above the age of 50. Some important lifestyle factors that have been shown to increase the risk of hypertension are too high body weight, too much salt, and alcohol and not enough fruit, vegetables, and potassium in the diet, and low levels of physical activity (Chobanian et al., 2003). In fact, Cornelissen and Smart (2013) has performed a systematic review and meta-analysis of studies that look at the effect of exercise on blood pressure and found that both endurance and resistance training lower the systolic and the diastolic blood pressure.

Many papers proposing and evaluating risk prediction models for hypertension have been published in the statistical literature. Sun et al. (2017) gives an overview of 26 such studies including a total of 48 risk prediction models for hypertension. The majority of the studies include traditional explanatory variables such as body mass index (BMI), age, systolic blood pressure, diastolic blood pressure and parental history of hypertension, etc., while only 6 studies include genetic risk scores. The studies have cohorts from the US, Europe, China, Japan, Korea, Iran, and India. Follow-up times vary from study to study, with the shortest at 3 years and the longest at 30 years. However, the majority of the studies have a follow-up time between 3 and 10 years. To predict the risk of hypertension the studies propose different methods, with logistic regression being the most common, followed by COX regression, and Weibull regression, and one case of linear regression. As a measure of the discrimination ability of the models, many of the papers report the area under the receiver-operator statistic (AUC) or the C-statistic (Harrell Jr. et al., 1996), and the results range from 60% to 90% for the C-statistic and from 0.64 to 0.97 for the AUC. As a measure of the calibration ability of the models the Hosmer-Lemeshow chi-square statistic (Hosmer and Lemeshow, 1989) is reported for 15 of the 48 models, and all of them report a value below 16. The Framingham model proposed by Parikh et al. (2008) has a good C-statistic and is one of four models with Hosmer-Lemeshow chi-square statistics below 5. In addition to this, only a few of the models have been externally validated, yet the Framingham model has been externally validated 7 times, the most times of any of the models in the review paper by Sun et al. (2017) by far. There were noticeable differences in the performances of the Framingham model on different populations (Sun et al., 2017).

The topic of this thesis was inspired by a project called ”A Digital Twin For Essential Hypertension Management And Treatment- My Medical Digital Twin”, hereafter shortened to MyMDT. It is a cross-disciplinary project lead by Prof. Ulrik Wisloeff, involving researchers from departments such as Medicine, Mathematics, Computer Science, etc. at the Norwegian University of Science and Technology (NTNU). To reach its goal of improving the prevention and treatment of hypertension, MyMDT will use machine learning to merge a physical model of the cardiovascular system with personal data collected from custom-made wearable sensors. The result will be a personalized digital representation of the user, called a medical digital twin, which can be used in a clinical decision support system (NTNU, 2020). MyMDT bases its models, in part, on data from the Troendelag Health Study, hereafter shortened to the HUNT Study.

The HUNT Study is a large longitudinal population health study in a county in Norway, which started in 1984 and is still ongoing. In total, the HUNT study has gathered health information and biological samples from over 230 000 participants. In addition to many other health variables, the HUNT study includes measurements of the systolic and diastolic blood pressure, and other variables related to the blood pressure. All the inhabitants in the county Troendelag in Norway who were over 20 years old at the time of the survey were invited to participate. This information, as well as more detailed information about the HUNT Study, can be found on the webpage of the HUNT Databank (`https://hunt-db.medisin.ntnu.no/hunt-db/#/`).

The goal of this thesis is to predict the systolic blood pressure at the time of HUNT3 for people with initially healthy blood pressure at HUNT2, based on data from HUNT2. We predict only the systolic blood pressure both for the sake of simplicity and because the review paper by He and Whelton (1999) found that there is a stronger association between systolic blood pressure and coronary heart disease, stroke, and end-stage renal disease. To reach this goal we use continuous generalized linear models, as well as a modified version of the Framingham model. We choose to compare our models to, and implement a modified version of, the Framingham model on our data because the Framingham model is a well-known model and has been externally validated many times. Even though the models proposed by us only predict the continuous systolic blood pressure, their ability to classify the binary systolic hypertension status of the participants at HUNT3 is also evaluated. In addition to this, we focus especially on the effect of the physical activity measurement PAI, proposed by Nes et al. (2017), on the predicted systolic blood pressure.

Both the MyMDT project and this thesis aim to create good prediction models of blood pressure and hypertension based on data from the HUNT Study. However, the MyMDT project also includes current data from wearable sensors, while the models in this thesis will base its predictions solely on information from HUNT2. In this respect, the results in this thesis can be seen as a benchmark for the MyMDT models.

In Chapter 2 we present the available data set, explain how we select the relevant data from the total data set, and perform an exploratory data analysis. The statistical framework is presented in Chapter 3, before we present the proposed prediction models, the Framingham model, and the evaluation schemes in Chapter 4. The numerical details of the models and their performances on the systolic blood pressure from HUNT3 are given in Chapter 5. We also compare the performances of the models in Chapter 5. The results are discussed, we reach a conclusion and suggest possible future work in Chapter 6.

Chapter 2

Data and exploratory analysis

This chapter aims to present the data set used in this thesis. This is done by presenting the available data set, explaining how the relevant data is selected from the total data set, and presenting the insights we gained through exploratory data analysis.

2.1 Available data set

In this thesis, we are working with data from the Troendelag Health Study, hereafter shortened to the HUNT Study, which is a large longitudinal population health study in a county in Norway. The study consists of questionnaire data, clinical measurements, and samples collected through four surveys named HUNT1 (1984-1986), HUNT2 (1995-1997), HUNT3 (2006-2008), and HUNT4 (2017-2019). All the inhabitants in the county over 20 years of age at the time of the survey were invited to participate in the surveys. This information, as well as more detailed information about the four HUNT surveys, can be found on the webpage of the HUNT Databank (`https://hunt-db.medisin.ntnu.no/hunt-db/#/`). The data available to us includes 237 variables for all the 78 962 people who participated in HUNT2 and/or HUNT3.

2.1.1 Relevant data

We are not interested in all of the available data. The reason for this is that our goal is to create prediction models of systolic blood pressure at HUNT3 from information from HUNT2, for people who are initially healthy with respect to blood pressure. To select the relevant data from the available data, we include participants who meet the inclusion criteria and exclude the rest. We have defined the inclusion criteria in cooperation with Emma Ingström, a PhD-student in the MyMDT-project. Our inclusion criteria, listed in order of importance, are

- the participant has participated in both HUNT2 and HUNT3

- the participant doesn't have any missing values of mean systolic or mean diastolic blood pressure measurements from HUNT3

- the participant is initially healthy with respect to blood pressure. In other words, the participant has no self-reported, or measured, history of cardiovascular disease, diabetes or hypertension at the time of HUNT2

- the participant has no missing values of cardiovascular disease, diabetes or hypertension at the time of HUNT2

- the participant doesn't have any missing values of their use of blood pressure medication at the time of HUNT3

- the participant has no missing values of the proposed explanatory variables (listed in Section 2.1.3)

- the participant has no missing values of cardiovascular disease or diabetes at the time of HUNT3

2.1.2 Selecting the population

We start by getting a quick overview of the available data set and find that it has 237 columns, one for each variable, and 78 962 rows, one for each participant. The columns are either factors or contain numeric values. Each participant is identified by a project person identification (PID) number, and there are no duplicates in this list, which means that there is a one-to-one correspondence between row and participant.

Since we are only interested in persons who participated in both HUNT2 and HUNT3, we remove the persons who only participated in one of the surveys. This was the case for 46 496 of the participants in the data set, and we are thus left with 32 466 participants.

We want to create prediction models of systolic blood pressure at HUNT3 for initially healthy people at HUNT2, so we remove participants who have missing blood pressure measurements at HUNT2, are not healthy with regard to blood pressure at the time of HUNT2, have missing blood pressure measurements from HUNT3, or missing information about their use of blood pressure medication at the time of HUNT3.

For the sake of clarity, we present our definition of cardiovascular disease, diabetes, and hypertension. Throughout this thesis we define cardiovascular disease, hereafter denoted CVD, as a self-reported history of either heart attack, angina pectoris, or stroke. A participant is defined as diabetic if they have a self-reported history of diabetes or if their measured non-fasting glucose level is above 11.1 mmol/L, as this probably indicates diabetes (Chobanian et al., 2003). In this thesis, we use a common definition of hypertension which is mean systolic blood pressure equal to or higher than 140 mmHg and/or mean diastolic blood pressure equal to or higher than 90 mmHg and/or current or previous usage of blood pressure medication (Sun et al., 2017).

In Figure 2.1 the health status of the participants at the time of HUNT2 are presented. It is clear from the figure that there are relatively few people with CVD or diabetes, while many of the participants are hypertensive. In fact, approximately 39% of the people who participated in both HUNT2 and HUNT3 were defined as hypertensive at the time of

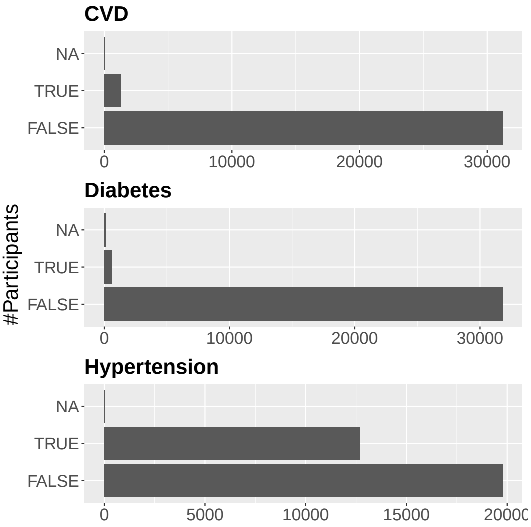


Figure 2.1: The health status at HUNT2 of the participants who participated in both HUNT2 and HUNT3, with regards to cardiovascular disease (CVD), diabetes, and hypertension. TRUE indicates that the participant has the illness, while FALSE indicates the opposite, and NA indicates a missing value.

HUNT2. We also observe that there seems to be a small number of missing values in these variables. To examine the missing values further, and get a closer look at the missing values in the other variables relevant for selecting the correct population, see Figure 2.2. After removing all the people who don't fulfill the health requirements and have missing values of blood pressure from HUNT2 or HUNT3, or the specified illnesses at HUNT2 or blood pressure medication use at HUNT3, we are left with 19 126 participants.

2.1.3 Considering explanatory variables

After selecting the population we want to study and use for our prediction model, a natural next step is to consider which explanatory variables to include in our model. Based on variables found to be important in Sun et al. (2017) and Parikh et al. (2008), and advice from Emma Ingström, a PhD-student also working on HUNT Study Data, we propose a set of variables from HUNT2 that we believe to be possibly important explanatory variables. The variables we consider are listed below with a short explanation. More detailed information can be found by searching for the variable name, given in parentheses, in the HUNT Databank (url: `https://hunt-db.medisin.ntnu.no/hunt-db/#/`). Proposed explanatory variables from HUNT2:

- **Mean systolic blood pressure** (`BPSystMn23@NT2BLM`) A numeric variable containing the rounded arithmetic mean of the second and third measurement of the systolic blood pressure. The measurements are given in mmHg, and were taken using a blood pressure cuff around the upper arm and a Dinamap device.

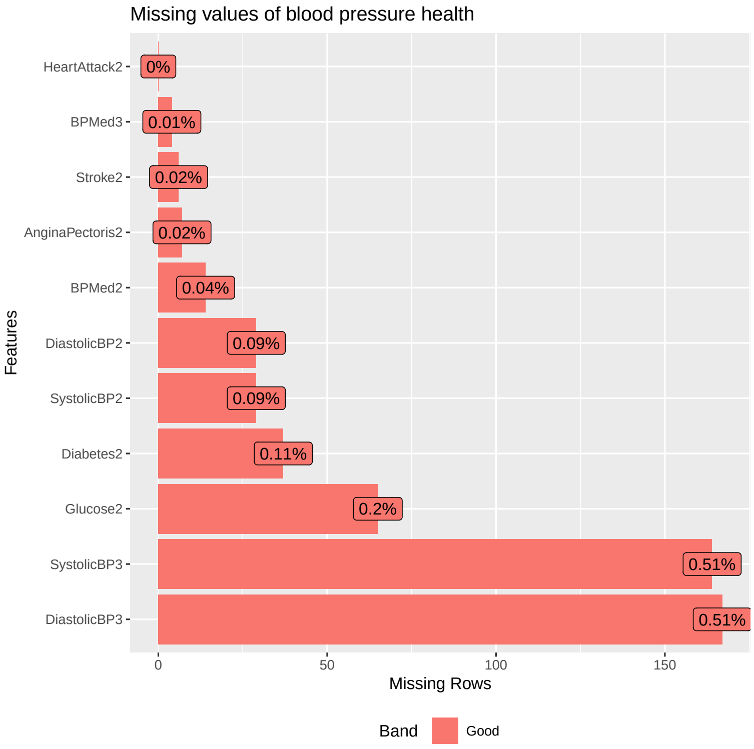


Figure 2.2: Missing values in variables of health regarding blood pressure of the people who participated in both HUNT2 and HUNT3. The percentage of missing values is shown for each relevant variable, and the number at the end of the variable name indicates whether the variable is from HUNT2 or HUNT3.

- **Mean diastolic blood pressure** (`BPDiasMn23@NT2BLM`) A numeric variable containing the rounded arithmetic mean of the second and third measurement of the diastolic blood pressure. The measurements are given in mmHg, and were taken using a blood pressure cuff around the upper arm and a Dinamap device.
- **Birth year** A numeric variable containing the year the participant was born. The values were found in The Norwegian National Registry.
- **Sex** A factor with two levels, ”Female” and ”Male”, describing the sex of the participants. These values were found in the Norwegian National Registry.
- **BMI (Body Mass Index)** (`Bmi@NT2BLM`) A numeric variable containing the Body Mass Index of the participant. This value is calculated by dividing a person’s weight in kilograms by the square of their height in meters (Keys et al., 1972), and is therefore measured in kg/m^2.
- **PAI (Personal Activity Intelligence)** A factor with the levels ”Low”, ”Moderate”

and ”High” describing the participants’ PAI score. A PAI score equal to or below 49 is categorized as ”Low”, a PAI score in the interval (49, 99] is categorized as ”Moderate” and a PAI score above 99 is categorized as ”High”. PAI, short for Personal Activity Intelligence, is a measure of physical activity defined by Nes et al. (2017). This score is calculated from HUNT variables describing the number of hours of self-reported light physical activity, `ExeLigDuLY@NT2BLQ1`, and hard physical activity, `ExeHarDuLY@NT2BLQ1`, per week during the last year.

- **RecPA** (Physical activity above/below recommended level) A boolean variable that describes whether the participant is meeting the recommended level of physical activity. It is `TRUE` if the physical activity of the participant is higher or equal to the recommended level, and `FALSE` if not. MVPA is a measure of physical activity defined by Ernstsen et al. (2016), and the recommended level of physical activity is defined as an MVPA score of 2.5. The MVPA score is derived from HUNT variables describing the number of hours of self-reported light physical activity, `ExeLigDuLY@NT2BLQ1`, and hard physical activity, `ExeHarDuLY@NT2BLQ1`, per week during the last year.

- **Hypertensive parents** A boolean variable which is `TRUE` if the participant has reported that one or both of their parents have ever been hypertensive, and `FALSE` otherwise. It is created from the HUNT variables `BPHigMothEv@NT2BLQ2`, `BPHigFathEv@NT2BLQ2`, `BPHigBrotEv@NT2BLQ2`, `BPHigSistEv@NT2BLQ2`, `BPHigChiEv@NT2BLQ2`, `BPHigFamNon@NT2BLQ2`, which describe the family history of hypertension.

- **Alcohol** A numerical variable that contains the total number of glasses of alcohol the participant has consumed during the last 14 days. This variable is created by adding the number of glasses of beer (`AlcBeL2WN@NT2BLQ1`), wine (`AlcWiL2WN@NT2BLQ1`) and spirits (`AlcLiL2WN@NT2BLQ1`) consumed during the last 14 days.

- **Smoking** (`SmoStat@NT2BLQ1`) A factor with the levels ”Never smoked daily”, ”Ex smoker daily”, and ”Current smoker daily”, which contains the self-reported smoking habits of the participant. For convenience, the levels are called Never, Previous, and Current, respectively, for the rest of this thesis.

- **Cholesterol** (`SeChol@NT2BLM`) A numerical variable which contains the cholesterol in a non-fasting blood sample from the participant. The measurements are given in mmol/L.

- **HDL Cholesterol** (`SeHDLChol@NT2BLM`) A numerical variable which contains the HDL cholesterol in a non-fasting blood sample from the participant. The measurements are given in mmol/L.

- **Non-fasting blood glucose** (`SeGluNonFast@NT2BLM`) A numerical variable which contains the glucose in a non-fasting blood sample from the participant. The measurements are given in mmol/L.

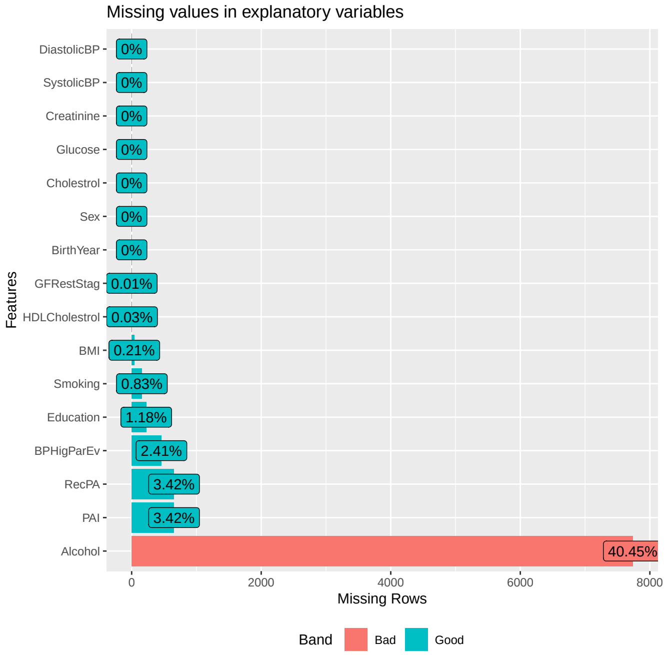


Figure 2.3: Missing values in the proposed explanatory variables. The percentage of missing values is shown for each explanatory variable. All the variables are from HUNT2.

- **GFR (Glomerular filtration rate)** (`SeGluNonFast@NT2BLM`) A factor with 5 levels ”Stage 1: GFREst $>$ 90 ml/min”, ” Stage 2: GFREst 60-89 ml/min”, ” Stage 3: GFREst 30-59 ml/min”, ”Stage 4: GFREst 15-29 ml/min” and ”Stage 5: GFREst $<$ 15 ml/min” which describes the estimated glomerular filtration rate stage of the participant. For convenience, the levels are called Stage 1, Stage 2, Stage 3, Stage 4, and Stage 5, respectively, for the rest of this thesis. The value is estimated from a blood sample from the participant.

- **Creatinine** (`SeCreaCorr.NT2BLM`) A numerical variable containing the creatinine level in a blood sample from the participant. The measurements are given in μmol/L.

- **Education level** (`Educ@NT2BLQ1`) A factor with five levels ”Primary school 7-10 years, continuation school, folk high school”, ”High school, intermediate school, vocational school, 1-2 years high school”, ”University qualifying examination, junior college, A levels”, ”University or other post-secondary education, less than 4 years”, ”University/college, 4 years or more”, which describes the participants highest level of education. For convenience, the levels are called Level 1, Level 2, Level 3, Level 4, and Level 5, respectively, for the rest of this thesis.

We would like to include only participants with no missing values of the explanatory variables included in the model. To examine if any of the variables have too many missing values, such that it might not be worth including it as an explanatory variable, the number and percentage of missing values in each proposed explanatory variable is shown in Figure 2.3. It is clear that alcohol has the biggest amount of missing values. In fact, more than 40% of participants have missing information about their alcohol consumption during the last 14 days. Removing these people would downsize our data set by a great amount. In addition to this, several studies have not listed alcohol as significant in relation to blood pressure prediction (Parikh et al., 2008; Sun et al., 2017). For these reasons, we decide not to include alcohol consumption as an explanatory variable in our prediction models. There are some missing values in some of the other variables as well, but small amounts relative to the total number of observations. Therefore, we choose to include these variables, and remove the participants with missing values in the explanatory variables listed above (excluding alcohol). We are left with 17 733 participants.

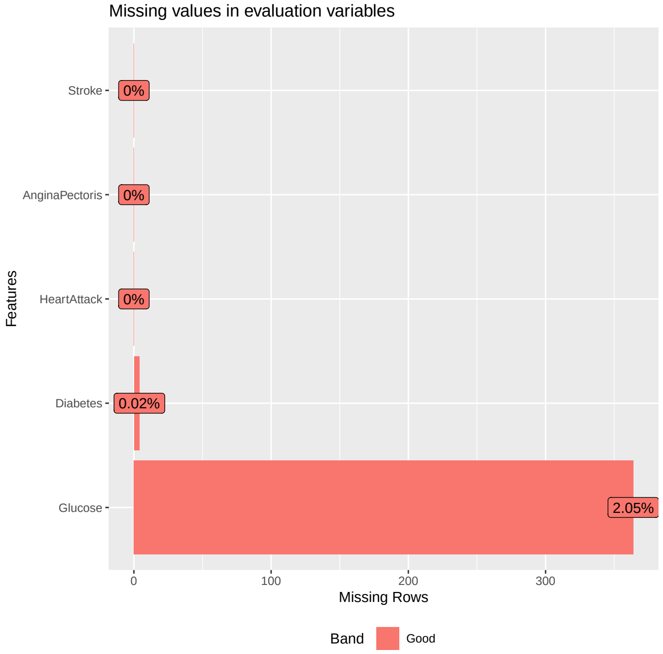


Figure 2.4: Missing values in variables on health regarding blood pressure at the time of HUNT3. The percentage of missing values is shown for each evaluation variable. All the variables are from HUNT3.

2.1.4 Health during HUNT3

We take a closer look at some variables that contain information about the blood pressure-related health of the participants at the time of HUNT3. The reason for this is that we want to know who is on blood pressure medication during HUNT3 such that we can correct the effect the blood pressure medication has on the measured blood pressure. In addition to this, we want to know who has a history of diabetes or CVD at the time of HUNT3. The reason for this is that these illnesses are associated with hypertension (Chobanian et al., 2003), and we want to have the opportunity to check how our prediction models perform on these subgroups.

We have already removed the participants with missing information about their use of blood pressure medication at the time of HUNT3 when we selected the population. According to Paz et al. (2016) it is reasonable to add 15 mmHg to the mean systolic blood pressure and 10 mmHg to the mean diastolic blood pressure to get a blood pressure value similar to what it would have been if the participant had not been using blood pressure medication.

In Figure 2.4 we see that there are no missing values in the history of CVD, and relatively few missing values in the history of diabetes, at HUNT3. Since there are so few participants with missing values, we conclude that it is worth removing these participants such that it is easier if we choose to evaluate the performance of the prediction models on these subgroups later on.

After removing the participants with missing values of CVD and diabetes at the time of HUNT3, we are left with 17 365 participants.

2.2 Exploratory data analysis

2.2.1 Response variable

We start our exploratory data analysis by looking at the blood pressure measurements from HUNT3, see Figure 2.5. It is clear that both the diastolic and systolic blood pressure seems to approximately follow a Gaussian distribution. The systolic blood pressure has a somewhat heavier right tail than the diastolic blood pressure. This might be because the diastolic blood pressure tends to decrease with age after one turns 60 years, while the systolic blood pressure tends to increase linearly with age (Franklin et al., 1997). Another interesting observation is that even though we excluded all the people who were hypertensive at the time of HUNT2, there is a relatively large portion of the participants who are hypertensive, ie. above the red line, at the time of HUNT3. We see this more clearly in Figure 2.6, where we observe that approximately 20% of the participants are systolic hypertensive at the time of HUNT3. Our criteria for systolic hypertension is that the mean systolic blood pressure, of measurements taken on two separate days, is above 140 mmHg.

Since we are using the systolic blood pressure from HUNT3 as the response variable in our prediction model, we want to examine it in more detail. From Figure 2.5 we know that the distribution looks approximately Gaussian with a heavier right tail. This is examined further in Figure 2.7, where it is clear that the systolic blood pressure has a lighter left tail and a heavier right tail than a normal distribution. However, it is not very far from a Gaussian distribution.

We move on to check if the correction of the blood pressure measurements from people using blood pressure medication, details in Section 2.1.4., is reasonable. In Figure 2.8 both the corrected and the uncorrected systolic blood pressure from HUNT3 is shown. We see that without correction the mean of the blood pressure of participants using blood pressure medication, marked by the blue line, is just slightly higher than the mean of the blood

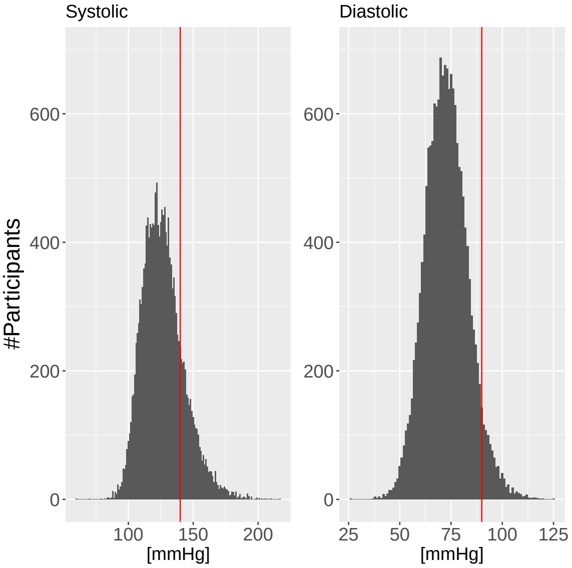


Figure 2.5: The systolic and diastolic blood pressure of the participants at HUNT3. The red line is marking the hypertension threshold, and is thus at 140 mmHg for the systolic blood pressure and 90 mmHg for the diastolic blood pressure.

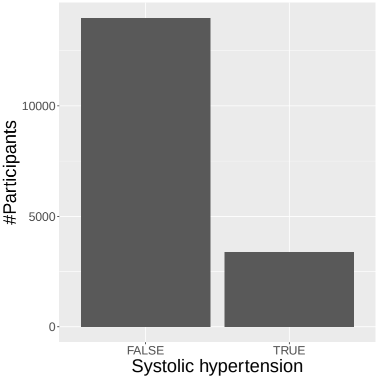


Figure 2.6: The number of participants with systolic hypertension at HUNT3. TRUE indicates systolic blood pressure $\geq$ 140 mmHg, and FALSE indicates systolic blood pressure $<$ 140 mmHg. Approximately 20% of the participants had systolic hypertension at the time of HUNT3.

pressure of the participants not using blood pressure medication, marked by the black line. The corrected systolic blood pressure values have a higher mean, yet the total distribution of the systolic blood pressure still seems reasonable. The distribution is still approximately Gaussian, and there are no big outliers nor multiple peaks. Since the correction seems reasonable, we use the corrected version of the systolic blood pressure from now on.

Before we move on to explore the explanatory variables, it is interesting to examine

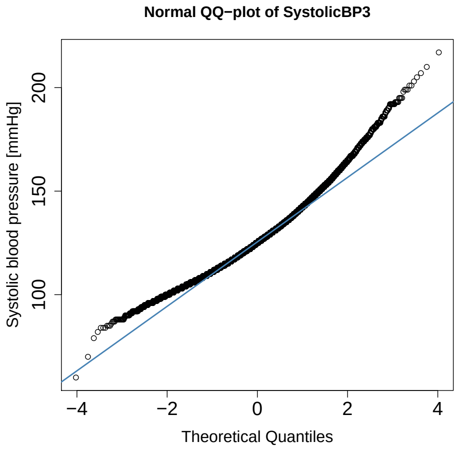


Figure 2.7: A normal QQ-plot of the systolic blood pressure of the participants at HUNT3.

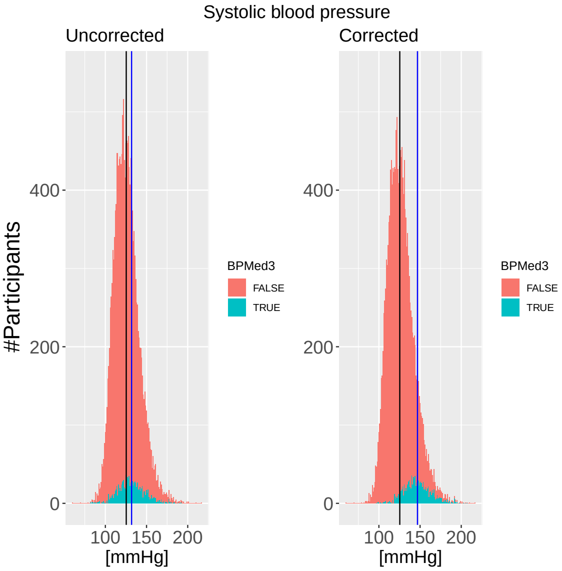


Figure 2.8: The plot to the left shows the uncorrected systolic blood pressure measurements from HUNT3. The plot to the right shows the systolic blood pressure at HUNT3 where the effect of blood pressure medication has been corrected. The participants who were currently on blood pressure medication at the time of HUNT3 are presented in turquoise, and other participants are presented in light red. The blue and black lines are marking the mean of the systolic blood pressure of the participants, respectively, using and not using blood pressure medication at HUNT3.

the systolic blood pressure of the participants with diabetes or CVD at the time of HUNT3, ie. the illnesses we excluded in HUNT2. In Figure 2.9 we see that for both diabetes and

CVD the mean of the systolic blood pressure of the affected participants is slightly higher than the mean of the non-affected participants.

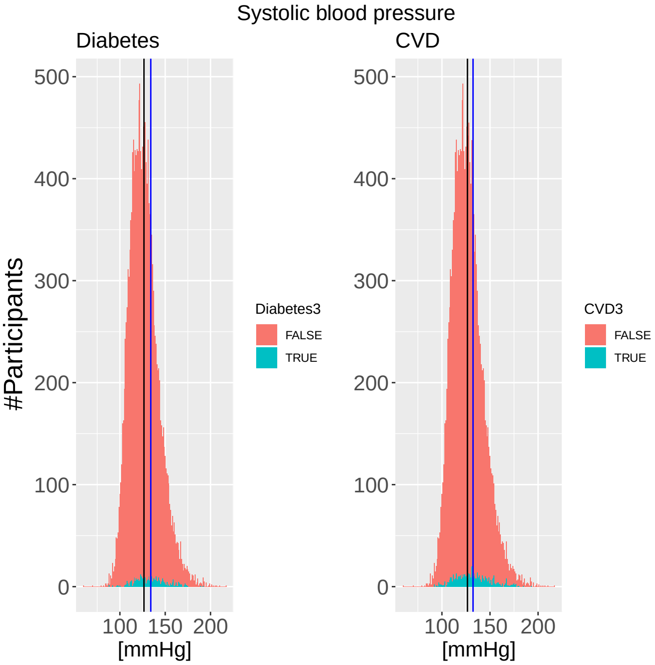


Figure 2.9: These plots show the systolic blood pressure of the participants at HUNT3. In the left plot, the participants with diabetes at the time of HUNT3 are presented in turquoise, and the other participants are presented in light red. In the right plot, the equivalent is true for participants with CVD. The blue and black lines in the left plot are marking the mean of the systolic blood pressure of the participants who are, respectively, suffering from diabetes or not suffering from diabetes at HUNT3. The same goes for the lines in the right plot, but for participants with CVD.

2.2.2 Explanatory variables

In this section, we want to examine the explanatory variables, and see how they are correlated with each other and the response. A figure showing the correlation between all the explanatory variables, both numerical and categorical, and the response can be found in the Appendix. To get a quick overview, we start by looking at the correlation between all the continuous variables and the continuous response, see Figure 2.10. Not surprisingly, the systolic blood pressure from HUNT2 has the highest positive correlation with the systolic blood pressure from HUNT3, followed closely by the diastolic blood pressure from HUNT2. More interestingly, we see a noticeable negative correlation between the response, ie. the systolic blood pressure from HUNT3, and birth year, and a somewhat smaller positive correlation between the response and both BMI and cholesterol. We examine these relations in more detail later in this section. Another interesting observation is the negative correlation between birthyear and cholesterol. We divide the explanatory variables into four categories: Basic information, Blood pressure, Lifestyle, and Blood samples, and start examining the explanatory variables in the Basic information category.

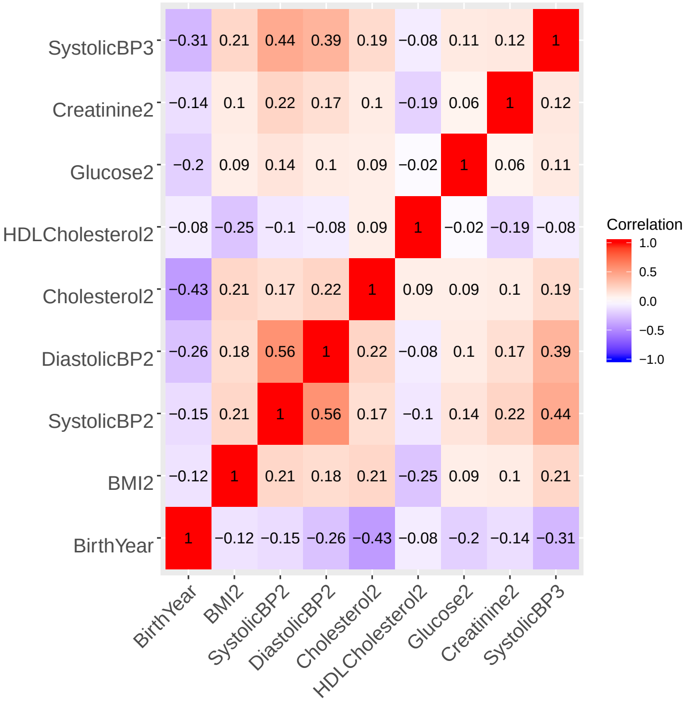
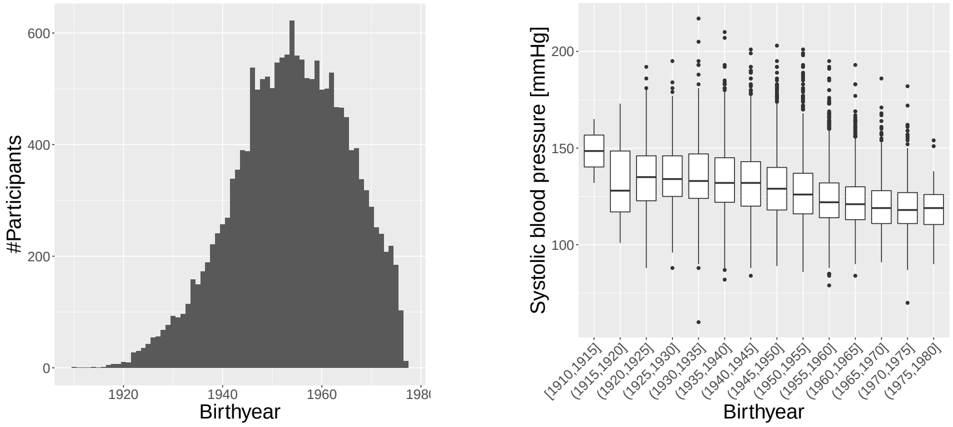


Figure 2.10: The correlation between all the continuous explanatory variables from HUNT2 and the continuous response, ie. the systolic blood pressure from HUNT3. The number at the end of the variable name indicates which HUNT survey the variable belongs to.

Basic information

In Figure 2.11 a) we see the distribution of the participants' birth year. The oldest participant was born in 1910, ie. 98 years old at the end of the HUNT3 survey. In contrast, the youngest participant was born in 1977 and turned 20 during the last year of the HUNT2 survey. This was the cut-off for being allowed to participate in HUNT2, and we see this cut-off clearly in Figure 2.11 a). Both mean and median birth year is 1954, and thus the median age of the participants is approximately 42 years during HUNT2 and 53 years during HUNT3. This is slightly higher than the median age of the general Norwegian population, which has a median age of 39.8 years, according to Worldometer (2020). This is expected due to the minimum age limit for HUNT2. If we don't consider the cut-off, the values seem to follow an approximately Gaussian distribution, with a small spike at the median 1954.

The relationship between the systolic blood pressure from HUNT3 and birth year is presented in Figure 2.11 b). In general, there seems to be a negative correlation, which



(a) **(b)**

Figure 2.11: a) The distribution of the birthyear of the participants. (b) The systolic blood pressure from HUNT3 versus the birthyear of the participants.

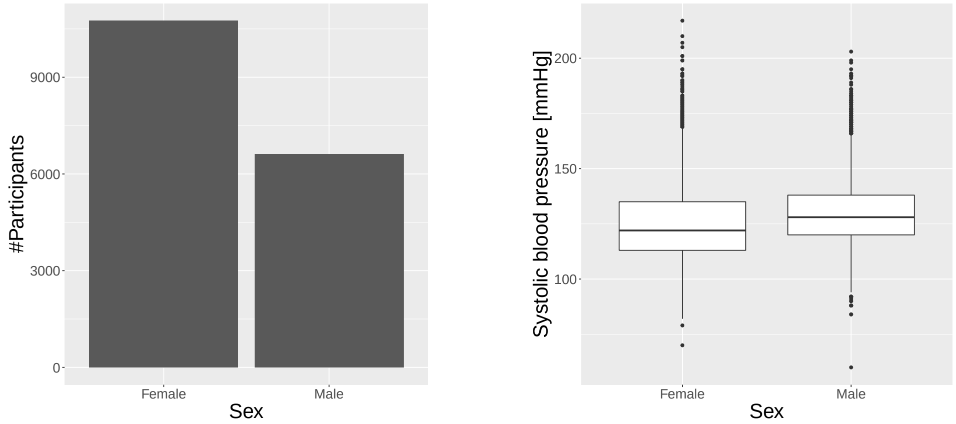
coincides with Figure 2.10 where the correlation is shown to be -0.31. It is not a completely linear relationship, which is particularly noticeable at the extremes of the birth year range. On the other hand, there are many fewer participants in the youngest and oldest age groups, which might explain why these age groups deviate from the trend. The variance in systolic blood pressure seems to increase slightly with age, ie. decrease with birth year. One could speculate that this increase in variation is because lifestyle choices accumulate with the years and contribute to either a stable systolic blood pressure stable or to an increase in systolic blood pressure.

We move on to study the amount of female versus male participants, and the relationship between sex and systolic blood pressure. From Figure 2.12 a) it is clear that there are significantly more women than men among our participants. In fact, 61.93% of the participants are female, while only 49.61% of the general Norwegian population is female, according to Statistics Norway (2018).

In Figure 2.12 b) we see the distribution of the systolic blood pressure from HUNT3 for each sex. Males have a clearly higher median systolic blood pressure than females. Another interesting observation is that it seems like the females have a larger variation in systolic blood pressure, but this might be due to the higher number of female participants.

Blood pressure

In this section, we explore the explanatory variables related to blood pressure. The distributions of the systolic and diastolic blood pressure from HUNT2 seem to be approximately Gaussian when disregarding the hypertension cut off, and can be found in Figure 6.2 in the Appendix. We already know from Figure 2.10 that systolic and diastolic blood pressure from HUNT2 are correlated with the systolic blood pressure from HUNT3. This can also be seen more clearly in Figure 2.13. From this figure, we also observe that there is a



(a)

(b)

Figure 2.12: a) Sex of the participants (b) Systolic blood pressure from HUNT3 versus sex of the participants

higher correlation between blood pressure measurements from the same survey, than the systolic blood pressure measurements from different surveys. The multicollinearity caused by the correlation between systolic and diastolic blood pressure from HUNT2 might lead to challenges if we include both as explanatory variables in the prediction models.

We are also interested in the hypertension status of the parents of the participants. In Figure 2.14 a) we see that approximately 32% of the participants have at least one parent with a history of hypertension. Parikh et al. (2008) concluded that parental hypertension is a significant predictor of hypertension. In contrast, in Figure 2.14 b) parental hypertension seems to have a very small effect on the systolic hypertension of the participant. The participants with at least one hypertensive parent have a very slightly higher median systolic blood pressure than the other group, but this is almost not noticeable. One could speculate that the reason for the small effect observed here is that we have excluded all the participants who were hypertensive at the time of HUNT2 and that we would have seen a bigger effect if these participants had been included.

Lifestyle

It is interesting to study the variables related to the lifestyle of the participants at the time of HUNT2, and see how these relate to the systolic blood pressure at the time of HUNT3. The underlying assumption in including variables describing the lifestyle of the participants approximately 11 years before the time of the predictions is that the lifestyle of the participants is fairly constant.

We start by examining the BMI of the participants at the time of HUNT2, see Figure 2.15 a). The distribution seems approximately Gaussian, but with a slightly heavier right tail and some outliers of BMI as high as 52.8 kg/m^2. The most common value is 23.8 kg/m^2, and the mean is 25.3 kg/m^2. We observed a positive correlation of 0.21 be-

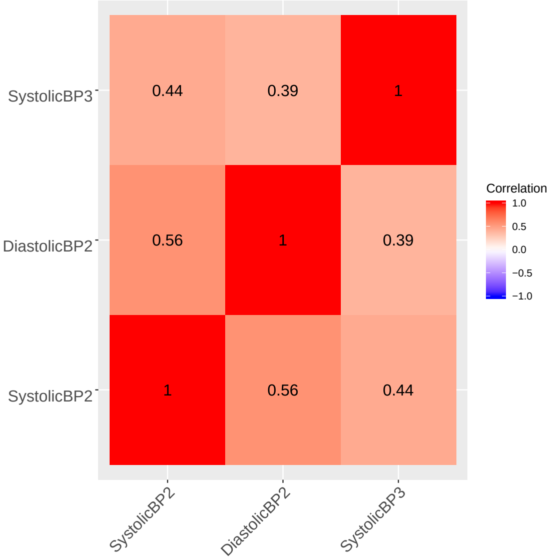
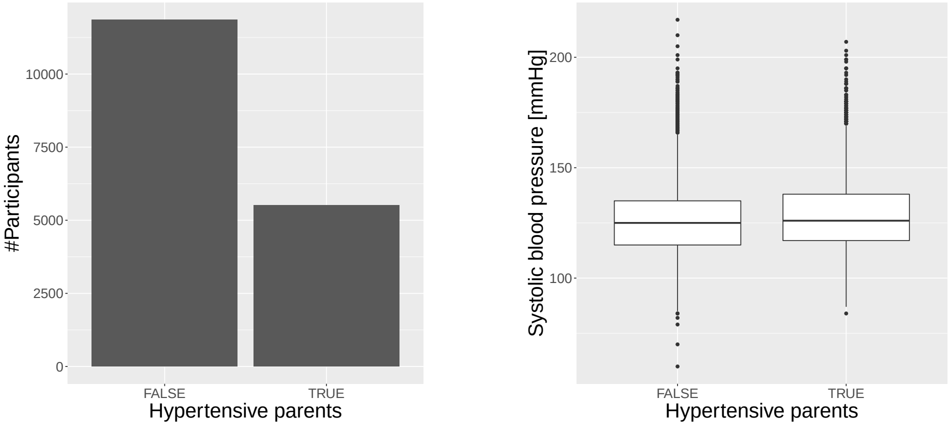


Figure 2.13: Correlation between systolic and diastolic blood pressure measurements from HUNT2 and systolic blood pressure measurements from HUNT3. The number at the end of the variable name indicates which HUNT survey the variable belongs to.

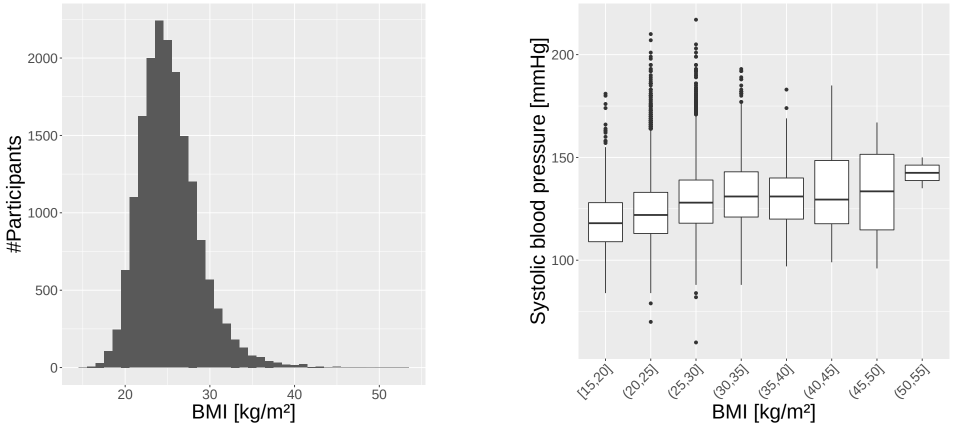


(a)

(b)

Figure 2.14: a) The distribution of participants with at least one parent with a history of hypertension at the time of HUNT2 ; (b) Systolic blood pressure from HUNT3 versus a parental history of hypertension from HUNT2.

tween BMI from HUNT2 and the systolic blood pressure from HUNT3 in Figure 2.10. We examine this relationship further in Figure 2.15 b) and notice that overall the median systolic blood pressure increases with the BMI. It seems to increase linearly for BMI from 15 to 35 kg/m^2, which includes most of the participants. The change in median systolic blood pressure flattens out between 35 kg/m^2 to 50 kg/m^2, and then increases again for the (50,55] kg/m^2 group. This might indicate a generally nonlinear relationship between systolic blood pressure and BMI. However, there are many fewer participants in the upper BMI range, and the variance seems to be bigger, so it is hard to say for certain. Another interesting observation is that the participants with the highest systolic blood pressure have a BMI between 20 and 30 kg/m^2, which is within the normal and overweight range (Ardern et al., 2004). This might just be because the majority of participants lie within this range.



(a) **(b)**

Figure 2.15: a) The distribution of BMI of the participants at the time of HUNT2 (b) Systolic blood pressure from HUNT3 versus BMI from HUNT 2.

In this thesis we are particularly interested in the effect physical activity might have on hypertension. We use two variables to measure the physical activity of the participants at the time of HUNT2, `PAI` and `RecPa`. See the Appendix for their distributions among the participants in Figure 6.3. The relationship between these two variables and the systolic blood pressure at the time of HUNT3 are shown in Figure 2.16. There seems to be a surprisingly small effect of a higher level of physical activity. Participants with a high PAI score have a slightly lower median blood pressure than participants with low PAI score, but the effect is barely noticeable. The same goes for participants with a physical activity level above the recommended level versus participants below the recommended level. One interesting observation is that the few participants with the lowest systolic blood pressure all have a high PAI score or physical activity above the recommended level.

The last variables we consider concerning lifestyle are daily smoking habits and highest education level achieved at the time of HUNT2. The distribution of these variables can be found in Figure 6.4 in the Appendix. There seems to be a correlation between daily smoking habits and systolic blood pressure, and between education level and systolic blood pressure, in Figure 2.17. However, since both education and smoking are lifestyle aspects that have changed a lot during the last decades, we check the correlation

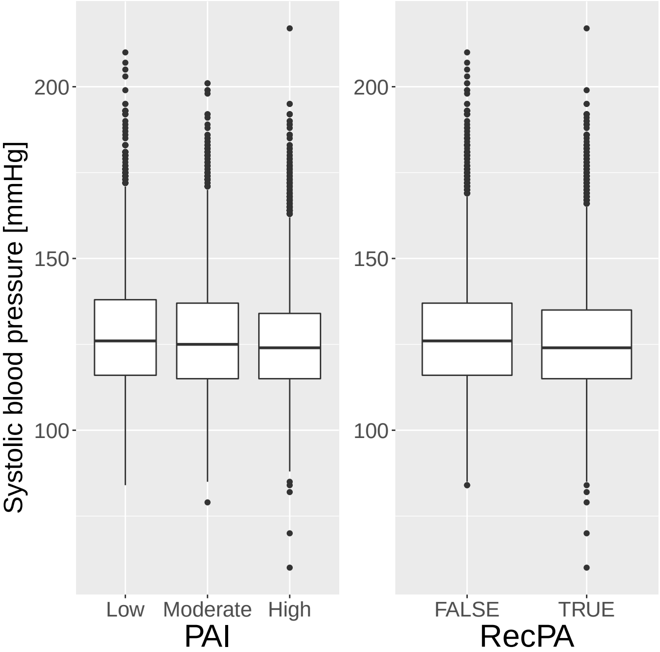


Figure 2.16: Systolic blood pressure from HUNT3 versus: PAI from HUNT2 (left); RecPA from HUNT2 (right).

between these variables and birthyear in Figure 2.18. In this figure, we observe a stronger and reversed version of the correlations in Figure 2.17. This indicates that the correlations between daily smoking habits and systolic blood pressure, and education level and systolic blood pressure, might be caused mostly by the negative correlation between birthyear and systolic blood pressure explored earlier in this chapter.

Blood samples

The last category of explanatory variables is the values found from blood samples taken from the participants during HUNT2. The distribution of the continuous blood sample variables HDL cholesterol, cholesterol, creatinine, and non-fasting glucose can be found in Figure 6.5 in the Appendix. All four variables seem approximately Gaussian distributed, but with slightly heavier right tails.

In Figure 2.19, the correlations between the response and the continuous blood sample explanatory variables are shown. Cholesterol is, as noted previously, somewhat positively correlated with the response, but it isn't a strong correlation. Another interesting observation is the equally large negative correlation between HDL cholesterol and creatinine. Other than this, there are no particularly strong positive or negative correlations.

The distribution of the categorical variable describing the GFR stage of the participants during HUNT2 is shown in Figure 2.20 a). It is clear that the number of participants within

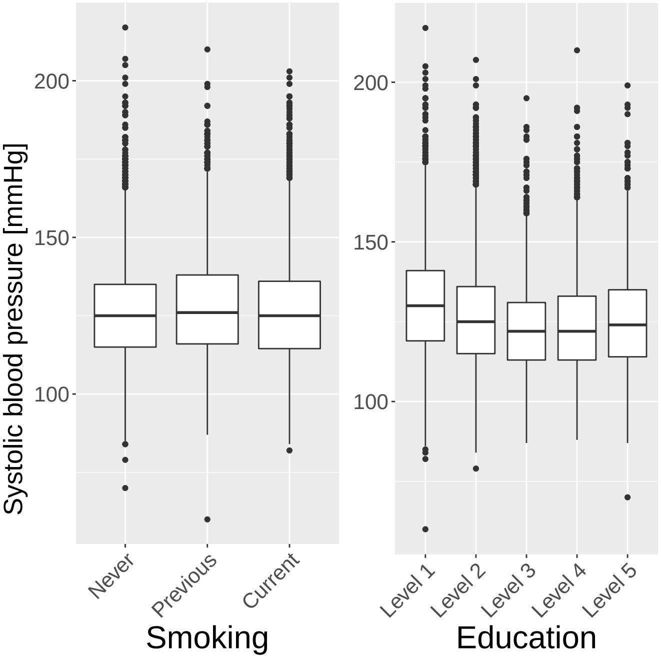


Figure 2.17: Systolic blood pressure from HUNT3 versus: Smoking at HUNT 2 (left); Education from HUNT2 (right). The levels of Smoking and Education are described in further detail in Section 2.3.

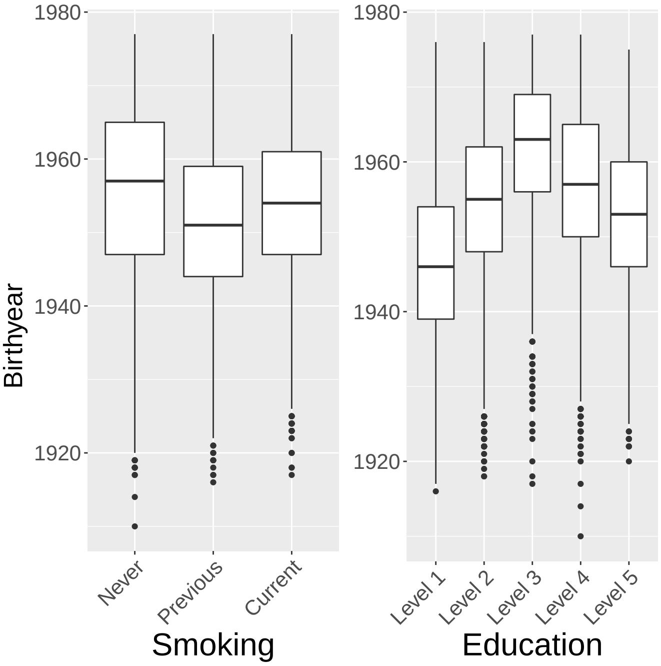


Figure 2.18: Birthyear versus: Smoking at HUNT 2 (left); Education from HUNT2 (right). The levels of Smoking and Education are described in further detail in Section 2.3.

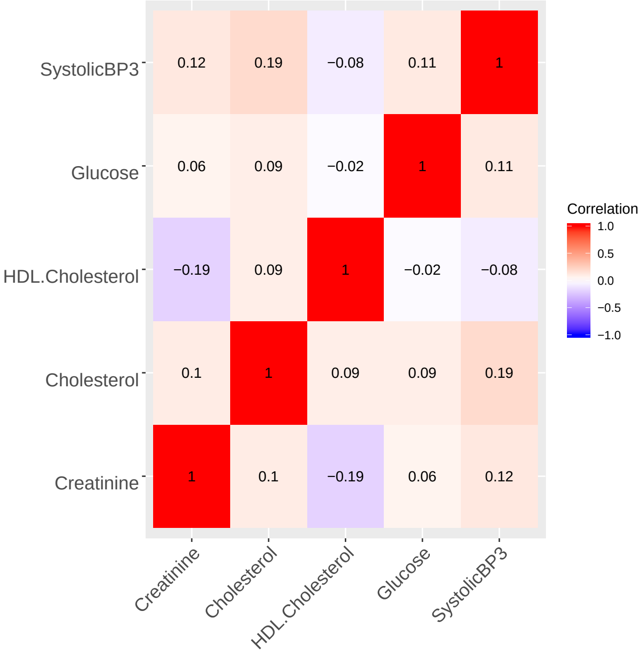
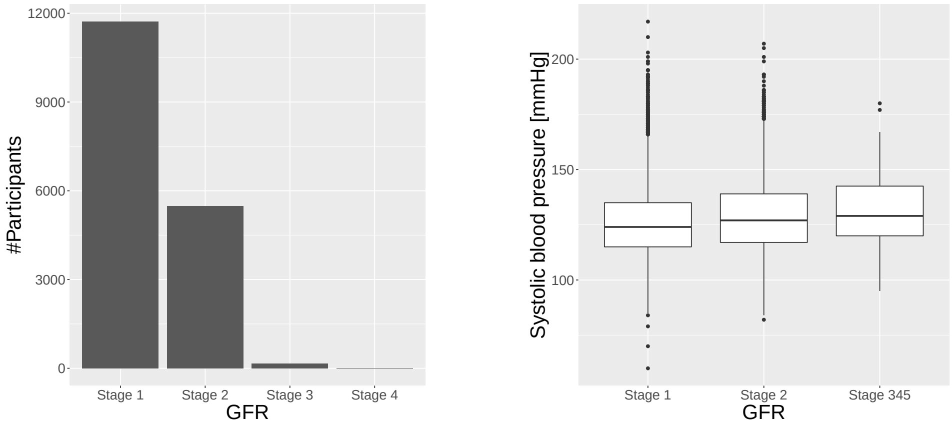


Figure 2.19: Correlation between the continuous explanatory variables from blood samples in HUNT 2, ie. Glucose, HDL Cholesterol, Cholesterol, and Creatinine, and systolic blood pressure from HUNT3.

each stage decreases as the GFR decreases (ie. as stage number increases). In fact, there are few participants with GFR in Stage 3, only one participant with GFR in Stage 4, and no participants with GFR in Stage 5. This makes GFR stand out since all categories in the other categorical explanatory variables have been well represented. Due to this, we put all the participants with GFR in Stage 3, Stage 4, and Stage 5 together in a new category called Stage 345.

There seems to be a small, but noticeable, positive correlation between systolic blood pressure from HUNT3 and GFR stage from HUNT2, see Figure 2.20 b). The median systolic blood pressure is increasing from Stage 1 to Stage 345. It is interesting to note that the participants with the highest systolic blood pressure have GFR in Stage 1 and Stage 2, but this is probably due to the fact that there are so many more participants with GFR in these stages than with GFR in Stage 345.



(a)

(b)

Figure 2.20: a) Distribution of participants among the different stages of GFR at time of HUNT2. (b) Systolic blood pressure from HUNT3 versus stages of GFR. Note that there are 5 possible stages of GFR, but the fifth stage is not represented among the participants, is therefore not shown in these plots.

2.3 Data transformation

Before using the data to fit the prediction models, we standardize the continuous explanatory variables by subtracting the mean of the variables and diving by the standard deviation of the variables. We do this for easier interpretation of the coefficients in the prediction models, and easier comparison of the importance of the continuous explanatory variables.

Chapter 3

Statistical Framework

3.1 Generalized Linear Models

The presentation of generalized linear models in this section is based on Fahrmeir et al. (2013). Generalized linear models are a generalization of the classical linear model. In the classical linear model, the response $\mathbf{y}$ is assumed to be continuous and follow an approximately Gaussian distribution, and have an expected value that can be written as a linear combination of the explanatory variables. In a generalized linear model, there are less restrictive assumptions, and as a consequence, generalized linear models can be used to model binary, continuous, categorical, or count data responses. We assume n observations and k explanatory variables, such that we have a $n \times 1$ response vector $\mathbf{Y} = [y_1, ..., y_n]^T$, and a $n \times p$ design matrix $\mathbf{X}$, with elements $[\mathbf{X}]_{\mathbf{ij}} = x_{ij}$. The main assumptions of a generalized linear model can be divided into distributional and structural assumptions.

Distributional assumptions

Consider the covariates $\mathbf{x_i} = [1, x_{i1}, ..., x_{ik}]^T$, where $p = k + 1$. Given these covariates, the response variables are independent and the density of the response variable y_i belongs to a univariate exponential family with

$$f(y_i|\theta_i) = \text{exp}\bigg(\frac{y_i\theta_i - b(\theta_i)}{\phi} w_i + c(y_i, \phi, w_i)\bigg), \tag{3.1}$$

where θ_i is the natural parameter, ϕ is the dispersion parameter, w_i is a weight function which equals 1 for ungrouped data, and b and c are known functions. The mean and variance of a univariate exponential family are given by

$$\text{E}(y_i) = \mu_i = b'(\theta_i), \quad \text{Var}(y_i) = \sigma_i^2 = \frac{\phi}{w_i} b''(\theta_i).$$

Structural assumptions

The linear predictor η_i is defined as

$$\eta_i = \mathbf{x_i}^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}.$$

The conditional mean μ_i of the response variable y_i can be found by setting the linear predictor as input to the response function h,

$$\mu_i = h(\eta_i) = h(\mathbf{x_i}^T \boldsymbol{\beta}).$$

Inversely, the linear predictor can be found from the conditional mean through the link function g

$$\eta_i = g(\mu_i).$$

We assume that the response function h is one-to-one and twice differentiable and that the link function g is the inverse of the response function $g = h^{-1}$.

Some examples of distributions which belong to the univariate exponential family are the Gaussian, gamma, Poisson, and binomial distributions.

Inference

Given the observed data $(y_i, \mathbf{x_i})$, $i = 1, ..., n$, the maximum likelihood estimate of the regression coefficients $\hat{\boldsymbol{\beta}}_{\text{ML}}$ is found by maximizing the likelihood function $L(\boldsymbol{\beta})$. From the distributional assumptions, it is known that the response variables y_i, $i = 1, ..., n$, are conditionally independent and thus the likelihood function can be written as the product of the likelihood of the individual observations y_i, $i = 1, ...n$,

$$L(\boldsymbol{\beta}) = f(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta}) \tag{3.2}$$

Since the natural log function is monotonically increasing, the estimate $\hat{\boldsymbol{\beta}}_{\text{ML}}$ that maximizes the the log of the likelihood function $l(\boldsymbol{\beta})$ also maximizes the likelihood function $L(\boldsymbol{\beta})$. The maximum likelihood estimate of the regression coefficients $\hat{\boldsymbol{\beta}}_{\text{ML}}$ is thus defined as the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ that maximize the log-likelihood function $l(\boldsymbol{\beta})$, where

$$l(\boldsymbol{\beta}) = ln(L(\boldsymbol{\beta})) = ln(\prod_{i=1}^{n} f(y_i|\boldsymbol{\beta})) = \sum_{i=1}^{n} ln(f(y_i|\boldsymbol{\beta})) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}). \tag{3.3}$$

The maximum likelihood estimate of the regression coefficients $\hat{\boldsymbol{\beta}}_{\text{ML}}$ is consequently found by solving the equation

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} s_i(\boldsymbol{\beta}) = \mathbf{s}(\boldsymbol{\beta}) = 0 \tag{3.4}$$

where the derivative of the log-likelihood function $l(\cdot)$ with regards to $\boldsymbol{\beta}$ is called the score function $s(\boldsymbol{\beta})$.

To solve the equations $\mathbf{s}(\hat{\boldsymbol{\beta}}) = 0$, it is common to use an iterative algorithm. One such algorithm is the iteratively reweighted least squares algorithm, often called IRWLS, see Fahrmeir et al. (2013) for details.

It can be shown that as the total sample size n goes to infinity, $n \to \infty$, the distribution of the maximum likelihood estimator of the regression coefficients $\hat{\boldsymbol{\beta}}_{\text{ML}}$ goes towards a p-variate Gaussian distribution with the true regression coefficients $\boldsymbol{\beta}$ as expected value and the inverse Fisher matrix, evaluated at the maximum likelihood estimate, as the estimated covariance matrix,

$$\hat{\boldsymbol{\beta}}_{\text{ML}} \approx \mathcal{N}_p(\boldsymbol{\beta}, \mathbf{F}^{-1}(\hat{\boldsymbol{\beta}}_{\text{ML}})). \tag{3.5}$$

The Fisher matrix $\mathbf{F}(\boldsymbol{\beta})$ is defined as

$$\mathbf{F}(\boldsymbol{\beta}) = \mathrm{E}\Big(- \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \Big). \tag{3.6}$$

3.2 Root mean square error

Root mean square error, hereafter denoted RMSE, is a commonly used measure of the distance between predicted or fitted values $\hat{\mathbf{y}} = [\hat{y}_1, ...\hat{y}_n]$ and the observed values $\mathbf{y} = [y_1, ..., y_n]$. In other words, it can be used to describe the goodness-of-fit of a regression model or prediction model. It is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(\hat{y}_i - y_i)^2]} \tag{3.7}$$

3.3 Brier score

The Brier score is a method for evaluating the accuracy of categorical probabilistic prediction models. It was proposed by Brier (1950), and according to Hersbach (2000) it is one of the oldest methods still in use for evaluating the accuracy of probabilistic models.

Consider a probabilistic prediction model with variable of interest Y, where Y can belong to one of r mutually exclusive categories. The probability that the i-th observation of Y belongs to category j is denoted by f_{ij}, where $i = 1, 2, ..., n$ and $j = 1, ..., r$. These probabilities must sum to 1 for each observation i,

$$\sum_{j=1}^{r} f_{ij} = 1, \quad \text{for } i = 1, 2, ..., n. \tag{3.8}$$

The Brier score (BS) of the prediction model is defined by Brier (1950) as

$$BS = \frac{1}{n}\sum_{j=1}^{r}\sum_{i=1}^{n}(f_{ij} - o_{ij})^2, \tag{3.9}$$

where o_{ij} is a binary variable which is either 0 or 1. It is 1 if the i-the observation of Y belonged to category j, and 0 if not. From the definition, it is clear that the Brier score (BS) has the range $(0, 2)$, where a Brier score of 0 can only be achieved by a model which predicts 100% probability for the correct category for all observations $i = 1, 2, ..., n$.

For binary prediction models, for example if the random variable Y describes whether an event occurs or not, a slightly altered version of the Brier score is often used. An example is Hersbach (2000), where they define a version of the Brier score (BS^*) as half the Brier score from the original paper by Brier (1950). This altered version BS^* can be formulated as

$$BS^* = \frac{1}{2n} \sum_{j=1}^{2} \sum_{i=1}^{n} (f_{ij} - o_{ij})^2 = \frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2, \tag{3.10}$$

where p_i is the probability of the event occurring at the i-th observation, and $o_i \in \{0, 1\}$ is a binary variable indicating whether the event actually occurred at the i-th observation. With this new formulation, the Brier score has the range $(0, 1)$. A Brier score of 0 indicates a perfect prediction model, and a Brier score of 1 indicates the worst possible prediction model.

Typically, in binary situations where BS^* is used, the event of interest is whether the i-th observation y_i of the random variable Y is below a given threshold value y_t. In other words, if $y_i \leq y_t$ then the event has occurred and $o_i = 1$, otherwise the event hasn't occurred and $o_i = 0$ (Hersbach, 2000).

From this point on in this thesis, we use the alternative definition BS^* of the Brier score.

3.4 Continuous rank probability score

The presentation of the continuous rank probability score in this section is based on the presentation in the article by Hersbach (2000), which is again based on the work of Brown (1974); Unger (1985); Matheson and Winkler (1976); Bouttier (1994). The continuous rank probability score, hereafter denoted CRPS, is a method for evaluating and comparing the accuracy of probabilistic forecast models. It can be viewed as a generalization of the mean absolute error, and for a deterministic forecast it is, in fact, equivalent to the mean absolute error. We consider a probabilistic forecast of the random variable Y, with cumulative distribution function $F(Y)$, and we denote the true observation by y. The CRPS is defined as

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} [F(x) - H(y - x)]^2 dx, \tag{3.11}$$

where CRPS has the same unit as the random variable Y, due to the multiplication with dx, and $H(z)$ is the Heavyside step function

$$H(z) = \begin{cases} 0 & \text{for } z < 0 \\ 1 & \text{for } z \geq 0. \end{cases} \tag{3.12}$$

There are several ways to interpret the CRPS. According to Hersbach (2000) it can be considered as a measurement of the distance between the cumulative distribution function of the probabilistic forecast and the empirical cumulative distribution function. This means that a well-calibrated probabilistic model has low CRPS. On the other hand, CRPS can also be understood through its relation to the Brier score, see the previous section for further details on the Brier score. If we consider all real threshold values, $y_t \in \mathbb{R}$, and take the integral over all the Brier scores of these thresholds, we get the CRPS.

$$\text{CRPS} = \int_{-\infty}^{\infty} BS^*(y_t)dy_t \tag{3.13}$$

3.5 Probability Integral Transformation

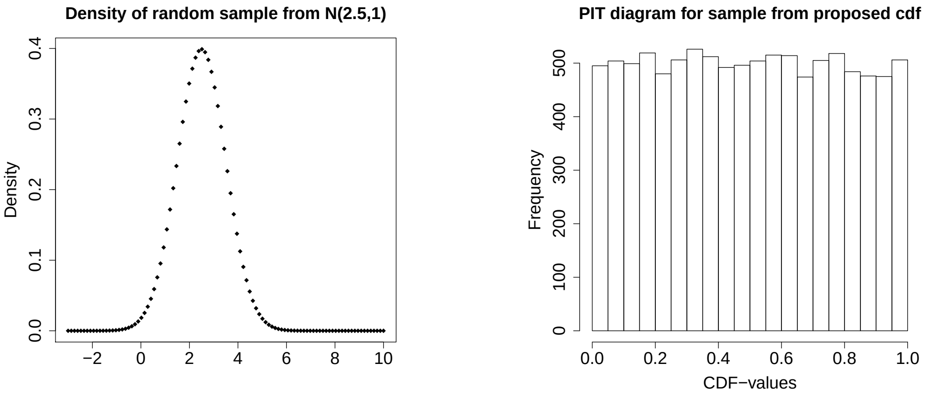
Angus (1994) presents and gives a proof of the Probability Integral Transformation theorem. We use slightly different notation than Angus (1994) for the sake of consistent notation in this thesis. The Probability Integral Transformation theorem states that if the cumulative density function F($\cdot$) of the real-valued random variable Y is continuous, then the random variable $Z = F(Y)$ is uniformly distributed on the interval (0,1).

If we have a random sample $y_1, ..., y_n$ and we don't know the true cumulative distribution function $F_t(\cdot)$ of this sample, we can use the Probability Integral Transformation theorem to check the goodness-of-fit of a proposed and known cumulative distribution function $F_k(\cdot)$. A way to visualize this is Probability Integral Transformation diagrams, hereby denoted PIT diagrams.

These diagrams are created by plotting a histogram of the proposed cumulative distribution functions applied to the observed values, $\{F_k(y_1), ..., F_k(y_n)\}$. If the sample $y_1, ..., y_n$ is truly from the cumulative distribution function $F_k(\cdot)$, then the set $\{F_k(y_1), ..., F_k(y_n)\}$ is uniformly distributed on the interval (0,1), and the bars in the histogram should be approximately the same height. This is illustrated in Figure 3.1. The density of a random sample of size 10 000 from a Gaussian distribution with mean 2.5 and standard deviation 1 is plotted in Figure 3.1 a), and in Figure 3.1 b) we see the cdf-values of applying the true cumulative distribution function on the random sample. It is clear that the bars have approximately equal height, which means that the cdf-values are approximately uniformly distributed.

If the true density function $F_t'(\cdot)$ of the sample has a heavier right tail than the proposed density function $F_k'(\cdot)$, then the bars closer to 1 are higher than the bars closer to 0. The reason for this is that the points from the sample that lie in the heavy right tail have higher values than expected by the proposed cdf-function $F_k'(\cdot)$. This is illustrated in Figure 3.2, where the random sample comes from a Gamma distribution with a heavy right tail, while the proposed distribution is the same symmetrical Gaussian distribution as shown in Figure 3.1 a). For heavy left tails, we would see the opposite trend where bars closer to 0 is much higher.

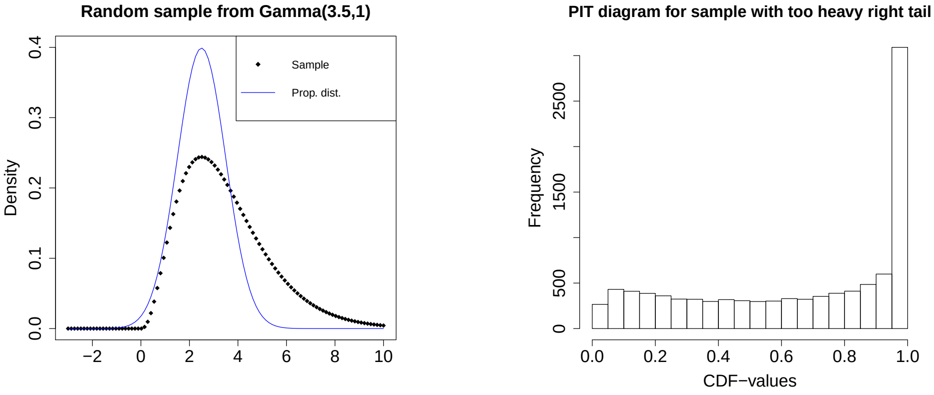
Another scenario is if the proposed density function $F_k'(\cdot)$ has bigger variation than the true density function $F_t'(\cdot)$. This can be seen in the PIT diagrams as shorter bars towards both ends of the interval (0,1), see Figure 3.3 b).



(a)

(b)

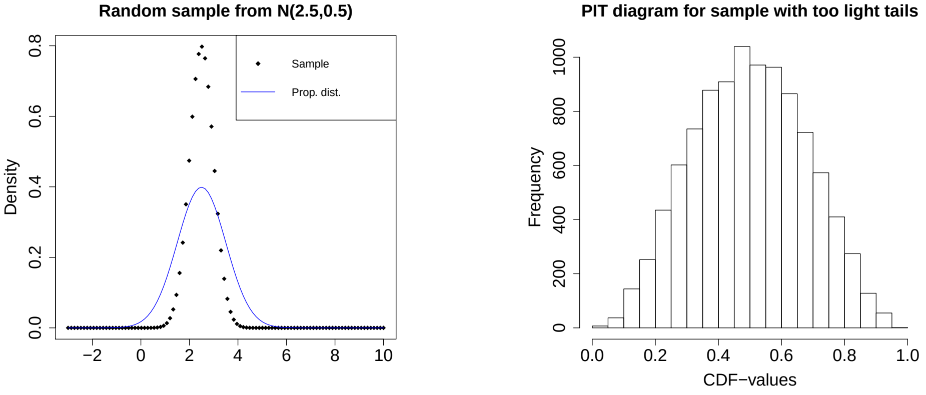
Figure 3.1: a) The distribution of random sample of size 10 000 from a Gaussian distribution $\mathcal{N}(2.5, 1)$; (b) PIT Diagram where the proposed cumulative distribution function is the true cumulative distribution function of the random sample.



(a)

(b)

Figure 3.2: a) The distribution of a random sample of size 10 000 from a Gamma distribution with shape parameter $k = 3.5$ and scale parameter $\theta = 1$, $\Gamma(3.5, 1)$ shown as dots. The proposed distribution, the Gaussian distribution $\mathcal{N}(2.5, 1)$, is shown as a blue line; (b) PIT Diagram where the proposed cumulative distribution function is the cumulative distribution of $\mathcal{N}(2.5, 1)$, while sample in reality is sampled from $\Gamma(3.5, 1)$.



(a) **(b)**

Figure 3.3: a) The distribution of random sample of size 10 000 from the Gaussian distribution $\mathcal{N}(2.5, 0.5^2)$. The proposed distribution, the Gaussian distribution $\mathcal{N}(2.5, 1)$, is shown as a blue line; (b) PIT Diagram where the proposed density function is the Gaussian distribution $\mathcal{N}(2.5, 1)$, while the true density function is the Gaussian distribution $\mathcal{N}(2.5, 0.5^2)$.

3.6 Sensitivity and specificity

Binary classification tests try to correctly classify observations into one of two categories. There are many examples of such tests. An example from medicine are tests where the goal is to figure out whether the patient has a certain illness or not. In other words, if the patient should be classified as ill or not. To measure how well such tests are able to classify the observations, it is common to use sensitivity and specificity. Simply put, sensitivity is a measure of how well the test is able to identify the people who are ill, and specificity measures how well the test identifies people who are not ill. In other words, sensitivity is the number of people correctly classified as ill (true positives), divided by all the people who are ill (positives), while specificity is the number of people who are correctly classified as not ill (true negatives), divided by all the people who are not ill (negatives) (Lalkhen and McCluskey, 2008).

$$\text{Sensitivity} = \frac{\text{true positives}}{\text{positives}}, \quad \text{Specificity} = \frac{\text{true negatives}}{\text{negatives}}$$

The perfect binary classification test classifies all observations perfectly and has sensitivity and specificity equal to 1. However, most tests do not achieve this goal. In these cases, it is important to balance sensitivity and specificity. The reason for this is that even a useless test, which just categorizes all observations as negative, has specificity equal to 1.

3.7 C-statistic

This presentation of the C-statistic is based on the presentation in Harrell Jr. et al. (1996). The C-statistic measures how discriminating a prediction model is by measuring the concordance between the predicted responses and the observed responses. We consider a binary outcome, for instance the presence of a disease. Specifically, we consider a prediction model which gives the probability, for each participant, of getting the disease before a given time t_g. Consider all possible pairs of participants where one has the disease at t_g and the other does not have the disease at t_g. The C-statistic is the proportion of such pairs where the participant who got the disease had a higher predicted probability of getting the disease than the participant who didn't get the disease. In other words, the proportion of pairs with concordant predicted and observed values. The C-statistic can be calculated by the formula

$$\text{C-statistic} = \frac{B + 0.5E}{D * ND}, \tag{3.14}$$

where B is the number of pairs with concordant predictions and observations, E is the number of pairs where the participants have the same predicted probability, D is the number of people with the disease at t_g and ND is the number of people without the disease at t_g. From the definition above, it is clear that a model with C-statistic equal to 1 has perfect discrimination, and always assigns higher probability to people who get the disease, than to people who remain healthy.

Note that for models with binary outcomes, the C-statistic is equivalent to the area under the receiver operating curve (ROC), often referred to as the AUC (Hanley and McNeil, 1982).

Chapter 4

Models and methods

In this chapter, we present the prediction models and evaluation schemes used in this thesis and describe how we implement them in `R` (R Core Team, 2020) using the integrated development environment `RStudio` (RStudio Team, 2016).

4.1 Prediction models

4.1.1 Full Gaussian model

The first model we consider is a Gaussian GLM with identity link function and the systolic blood pressure at the time of HUNT3 as the response variable, denoted $\mathbf{Y} = [y_1, ..., y_n]$, where n =17 365 is the number of participants. The Gaussian GLM only predicts the continuous systolic blood pressure at the time of HUNT3, yet its ability to identify the systolic hypertension status of the participants at HUNT3 is also evaluated. The explanatory variables, all from HUNT2, included in the model are:

- Mean systolic blood pressure
- Mean diastolic blood pressure
- Birth year
- Sex
- BMI (Body Mass Index)
- PAI (Personal Activity Intelligence)
- RecPA (Recommended Physical Activity)
- Hypertensive parents
- Smoking

- Cholesterol
- HDL Cholesterol
- Non-fasting blood glucose
- GFR (Glomerular filtration rate)
- Creatinine
- Education level.

See Chapter 2.1.3 for the reasoning behind this choice of explanatory variables and further details on each variable. The response vector $\mathbf{Y} = [y_1, ..., y_n]$ is Gaussian distributed, $\mathbf{Y} = [y_1, ..., y_n] \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2\mathbf{I_n})$, with mean $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}$ and covariance matrix $\mathrm{Var}(\mathbf{Y}) = \sigma^2\mathbf{I_n}$. The linear predictor $\eta_i = \mathbf{x_{F,i}}^T\boldsymbol{\beta}$ of participant i is connected to the mean μ_i through an identity link function, for all $i = 1, ..., n$. Consequently, for each systolic blood pressure measurement from HUNT3 y_i, $i = 1, ...,$17 365, the model can be written on the form

$$\begin{aligned} y_i &\sim \mathcal{N}(\mu_i, \sigma^2), \\ f_{\mathcal{N}}(y_i; \mu_i, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right), \\ \eta_i &= \mathbf{x_{F,i}}^T\boldsymbol{\beta} = \mu_i, \end{aligned} \tag{4.1}$$

where μ_i is the mean, σ^2 is the variation of the observation noise, and $f_{\mathcal{N}}(\cdot)$ is the probability density function of the Gaussian distribution (Weisstein, 2020b). Using the same notation as in section 3.1, $\mathbf{x_{F,i}}^T$ is the i-th row vector of the design matrix X_F, and thus contains all of participant i's measurements of the explanatory variables listed above. $\boldsymbol{\beta}$ is the vector of true regression coefficients.

The true values of the regression coefficients $\boldsymbol{\beta}$ and the variation of the observation noise σ^2 are unknown. The iterative reweighted least squares (IRWLS) method is used to make an estimate $\hat{\boldsymbol{\beta}}$ of the true regression coefficient, see Fahrmeir et al. (2013) for details on IRWLS. To estimate the true variation of the observation noise σ^2, we use the sample variance of the residuals,

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon_i} - \bar{\epsilon})^2, \tag{4.2}$$

where $\hat{\epsilon_i} = y_i - \hat{y_i}$ are the residuals.

The predicted value of the i-th response is denoted by $\hat{y_i}$ and has the distribution

$$\hat{y_i} \sim \mathcal{N}(\mathbf{x_{F,i}}^T\hat{\boldsymbol{\beta}}, \mathrm{Var}(\hat{y_i})), \tag{4.3}$$

where the expression for the variance is found, by applying general rules of variance, to be

$$\text{Var}(\hat{y_i}) = \mathbf{x_{F,i}}^T \text{Var}(\hat{\boldsymbol{\beta}}) \mathbf{x_{F,i}} + \hat{\sigma^2}. \tag{4.4}$$

The residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$ of a Gaussian GLM are thus assumed to have zero mean and be homoscedastic (Fahrmeir et al., 2013).

The predicted probability of systolic hypertension for participant i at HUNT3 is consequently the integral from 140 to infinity of the probability density function of $\hat{y_i}$.

$$P(Sys.hyp) = \int_{140}^{\infty} pdf_{\hat{y_i}}(x)dx \tag{4.5}$$

We implement this model using the `R`-function `glm()` in `RStudio`.

4.1.2 Small Gaussian model

When implementing the full Gaussian GLM described in the previous section, it became apparent that many of the explanatory variables included in that model were not found to be significant when using a 0.05 significance level, see Chapter 5 for detailed results. To check whether a smaller model would perform as well or better, we implement a Gaussian GLM with only a selection of the explanatory variables from the full model. We choose to include the variables that were found to be significant, for significance level 0.05, as well as the PAI variable. The reason we also include PAI, even though it was not found to be significant in the larger model, is that we are especially interested in the effect of physical activity on systolic blood pressure. The explanatory variables, all from HUNT2, included in the smaller Gaussian GLM are

- Mean systolic blood pressure
- Mean diastolic blood pressure
- Birth year
- BMI (Body Mass Index)
- PAI (Personal Activity Intelligence)
- Hypertensive parents
- HDL Cholesterol
- Education level.

This second model is just a smaller version of the full model described in the previous section and can thus be formulated in the same way, except with a different design matrix X_S.

Using the same implementation method as for the full Gaussian model, we implement the small Gaussian model using the `R`-function `glm()` in `RStudio`.

4.1.3 Full gamma model

From Figure 2.5 it is clear that the distribution of the systolic blood pressure from HUNT3 is not symmetric. We observe that the right tail is somewhat heavier than the left tail of the distribution. The Gaussian distribution, however, is symmetric and is therefore not able to model this aspect of the response. In an effort to include the observed skewness in our prediction model, we consider a gamma GLM with the systolic blood pressure from HUNT3 as response variable $Y = [y_1, ..., y_n]$, the same explanatory variables as listed in the full Gaussian model in Section 4.1.1, and identity link function. It is reasonable to use an identity link function since the response values are positive and much bigger than zero. Same as the Gaussian GLM, the gamma GLM only predicts the continuous systolic blood pressure at the time of HUNT3, but its ability to identify the systolic hypertension status of the participants at HUNT3 is still evaluated. For each measurement of the systolic blood pressure from HUNT3 y_i, $i = 1, ..., n$, we have

$$\begin{aligned} y_i &\sim \Gamma(k, \lambda_i), \\ f_\Gamma(y_i; k, \lambda_i) &= \frac{\lambda_i(\lambda_i y_i)^{k-1}}{(k-1)!} e^{-\lambda_i y_i}, \\ \eta_i &= \mathbf{x_{F,i}}^T \boldsymbol{\beta} = \mu_i, \end{aligned} \tag{4.6}$$

where k is the shape parameter and λ_i is the rate parameter, and $f_\Gamma(\cdot)$ is the probability density function of the gamma distribution. Same as for the full Gaussian model, $\mathbf{x_{F,i}}^T$ is the i-th row vector of the design matrix X_F, and thus contains all of participant i's measurements of the explanatory variables. $\boldsymbol{\beta}$ is the vector of true regression coefficients. Since $\boldsymbol{\beta}$ is unknown, we use the iterative reweighted least squares (IRWLS) method to make an estimate $\hat{\boldsymbol{\beta}}$ of the true regression coefficients, see Fahrmeir et al. (2013) for details on IRWLS.

The true values of the shape and scale parameters are also unknown. However, we know the mean $\mathrm{E}(y_i) = \mu_i = \frac{k}{\lambda_i}$ and the variance $\mathrm{Var}(y_i) = \frac{\mu_i^2}{k} = \frac{k}{\lambda_i^2}$ of a gamma distribution (Weisstein, 2020a), and we also know that the gamma distribution belongs to the exponential family. By comparing the expression for the density function of the gamma distribution and the general formula for the density function for the exponential family in Equation (3.1), we see that the dispersion parameter is the reciprocal of the shape parameter $\phi = \frac{1}{k}$ (UIO, 2014). An estimate of the shape parameter, $\hat{k}$, is the thus the inverse of the estimated dispersion parameter, $\hat{\phi}$,

$$\hat{k} = \frac{1}{\hat{\phi}}.$$

From the expression of the expected value μ_i of y_i, it is clear that an estimate of the rate parameter $\hat{\lambda}_i$ can be found by dividing the estimated shape parameter $\hat{k}$ by the fitted values $\hat{\mu_i} = \mathbf{x_{F,i}}^T \hat{\boldsymbol{\beta}}$,

$$\hat{\lambda}_i = \frac{\hat{k}}{\hat{\mu_i}}.$$

Consequently, the predicted value of the i-th response is denoted by $\hat{y}_i$ and has the distribution

$$\hat{y}_i \sim \Gamma(\hat{k}, \hat{\lambda}_i). \tag{4.7}$$

The residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$ of a gamma GLM are thus assumed to have zero mean. The variance of the residuals can be expressed by

$$\text{Var}(y_i - \hat{y}_i) = \text{Var}(y_i) + \text{Var}(\hat{y}_i) - 2\text{Cov}(y_i, \hat{y}_i).$$

However, since the sample is so large, n=17365, Var$(\hat{y}_i)$ is so small that we can ignore it and thus also ignore Cov$(y_i, \hat{y}_i)$. In other words, we can estimate the variance of the residuals by the variance of y_i, Var$(y_i) = \frac{\mu_i^2}{k}$.

As for the Gaussian GLM, the predicted probability of systolic hypertension for participant i at HUNT3 is the integral from 140 to infinity of the probability density function of $\hat{y}_i$.

$$P(Sys.hyp) = \int_{140}^{\infty} pdf_{\hat{y}_i}(x)dx \tag{4.8}$$

We implement this model using the R-function `glm()` in `RStudio`.

4.1.4 Small gamma model

Similarly to the full Gaussian model, the full gamma model has many explanatory variables that aren't found to be significant on a 0.05 significance level, see Chapter 5 for detailed results. Consequently, we consider a smaller gamma model with systolic blood pressure from HUNT3 as the response variable, which only includes and PAI and the explanatory variables that were found to be significant on a 0.05 significance level in the full gamma model. As previously mentioned in the section about the small Gaussian model, we include PAI because we are especially interested in the effect of physical activity on systolic blood pressure. It turns out that the explanatory variables included in the small gamma model are the same as the explanatory variables in the small Gaussian model, see Section 4.1.2.

The small gamma model is just a smaller version of the full gamma model described in the previous section and can thus be formulated in the same way, except with a different design matrix X_S.

Same as for the full gamma model, we implement the small gamma model using the R-function `glm()` in `RStudio`.

4.1.5 Framingham model

We want to consider the Framingham model proposed by Parikh et al. (2008), and compare its performance on our cohort to the performance of the Gaussian and gamma prediction models proposed in the previous sections. According to Sun et al. (2017), the Framingham model has been externally validated by seven studies. These seven studies have been conducted on cohorts of different ethnicities, but the Framingham model has not been externally validated on a Scandinavian cohort before.

The Framingham model was originally conducted on a cohort consisting of 1717 white participants who were 20-69 years of age. Participants with hypertension or diabetes were excluded. In contrast to the models proposed in the previous sections, which give a prediction distribution for systolic hypertension after approximately 11 years, the Framingham model outputs the probability of getting hypertension within the next 1, 2, or 4 years. Parikh et al. (2008) uses the same definition of hypertension as we use in this thesis, namely systolic blood pressure $\geq$ 140 mmHg and/or diastolic blood pressure $\geq$ 90 mmHg and/or use of blood pressure medication.

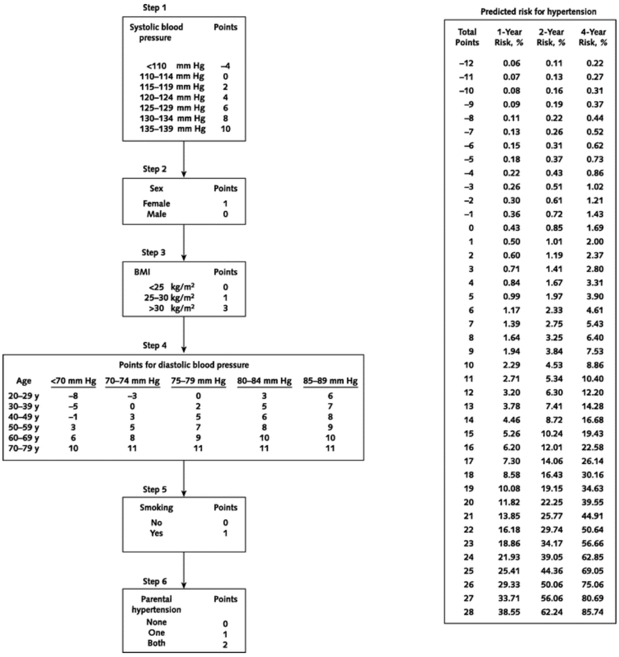
An illustration of the Framingham model taken from the original paper by Parikh et al. (2008) is shown in Figure 4.1. The 4-year risk of hypertension, given in percentages, is found by adding the points of the relevant levels of the explanatory variables and then finding the 4- year risk corresponding to the total score in the table to the right. Parikh et al. (2008) developed the point system version of the Framingham model shown in Figure 4.1 by using methods described in Sullivan et al. (2004) for a multivariable Weibull model with the explanatory variables listed below:

- sex
- systolic blood pressure
- BMI
- parental hypertension
- cigarette smoking
- interaction between age and diastolic blood pressure.

For the cohort in the paper by Parikh et al. (2008), the original Framingham model predicted a 4-year risk of hypertension below 5% for 34% of the participants, between 5% and 10% for 19% of the participants, and above 10% for 47% of the participants. The discrimination of the 4-year risk was measured by using the overall C-statistic (Harrell Jr. et al., 1996), and found to be good, with a C-statistic of 78.8% with 95% confidence interval (73.3,80.3). The calibration of the 4-year risk was also good, as the modified version of the Hosmer-Lemeshow chi-square statistic (Hosmer and Lemeshow, 1989) was found to be 4.35. We note that Parikh et al. (2008) do not specify in what manner the Hosmer-Lemeshow chi-square statistic was modified.

To make the comparison between the Framingham model and our models more reasonable, we modify the Framingham model by adding 7 years to the age of the participants at HUNT2 before using the Framingham model to find the probability of being hypertensive in 4 years. By adding 7 years to the age of the participants at HUNT2, the 4-year risk is more similar to the predictions of the Gaussian and gamma prediction models proposed by us, which predict systolic blood pressure approximately 11 years into the future. In practice, this means that we add 7 years to the age in the interaction term between age and diastolic blood pressure by subtracting 7 years from the birth year of the participants,

$$\text{Modified age at HUNT2} = \text{Year of HUNT2 - (BirthYear-7)}$$



Step 1

Systolic blood pressure	Points
<110 mm Hg	–4
110–114 mm Hg	0
115–119 mm Hg	2
120–124 mm Hg	4
125–129 mm Hg	6
130–134 mm Hg	8
135–139 mm Hg	10

Step 2

Sex	Points
Female	1
Male	0

Step 3

BMI	Points
<25 kg/m²	0
25–30 kg/m²	1
>30 kg/m²	3

Step 4

Points for diastolic blood pressure

Age	<70 mm Hg	70–74 mm Hg	75–79 mm Hg	80–84 mm Hg	85–89 mm Hg
20–29 y	–8	–3	0	3	6
30–39 y	–5	0	2	5	7
40–49 y	–1	3	5	6	8
50–59 y	3	5	7	8	9
60–69 y	6	8	9	10	10
70–79 y	10	11	11	11	11

Step 5

Smoking	Points
No	0
Yes	1

Step 6

Parental hypertension	Points
None	0
One	1
Both	2

Predicted risk for hypertension

Total Points	1-Year Risk, %	2-Year Risk, %	4-Year Risk, %
–12	0.06	0.11	0.22
–11	0.07	0.13	0.27
–10	0.08	0.16	0.31
–9	0.09	0.19	0.37
–8	0.11	0.22	0.44
–7	0.13	0.26	0.52
–6	0.15	0.31	0.62
–5	0.18	0.37	0.73
–4	0.22	0.43	0.86
–3	0.26	0.51	1.02
–2	0.30	0.61	1.21
–1	0.36	0.72	1.43
0	0.43	0.85	1.69
1	0.50	1.01	2.00
2	0.60	1.19	2.37
3	0.71	1.41	2.80
4	0.84	1.67	3.31
5	0.99	1.97	3.90
6	1.17	2.33	4.61
7	1.39	2.75	5.43
8	1.64	3.25	6.40
9	1.94	3.84	7.53
10	2.29	4.53	8.86
11	2.71	5.34	10.40
12	3.20	6.30	12.20
13	3.78	7.41	14.28
14	4.46	8.72	16.68
15	5.26	10.24	19.43
16	6.20	12.01	22.58
17	7.30	14.06	26.14
18	8.58	16.43	30.16
19	10.08	19.15	34.63
20	11.82	22.25	39.55
21	13.85	25.77	44.91
22	16.18	29.74	50.64
23	18.86	34.17	56.66
24	21.93	39.05	62.85
25	25.41	44.36	69.05
26	29.33	50.06	75.06
27	33.71	56.06	80.69
28	38.55	62.24	85.74

Figure 4.1: The original illustration of the Framingham model. Figure 2 in the paper by Parikh et al. (2008).

In other words, we use a modified version of Figure 4.1, taken from Parikh et al. (2008), as pseudocode when implementing the Framingham model on our cohort in `R` using `RStudio`.

4.2 Evaluation methods

One of the important steps of evaluating prediction models is observing how well the predictions of the models match the observed values, ie. the goodness-of-fit. In addition to this, it is interesting to check how well-calibrated the prediction models are by comparing the prediction distributions given by the models to the observed values. Another important step is to study the discrimination of the prediction model. In other words, the prediction models' ability to separate the participants who become systolic hypertensive from the participants whose systolic blood pressure remains at a healthy level.

4.2.1 Root mean square error

We implement the formula for the RMSE, see Equation (3.7), in `RStudio` and apply it to the full and smaller version of both the Gaussian and gamma models. The Framingham method only predicts the risk of hypertension and does not provide a predicted value of the systolic blood pressure, so it is not possible to calculate the RMSE for this model.

4.2.2 Brier score

To find the Brier score of the models we use the `BrierScore` function from the `DescTools` R-package (Signorell et al., 2020). The observed systolic blood pressure from HUNT3 and the probability of systolic hypertension predicted by the model are given as arguments. The probability of systolic hypertension is given directly as the response for the Framingham model, and can be found by Equation (4.5) for the Gaussian GLM and (4.8) for the gamma GLM.

4.2.3 Continuous rank probability score

To find the CRPS of the models we use the `crps` and `crps_gamma` functions from the `scoringRules` R-package (Jordan et al., 2019). The observed systolic blood pressure from HUNT3 and the parameters of the prediction distributions, ie. $(\hat{\mu_i}, \mathrm{SD}(\hat{y_i}))$ for the Gaussian model and $(\hat{k}, \hat{\lambda_i})$ for the gamma model, are given as arguments. Note that the prediction models we propose do not provide a single cumulative distribution function, but a different predicted cumulative distribution function for each participant. This means that each predicted cumulative distribution function is evaluated based on its corresponding observed value. It is not possible to find the CRPS of the Framingham model since it does not produce a prediction distribution.

4.2.4 Probability Integral Transformation diagrams

The Probability Integral Transformation diagrams are implemented in `RStudio` by evaluating the predicted cumulative distribution functions at the observed values of the systolic blood pressure from HUNT3 and plotting the result in a histogram. Note that in contrast to the case described in Chapter 3.5, the prediction models we propose do not provide a single cumulative distribution function, but a different predicted cumulative distribution function for each participant.

4.2.5 Sensitivity and specificity

We find the sensitivity of the proposed models by dividing the number of participants who were both systolic hypertensive at HUNT3 and had a predicted value of systolic blood pressure equal to or above 140 mmHg, by all the participants who were systolic hypertensive at HUNT3. The specificity of the proposed models is found by dividing the number of participants who were not systolic hypertensive at HUNT3 and had a predicted value of systolic blood pressure below 140 mmHg, by all the participants who were not systolic hypertensive at HUNT3. Since the Framingham model doesn't produce any predicted systolic blood pressure values, we create our own definition of systolic hypertension in the calculation of sensitivity and specificity of the Framingham model. Namely that a predicted probability of hypertension above 0.5 equal predicted hypertension.

4.2.6 C-statistic

The Framingham model and the Gaussian and gamma prediction models we consider can all be viewed as prediction models that predict the probability of participants getting a certain disease, systolic hypertension, before a given time t_g, the time of HUNT3. This is the same situation as described in Section 3.7, and we can, thus, use the same formula, Equation (3.14), to calculate the C-statistic for these models. We implement the simple formula in `RStudio` for all 5 models.

4.3 Implementation

The `R` code used in this thesis is available at `https://github.com/fridentnu/masterthesis`. This repository mainly contains the code used to clean, explore, and transform the data as described in Chapter 2, and to implement, inspect, and evaluate the prediction models as described in this chapter. In addition, there is some code used to create illustrative figures to explain the PIT diagrams in Chapter 3.5 in the file `Illustrations.R`. In general, we have used base `R` functions and functions from the `ggplot2` (Wickham, 2016) and `DataExplorer` (Cui, 2020) `R`-packages to create figures.

All of the code is written by Fride Nordstrand Nilsen, except the code used to calculate the PAI-level, `PAI.R`, and the MVPA- score, `MVPA.R`. These scripts are written by Emma Ingström, a Ph.D. student who also works with data from the HUNT study. The code in `PAI.R` and `MVPA.R` is based, respectively, on the papers by Kieffer et al. (2019) and Ernstsen et al. (2016).

The code files used to clean and explore the data are named `DataCleaning.R` and `EDA.R`, respectively. The data transformation and implementation of the GLM models are located in `Models.R`. The modified Framingham model is implemented in `Framingham.R`. The code creating the figures of the residuals is located in the file `Residuals`, while the code exploring the prediction distributions of individual participants is in the file `EvalParticipant.R`. All the evaluation methods are implemented and applied to the prediction models in `Evaluation.R`.

Note that due to privacy reasons the data used in this thesis is not available on the GitHub-webpage. However, the format of the original data is described in Chapter 2 and in `DataCleaning.R`.

Chapter 5

Results

In this chapter, we present the main results of the prediction models, and we evaluate their performance on the observed systolic blood pressure at HUNT3. We start by presenting the fitted Gaussian and gamma models, before evaluating and comparing the models, including the Framingham model.

5.1 Presenting main results of the models

Figure 5.1 gives a quick overview of the prediction models by showing the predicted values of systolic blood pressure at HUNT3, in addition to the observed values of the systolic blood pressure at HUNT2 and HUNT3. A perfect prediction model has predicted values equal to the observed values from HUNT3. The predicted values given by our proposed prediction models, on the other hand, seem more similar to observed systolic blood pressure at HUNT2, and very similar to each other.

5.1.1 Full and small Gaussian models

The assumptions and implementation of the full and small Gaussian model are described in Chapter 4.1.1-4.1.2, and the values, standard deviations, and p-values of the regression coefficients are presented in Table 5.1. As noted earlier, only PAI and the explanatory variables that were found to be significant on a 0.05 significance level in the full Gaussian model are included in the small Gaussian model. From Table 5.1, it is clear that the regression coefficients of the explanatory variables included in both models are very similar.

The variable with the biggest positive, and biggest absolute, influence on the predicted value is not surprisingly the systolic blood pressure at HUNT2. In these models, the predicted systolic blood pressure will increase by approximately 5 mmHg for each increase the size of a standard deviation of the systolic blood pressure at HUNT2, when other explanatory variables are held constant. The second most influential variable is birth year, with a negative regression coefficient. This means that for two people with equivalent

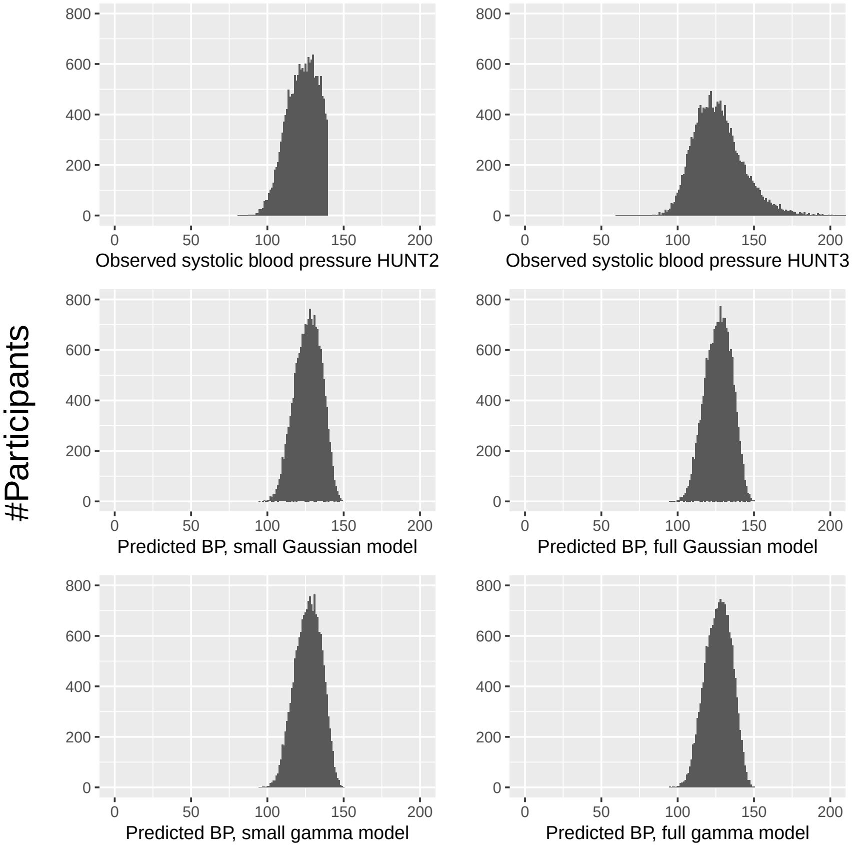


Figure 5.1: Histograms of the observed and predicted values of the systolic blood pressure of the participants.

values of the other variables, the oldest person will get a higher predicted systolic blood pressure. Other significant explanatory variables listed in decreasing order of influence are diastolic blood pressure, parental hypertension, BMI, Education Level 4 and 5, HDL Cholesterol, and PAI.

The most noticeable differences between the small and full model are that PAI has more influence and is significant in the small Gaussian model. Another difference between the models is that the standard deviations of the regression coefficients are slightly smaller in the small model.

5.1.2 Full and small gamma models

The assumptions and implementation of the full and small gamma model are described in Chapter 4.1.3-4.1.4, and the values, standard deviations, and p-values of the regression coefficients are presented in Table 5.2. The full and small gamma models have the same

Exp.Variable	FM.Est	SM.Est	FM.SD	SM.SD	FM.p.val	SM.p.val
(Intercept)	126.644	126.694	0.386	0.277	0	0
BirthYear	-3.623	-3.594	0.135	0.118	0	0
SexMale	0.176	NA	0.363	NA	0.628	NA
BMI2	1.412	1.397	0.114	0.111	0	0
SystolicBP2	5.028	5.005	0.133	0.128	0	0
DiastolicBP2	2.183	2.163	0.131	0.13	0	0
PAI2Moderate	-0.074	-0.119	0.32	0.271	0.817	0.661
PAI2High	-0.603	-0.72	0.498	0.251	0.226	0.004
RecPA2TRUE	-0.067	NA	0.42	NA	0.873	NA
BPHigPar2TRUE	1.906	1.933	0.226	0.224	0	0
Smoking2Previous	-0.387	NA	0.261	NA	0.139	NA
Smoking2Current	-0.141	NA	0.262	NA	0.591	NA
Cholesterol2	-0.061	NA	0.12	NA	0.612	NA
HDLCholesterol2	-0.653	-0.624	0.118	0.109	0	0
Glucose2	0.043	NA	0.107	NA	0.691	NA
GFR2Stage 2	0.336	NA	0.35	NA	0.337	NA
GFR2Stage 345	0.058	NA	1.326	NA	0.965	NA
Creatinine2	-0.316	NA	0.202	NA	0.118	NA
Education2Level 2	-0.178	-0.194	0.295	0.293	0.547	0.508
Education2Level 3	0.15	0.164	0.412	0.409	0.717	0.688
Education2Level 4	-0.725	-0.704	0.363	0.357	0.046	0.049
Education2Level 5	-0.837	-0.822	0.407	0.399	0.04	0.039

Table 5.1: The value, standard deviation, and p-value of the regression coefficients of the full and small Gaussian model. FM denotes the full model, and SM denotes the small model. SD is the standard deviation and p.val denotes the p-value. NA signifies that the corresponding variable isn't included in the small Gaussian model

explanatory variables as the corresponding Gaussian models. It is clear that the regression coefficients of the full and small gamma model are very similar, and that they are also very close to the regression coefficients of the Gaussian prediction models, presented in Table 5.1. In fact, the list of explanatory variables in decreasing order of influence for the gamma prediction models is identical to the corresponding list for the Gaussian models. For clarity, the explanatory variables listed in order of biggest to smallest influence on the predicted systolic blood pressure are systolic blood pressure at HUNT2, birth year, diastolic blood pressure at HUNT2, parental hypertension, BMI, Education Level 4 and 5, HDL Cholesterol, and PAI.

Similar to the Gaussian models, we observe that PAI has more influence in the small gamma model than in the full gamma model and that a High level of PAI is significant only in the small model. The standard deviations of the regression coefficients are slightly smaller in the small model, like in the Gaussian models.

Exp.Variable	FM.Est	SM.Est	FM.SD	SM.SD	FM.p.val	SM.p.val
(Intercept)	126.636	126.744	0.382	0.276	0	0
BirthYear	-3.598	-3.587	0.134	0.117	0	0
SexMale	0.446	NA	0.357	NA	0.212	NA
BMI2	1.437	1.425	0.113	0.11	0	0
SystolicBP2	4.931	4.942	0.129	0.124	0	0
DiastolicBP2	2.155	2.134	0.128	0.127	0	0
PAI2Moderate	-0.1	-0.127	0.315	0.268	0.751	0.635
PAI2High	-0.586	-0.652	0.49	0.246	0.231	0.008
RecPA2TRUE	-0.043	NA	0.414	NA	0.918	NA
BPHigPar2TRUE	1.909	1.915	0.223	0.221	0	0
Smoking2Previous	-0.433	NA	0.258	NA	0.093	NA
Smoking2Current	-0.207	NA	0.257	NA	0.421	NA
Cholesterol2	-0.032	NA	0.118	NA	0.785	NA
HDLCholesterol2	-0.671	-0.672	0.116	0.106	0	0
Glucose2	0.049	NA	0.106	NA	0.644	NA
GFR2Stage 2	0.34	NA	0.344	NA	0.323	NA
GFR2Stage 345	0.468	NA	1.333	NA	0.725	NA
Creatinine2	-0.352	NA	0.197	NA	0.074	NA
Education2Level 2	-0.284	-0.279	0.295	0.293	0.336	0.342
Education2Level 3	0.111	0.126	0.403	0.4	0.782	0.752
Education2Level 4	-0.867	-0.835	0.359	0.352	0.016	0.018
Education2Level 5	-0.967	-0.924	0.401	0.393	0.016	0.019

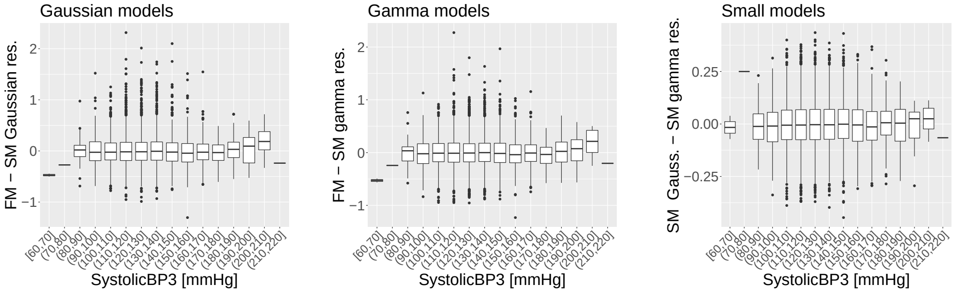
Table 5.2: The value, standard deviation, and p-value of the regression coefficients of the full and small gamma model. FM denotes the full model, SM denotes the small model, SD denotes standard deviation and p.val denotes the p-value. NA signifies that the corresponding variable isn't included in the small gamma model.

5.1.3 Residuals

From Table 5.1 and Table 5.2 we observe that the regression coefficients of all four models are quite similar. Since the regression coefficients, and thus also the predicted values, of the models, are similar, it follows that the residuals of the models should be similar too. To check this we study the difference between the residuals of the small and full Gaussian model, between the small and full gamma model, and between the small Gaussian and small gamma model in Figure 5.2.

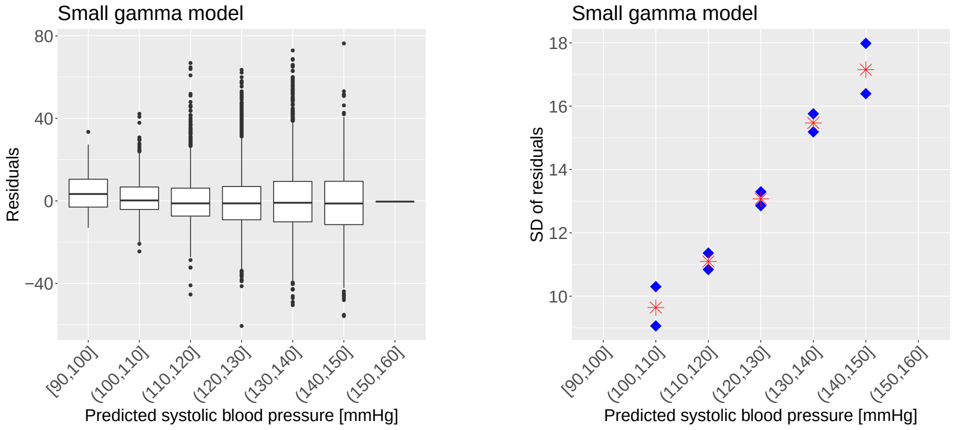
Not surprisingly, we observe that the residuals are very similar in all three cases. Comparing the full and small models the majority of residuals are within 1 mmHg of each other. The difference is even smaller between the small Gaussian and small gamma model, where the majority of residuals are within 0.25 mmHg of each other. Consequently, we only study the residuals of the small gamma model from now on, since any trends we find in the residuals is representative for the residuals from all four prediction models.

In Figure 5.3 a) the relationship between the residuals of the small gamma model and the predicted systolic blood pressure is shown. As mentioned in Chapter 4, the residuals of both the Gaussian GLM and the gamma GLM are assumed to have mean 0, which seems to



(a) **(b)** **(c)**

Figure 5.2: Difference between residuals from two different models versus observed systolic blood pressure from HUNT3. SystolicBP3 is short for observed systolic blood pressure at HUNT3. a) Full Gaussian model residuals minus small Gaussian model residuals; b) Full gamma model residuals minus small gamma model residuals; c) Small Gaussian model residuals minus small gamma model residuals



(a) **(b)**

Figure 5.3: The a) values; and b) standard deviations and corresponding 95% confidence intervals of the standard deviation,; of the residuals of the small gamma model versus the predicted systolic blood pressure. The standard deviations are not plotted for intervals containing less than 15 participants.

be approximately true for most values of the predicted systolic blood pressure. Observing the figure more closely, we see that for predicted systolic blood pressure between 90 and 100 mmHg, the median of the residuals is somewhat higher than 0. On the other hand, the median of the residuals is very slightly below zero for predicted blood pressure between 110 and 150 mmHg. For all values of the predicted blood pressure, there are more positive than negative outliers. This fits with the lack of a heavy right tail in the predicted values compared to the observed values shown in Figure 5.1.

The relationship between the standard deviations of the residuals and the predicted systolic blood pressure given by the small gamma model is presented in Figure 5.3 b). The approximate 95% confidence interval for the standard deviations of the residuals are plotted in blue in the figure. Note that the 95 % confidence intervals are calculated by taking the square roots of the limits of the 95% confidence interval of the variance, which are found using the critical values in the χ^2-distribution. In other words, the 95% confidence intervals only indicate the approximate uncertainty in the standard deviation estimates.

The Gaussian GLM assumes that the standard deviation of the residuals should be constant with the response, see Chapter 4.1.1. This is clearly not the case in Figure 5.3 b). The gamma GLM however, assumes that the standard deviation should increase linearly with the expected value of the response, see Chapter 4.1.3, which suits the results much better.

Figure 5.4 presents the residuals of the small gamma model versus the explanatory variables in the small gamma model. The median of the residuals is mostly independent of the value of the explanatory variables, and approximately 0. In the intervals with the fewest participants, the median tends to deviate slightly from 0. An example is the interval between 80 to 90 for systolic blood pressure from HUNT2 where the median of the residuals is close to 10. The most extreme deviation of the median from 0 is for diastolic blood pressure below 40 mmHg. However there are so few participants in this category, and our main focus is not on the participants with extreme values of blood pressure. We move on to study the standard deviations of the residuals in relation to the explanatory variables.

Figure 5.5 presents the standard deviations, as well as the 95% confidence intervals of the standard deviations, of the residuals of the small gamma model versus the explanatory variables in the small gamma model. The standard deviation is not plotted if there are fewer than 15 participants in an interval/category. We already know that the standard deviation of the residuals is assumed to depend linearly on the expected value of the response in a gamma GLM. Since we use the identity link function, the standard deviation of the residuals should also depend linearly on the linear predictor, which is a linear combination of the explanatory variables.

We immediately notice that the standard deviations of the residuals depend on the values of the explanatory variables. There is a significant and linear increase in the standard deviation of the residuals as the systolic blood pressure from HUNT2 increases, see Figure 5.5 a). There is a similar trend for the diastolic blood pressure from HUNT2 in Figure 5.5 b), except for the first two intervals. The reason for the high standard deviation in these two intervals might be that are so few participants in these intervals, see Figure 6.2 in the Appendix. The explanatory variable with the biggest change in the standard deviation of the residuals is birth year, see Figure 5.5 c). The older the participant, the bigger the standard deviation of the residuals is. However, it is important to note that there are fewer participants born before 1930, see Figure 2.11.

Once again, the effect of different levels of PAI in Figure 5.5 d) is small, but visible. The standard deviation decreases slightly with an increased activity level. If we disregard the interval with the highest BMI values, there seems to be a linear increase in the standard deviation of the residuals as a function of BMI in Figure 5.5 e). The standard deviation for the interval is above the line. This is probably caused by the small number of participants in this category, see Figure 2.15. The standard deviation is slightly bigger for participants

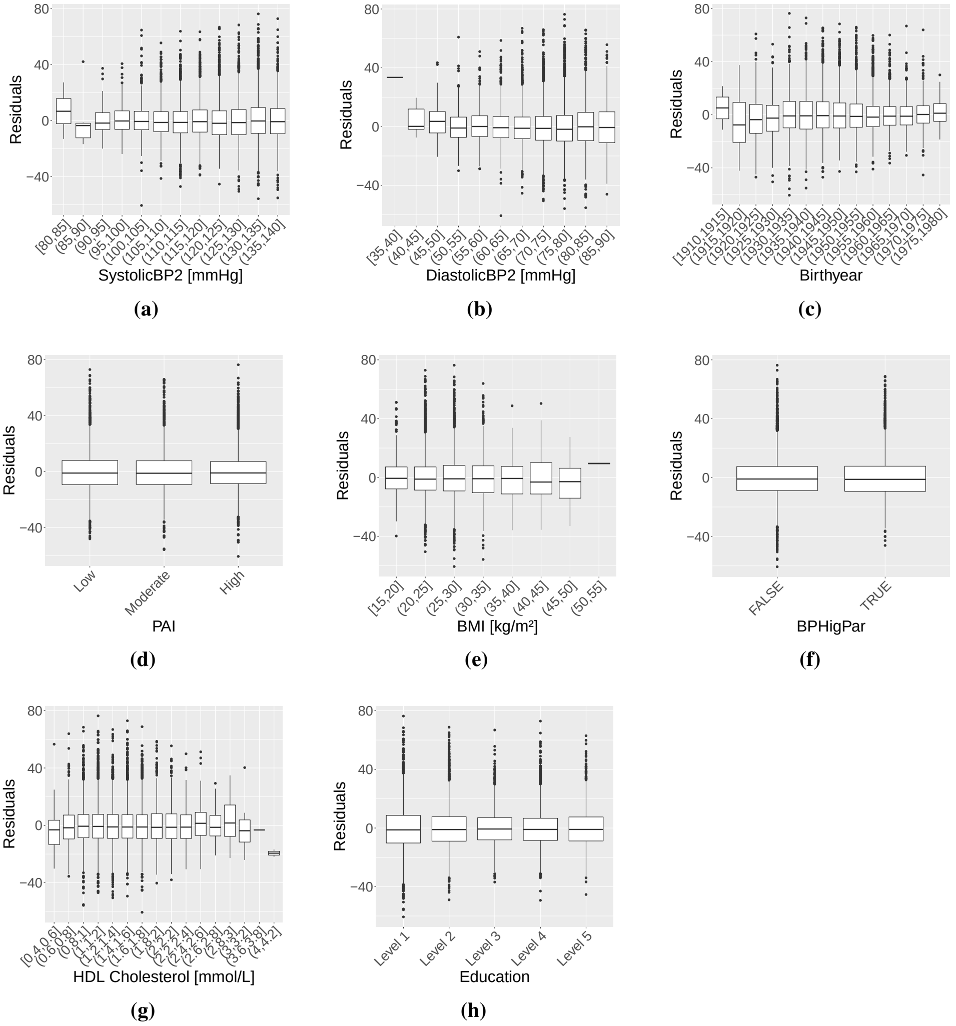
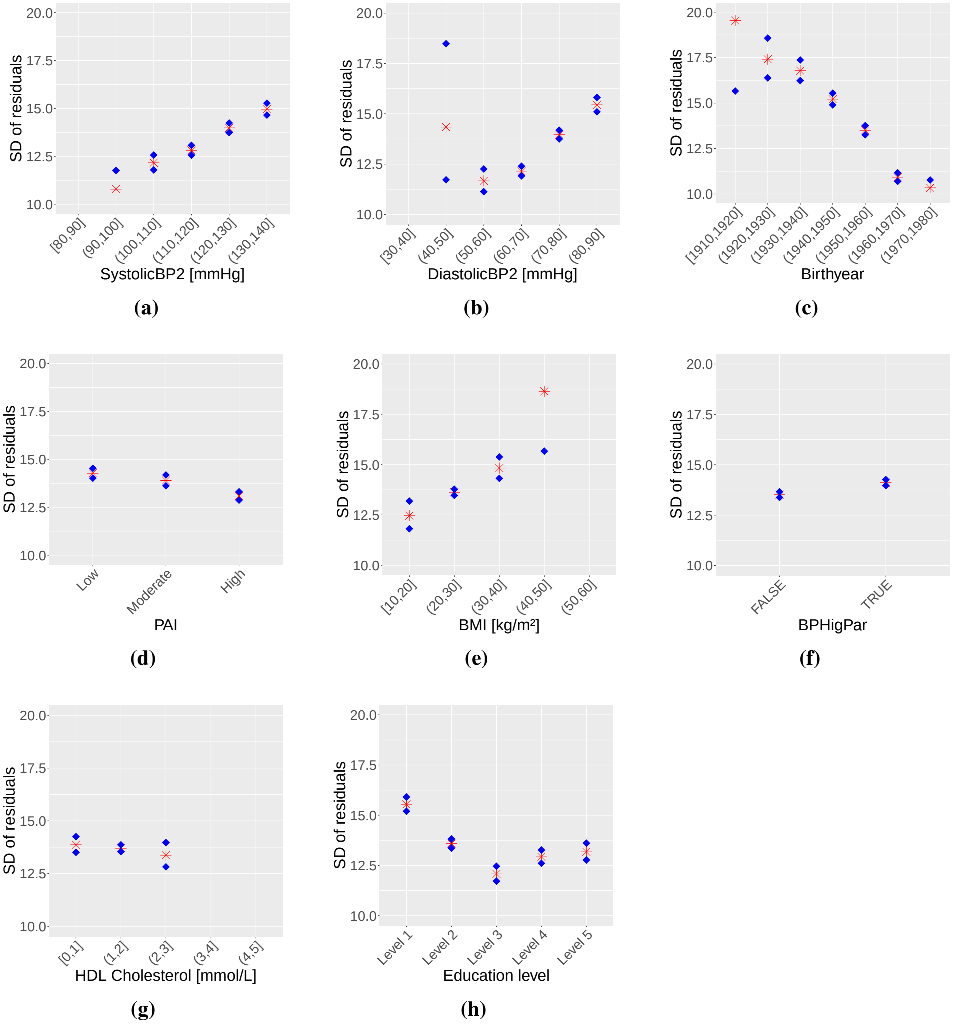


Figure 5.4: Residuals from the small gamma model versus a) Systolic blood pressure at HUNT2; (b) Diastolic blood pressure at HUNT2; (c) Birthyear; (d) PAI; (e) BMI; (f) Parental hypertension; (g) HDL Cholesterol; (h) Highest education level achieved at HUNT2

with parental history of hypertension, but the effect is quite small, see Figure 5.5 f). There is a small decrease in the standard deviations as the HDL Cholesterol increases, and in the last two intervals there are fewer than 15 participants so the standard deviation is not plotted in Figure 5.5 g). Since education level is a categorical variable not connected to a continuous scale, it is perhaps not surprising that there isn't a clear trend for the standard deviations in Figure 5.5 h). We know from Figure 2.18 that Education level and birth year are quite correlated, and we see the same trend here. The difference in standard



(a) **(b)** **(c)** **(d)** **(e)** **(f)** **(g)** **(h)**

Figure 5.5: SD of residuals from small gamma model versus a) Systolic blood pressure at HUNT2; (b) Diastolic blood pressure at HUNT2; (c) Birthyear; (d) PAI; (e) BMI; (f) Parental hypertension; (g) HDL Cholesterol; (h) Highest education level achieved at HUNT2. The standard deviation is not plotted if there are fewer than 15 participants in an interval/category.

deviation for the different education levels is therefore probably caused by the decrease of the standard deviation of the residuals as birth year increases, see Figure 5.5 c).

Generally, the standard deviation of the residuals seems to depend approximately linearly on the explanatory variables. We also notice that the sign of the slope of the line matches the sign of the regression coefficient of the corresponding explanatory variable. However, the value of the slope does not correspond to the value of the regression coefficient. This is clear since the systolic blood pressure from HUNT2 has the largest regression coefficient in absolute value, while birth year has the biggest change in standard deviation.

5.1.4 Prediction distributions of individual participants

Before we evaluate the total performance of prediction models, we present the prediction distributions from the four models for two individual participants, see Figure 5.6. Neither participant had diabetes, CVD, or was taking blood pressure medication at the time of HUNT3.

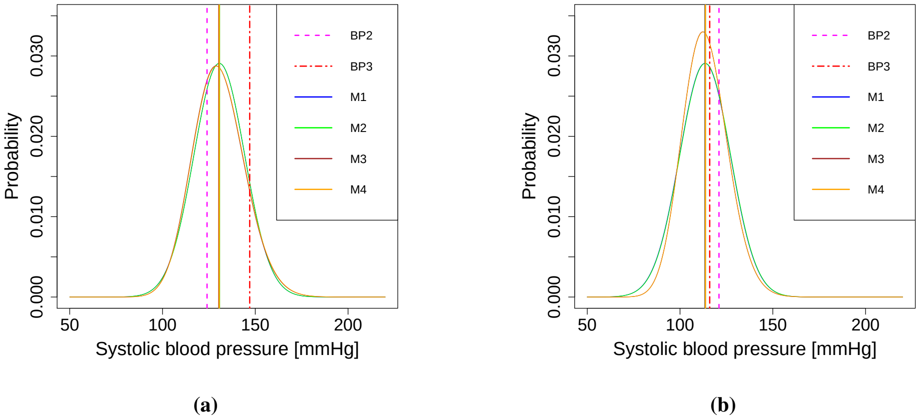


Figure 5.6: Prediction distributions of the different models for two different participants. M1 is the full Gaussian model, M2 is the small Gaussian model. M3 is the full gamma model, and M4 is the small gamma model. BP2 is the systolic blood pressure from HUNT2, while BP3 is the systolic blood pressure from HUNT3. Note that due to the similarity of the models M1 is behind M2, and M3 is behind M4, and so M1 and M3 are not visible.

Participant a) is a woman who at the time of HUNT2 was in her 40s, had a Low PAI level, and a BMI of approximately 25 kg/m^2. Her systolic blood pressure, see Figure 5.6 a), increased from 124 mmHg at HUNT2 to 147 mmHg at HUNT3. All the models predicted an increase in systolic blood pressure, but only to a predicted value of approximately 130 mmHg at the time of HUNT3. In other words, the predicted values given by the prediction models are very similar for all the models, which can also be seen in Figure 5.6 a). From the figure, it is clear that the prediction distributions for the small and full versions of the models are practically identical. The gamma and Gaussian prediction distributions are also close, but the gamma distributions have a slightly heavier right tail.

Participant b) is a man who at the time of HUNT2 was in his 20s, had a High PAI level, and a BMI of approximately 21 kg/m^2. From Figure 5.6 b), we observe that he had decreasing systolic blood pressure from HUNT2, 121 mmHg, to HUNT3, 116 mmHg. The prediction models all predicted a decrease in the systolic blood pressure, yet they predicted a slightly bigger decrease, with predicted value at approximately 114 mmHg. Similarly to Participant a) the differences between the prediction distributions of the small and full versions of the models are not visible. The prediction distributions for the gamma and Gaussian model are not as close for Participant b) as for Participant a). In this case, the Gaussian distribution actually has heavier tails in both directions.

For both Participant a) and Participant b) the prediction distributions from all four prediction models have large variances. In other words, the uncertainties in the predictions are big. To check if this is a general trend we plot a histogram of the standard deviations of all the individual prediction distributions from the small gamma model in Figure 5.7. The standard deviations of the small gamma prediction distributions are calculated based on the estimated shape and rate parameters of the small gamma model. The values of the standard deviations vary from 10 up to 16, with 13 to 14 being the most common values.

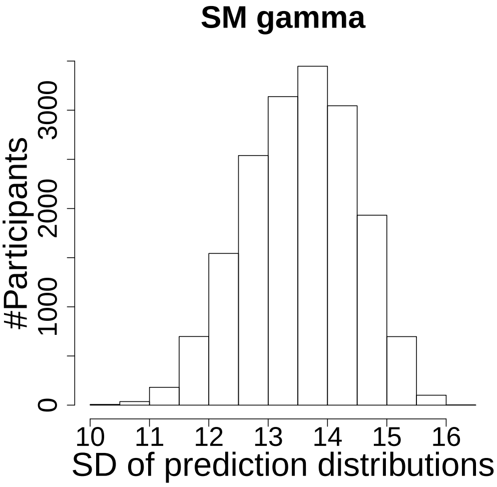


Figure 5.7: The standard deviations of the individual prediction distributions of systolic blood pressure given by the small gamma model.

5.2 Evaluation of model performance

In this section, we study the performance of the prediction models by using the evaluation methods described in Chapters 3 and 4. In particular, we are interested in the goodness-of-fit, calibration, and discrimination of the models. In addition to the previously described evaluation methods, we study the distribution of the probabilities of systolic hypertension and check whether the expected percentage of systolic hypertension matches the observed percentage.

In the previous sections, it has been shown that the regression coefficients and the residuals of the full and small versions of the same GLM are nearly identical. From prediction distributions of two individual participants in Chapter 5.1.4, it might seem like the same similarities can be found in the prediction distributions. Table 5.3 presents the nu-

Eval.Method	Full.Gauss	Small.Gauss	Full.Gamma	Small.Gamma	Framingham
Exp. Hyp	20.782	20.778	20.094	20.089	16.155
RMSE	13.705	13.708	13.706	13.708	NA
BrierScore	0.13246	0.13251	0.13244	0.13242	0.1348
CRPS	7.5498	7.5506	7.5022	7.5035	NA
Sensitivity	15.11	15.081	14.875	14.639	11.723
Specificity	96.979	97.008	96.979	97.101	97.688
C-statistic	78.082	78.053	78.049	78.05	77.483

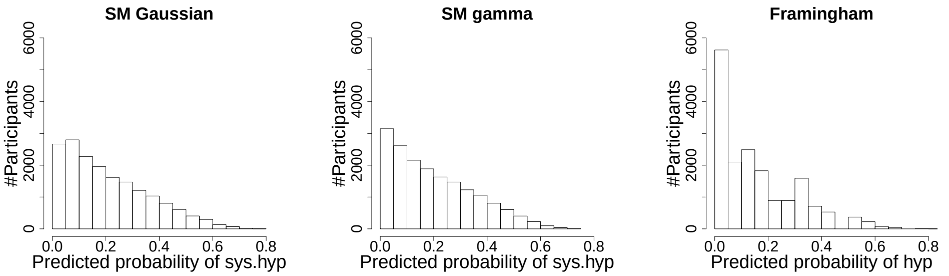
Table 5.3: Overview of the numerical results of the evaluation methods applied to the 4 prediction GLMs and the Framingham model. Exp.Hyp is short for expected percentage of systolic hypertensives. Sensitivity, specificity and C-statistic are also given as percentages.

merical results of the evaluation methods applied to all the prediction models, including the modified Framingham model. The performance of the small and full versions of the GLMs are very similar, which is as expected due to all the other similarities previously discussed. Consequently, we only evaluate the small Gaussian and the small gamma model from now on. To compare the performance of these models with a prediction model from literature, we also evaluate a modified version of the Framingham model, see Chapter 4.1.5 for details.

5.2.1 Predicted probability of systolic hypertension

We are interested in the predicted probability given to each participant of getting systolic hypertension, and how this matches the observed systolic hypertension at HUNT3. Figure 5.8 presents histograms of the predicted probabilities of getting systolic hypertension given by the small Gaussian and gamma GLMs and the predicted probability of getting general hypertension given by the Framingham model. All three models have predicted a low probability of hypertension for more participants than they have predicted high probability for. However, from the figure, it is clear that the Framingham model predicts close to 0 probability of hypertension for almost twice as many participants as the GLMs. The small Gaussian and small gamma models have a more linear decline in the number of participants as the predicted probability increases than the Framingham model. There is a large portion of the participants with a predicted probability of systolic hypertension relatively close to 50% for all three models, especially the GLMs. None of the models predict over 80% probability of hypertension for a significant amount of participants.

The observed systolic hypertension at HUNT3 was that 19.551% of the cohort were systolic hypertensive. Consequently, the fact that the models predict more instances of low probability of hypertension than high probability of hypertension matches the observed data. The expected percentage of systolic hypertension at HUNT3 according to the prediction models are presented in Table 5.3. The small Gaussian model expects 20.778% of the cohort to be systolic hypertensive, while the small gamma model was a little closer to the truth with an expected percentage of 20.089%. The Framingham model had a more conservative estimate of 16.155%.



(a) **(b)** **(c)**

Figure 5.8: The predicted probabilities of systolic hypertension at HUNT3 for the participants given by the a) Small Gaussian model; b) Small gamma model; c) Framingham model. Note that the Framingham model gives the probability for general hypertension, not just systolic hypertension.

5.2.2 RMSE

The RMSE of both the small Gaussian and the small gamma model is 13.708, see Table 5.3. This indicates that neither model has predicted values of the systolic blood pressure that fit the observed systolic blood pressure at HUNT3 particularly well. In other words, the goodness-of-fit of the models are not good. Since the Framingham model doesn't provide a predicted value of systolic hypertension, it isn't possible to find the RMSE of the Framingham model.

5.2.3 Brier score

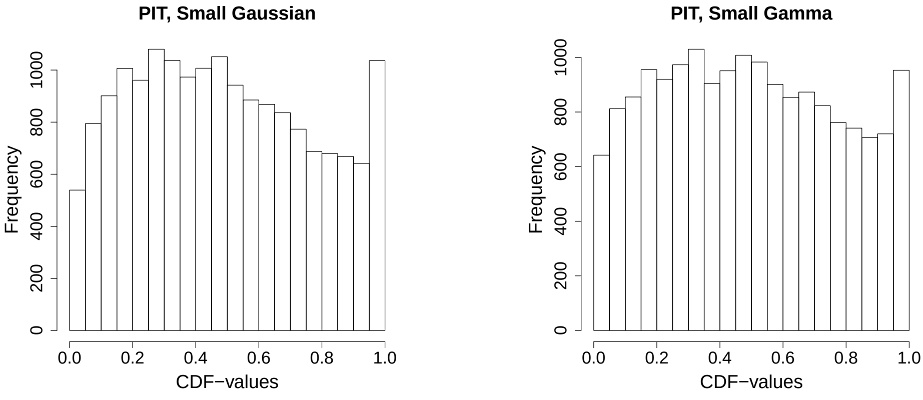
The Brier score measures the accuracy of the predictions of each participant's hypertension status at HUNT3 made by the prediction models. From Table 5.3 we observe that the small Gaussian model has a Brier score of 0.13251, while the small gamma model has a slightly smaller Brier score of 0.13242 and the Framingham model a somewhat higher Brier score of 0.13480. The GLMs have a similar performance to the well-known Framingham model, which indicates the validity of our models in this aspect. Since a perfectly accurate prediction model has Brier score 0, and the worst possible model has Brier score 1, the models perform reasonably well. The small gamma model has the highest accuracy of predicting whether a participant becomes hypertensive.

5.2.4 CRPS

The CRPS is a measure of how well the models are calibrated and can be interpreted as a measure of the distance between the predicted cumulative distribution function and the empirical cumulative distribution function of the observed systolic blood pressure at HUNT3. The small gamma model has a CRPS of 7.5035, which is somewhat smaller than the CRPS of the small Gaussian model at 7.5506. This indicates that the distributions predicted by the small gamma model are a little closer to the empirical distribution of the systolic blood pressure. The reason for this might be that the gamma distribution has a slightly heavier right tail, and we also observe a heavy right tail in the observed systolic blood pressure in Figure 5.1. Since the Framingham model doesn't provide a predicted cumulative distribution function of systolic hypertension, it isn't possible to find the CRPS of the Framingham model.

5.2.5 PIT Diagram

PIT diagrams are a visual way to check the goodness-of-fit of the predicted cumulative distribution functions given by the prediction models compared to the empirical distribution function of the observed systolic blood pressure. The PIT diagrams of the small Gaussian and small Gamma model are presented in Figure 5.9. As explained in detail in Chapter 3.5, a PIT diagram where the observations come from the proposed cumulative distribution should have bars of equal height. From Figure 5.9 we observe that both models have a higher bar closest to 1, which indicates that the empirical cumulative distribution function has a heavier tail than the predicted cumulative distribution function of the small gamma and small Gaussian models. In contrast, the bar closest to 0 is noticeably shorter than the other bars, especially so for the small Gaussian model. This indicates that the predicted cumulative distributions have a heavier left tail than the empirical distribution function. We also notice that the bars between approximately 0.15 and 0.7 for the small Gaussian model, and approximately 0.15 and 0.6 for the small gamma model are taller than the other bars. The bars in the PIT diagram of the small gamma model are more equal in height than the bars in the PIT diagram of the small Gaussian model, which indicates that the cumulative distribution functions of the small gamma model are closer to the empirical distribution function of the observed systolic blood pressure at HUNT3.



(a)

(b)

Figure 5.9: PIT diagrams of a) the small Gaussian model; and (b) the small gamma model.

As with the CRPS, it is not possible to create a PIT diagram of the Framingham model since it does not provide a predicted cumulative distribution function of the systolic blood pressure.

5.2.6 Sensitivity and specificity

The sensitivity of our tests measures how well they are able to identify the people who become systolic hypertensive, while the specificity of the tests measure how well they are able to identify the people with a systolic blood pressure which remains at a healthy level. The small Gaussian model is able to identify 15.081 % of the people who become hypertensive, the small gamma model identifies 14.639 %, and the Framingham model identifies 11.723%. As previously seen in the expected percentage of hypertension, the Framingham model predicts fewer people will get hypertension than the GLMs predict will get systolic hypertension. However, neither of the three models are able to identify anywhere near the true number of people who become systolic hypertensive. The specificity of the models is much better. All three models have a specificity of 97 % or higher, with the small Gaussian model at 97.008%, the small gamma model at 97.101%, and the Framingham model at 97.688%. In other words, very few people who remain non-hypertensive is predicted to become hypertensive or systolic hypertensive. In conclusion, all three models are much better at identifying the participants who remain healthy, than the participants who get hypertension and systolic hypertension. This might be because there are so many more people who remain healthy compared to people who become hypertensive. Note that good specificity is not necessarily an indication of a reasonable model in itself, since a model which predicts that everyone remains non-hypertensive, regardless of the values of the explanatory variables, would have a specificity of 100%.

5.2.7 C-statistic

The C-statistic measures the concordance between the predicted blood pressure and the observed blood pressure. In other words, it measures to what degree the prediction models assign a higher probability of hypertension to the people who become systolic hypertensive than to the people who still have healthy systolic blood pressure at HUNT3. The C-statistic of the small Gaussian model is 78.053, it is 78.050 for the small gamma model and 77.483 for the Framingham model. The performances of the models are very similar and quite good. In most cases, the models assign a higher probability of hypertension to the people who become hypertensive. The small Gaussian model has a slightly better discrimination ability than the two other models.

Chapter 6

Discussion and conclusion

In this chapter, we discuss the model assumptions, the performance of the models, the effect of physical activity on the predicted systolic blood pressure, and compare the performance of the models to the original Framingham model. Lastly, we reach a conclusion and suggest some ideas for possible future work.

6.1 Discussion

In Chapter 2 we discovered that the percentage of women in our cohort, 61.93%, is noticeably higher than the percentage of women in the total Norwegian population, 49.61% (Statistics Norway, 2018). This calls into question whether our cohort is representative of the total Norwegian population. However, the question of representability is out of the scope of this thesis. Luckily, our cohort is quite large which makes certain deviations from the total Norwegian population less of a problem. In fact, we include more participants in this thesis than 25 of the 26 studies on blood pressure prediction reviewed in the paper by Sun et al. (2017).

The first issue we noticed when presenting the fitted models in Chapter 5.1 is that there are small differences between the predictions from the different models. Since the small versions only contain PAI and the explanatory variables that were found to be significant on a 0.05 significance level in the full models, it is as expected that the small and full versions of the same GLM are quite similar. The similarities between the Gaussian and the gamma models are not surprising either. One reason for this is that the models include the same explanatory variables. The other reason is that a gamma distribution goes towards a Gaussian distribution, with identical mean and variance, when the shape parameter goes towards infinity (Leemis and McQueston, 2008). The estimated shape parameter $\hat{k}$ for both the small and the full gamma model is approximately 88, which is quite large.

In Figures 5.3b and 5.5 we observe that the standard deviation of the residuals of the small gamma model depends approximately linearly on both the predicted systolic blood pressure and the explanatory variables. Since the residuals are shown to be very similar in Figure 5.2, it is reasonable to assume that this trend is true for the residuals from all the prediction GLMs. This trend deviates from the assumptions in the Gaussian GLM, where the standard deviation of the residuals is assumed to be constant and independent of the predicted response and the explanatory variables, see Chapter 4.1.1. On the other hand, the gamma GLM with identity link does assume a linear dependence between the standard deviation of the residuals and both the predicted systolic blood pressure and the linear predictor, see Chapter 4.1.3. In Figure 5.5 the sign of the slope of the dependence between the standard deviation of the residuals and the explanatory variables matches the sign of the corresponding regression coefficient. However, the values of the slopes do not correspond to the values of the corresponding regression coefficients. An example of this is that we see the biggest difference in the standard deviation of the residuals for birth year, but the largest absolute regression coefficient belongs to the systolic blood pressure from HUNT2. In other words, the standard deviation of the residuals does not completely fit the model assumptions of either prediction GLM.

We move on to discuss the performance of the models. Based on Figure 5.1 and an RMSE of approximately 13.7 for all the prediction GLMs, it is clear that the prediction models proposed by us do not give accurate predictions of the systolic blood pressure at HUNT3. The prediction distributions for individual participants in Figure 5.6 indicate that the variance in each individual prediction distribution is relatively large. This is confirmed in Figure 5.7. As a consequence, it is reasonable to assume that even though the observed value is far from the predicted value, the observed value often lies well within the prediction distribution. In spite of a large RMSE we find from CRPS and Brier score coverage that the prediction distributions reflects the uncertainty in the predictions.

A possible reason for the poor accuracy and large individual uncertainty is that 11 years is a too long time period to assume a constant lifestyle. In our prediction models, we only include explanatory variables with information from HUNT2, and the lifestyle might have changed drastically during the 11 years before the prediction time point at HUNT3. Following this line of argumentation, we would perhaps achieve better accuracy if we included explanatory variables containing information about the lifestyle of the participants at several times during the 11 years. Another possible reason is that 11 years may be simply too long to be able to accurately predict blood pressure. Life itself is unpredictable, and perhaps there are too many factors that affect a person's life and blood pressure in 11 years for it to be possible to include all the relevant explanatory variables in a general prediction model and expect accurate predictions for most individuals.

This line of thought may also explain the surprisingly small observed effect of physical activity level on the predicted systolic blood pressure. There are some notable effects of physical activity level, measured by PAI level, such as the fact that the High level of PAI is significant on a 0.05 significance level in the small Gaussian and small gamma model, the standard deviation of the residuals decrease slightly as PAI increases and the few participants with the lowest observed systolic blood pressure all have a High PAI level, see Figure 2.16. Otherwise, the data exploration analysis in Chapter 2 and the small regression coefficient of PAI in Chapter 5 shows that PAI has a small effect on the predicted systolic blood pressure.

As discussed previously, the small effect of physical activity may be due to the fact that lifestyle, including physical activity level, probably changes during the 11 years between the surveys. We may see a bigger effect of physical activity if we included information about the physical activity level of the participants at several time points between the studies. According to this argument we would expect better results if, as in the MyMDT study, the current PAI score was found through wearable sensors. It is also important to note that the PAI level used in this thesis is calculated from self-reported variables on physical activity during the year before HUNT2. In other words, the PAI level used here is at best a rough estimate of the participants' physical activity, and at worst may be incorrect due to poor reporting. Eventually, it might just be that physical activity just isn't that important for the systolic blood pressure, and that there are other factors that are more important, such as age, BMI, and parental hypertension.

In addition to evaluating the models' performance on predicting the systolic blood pressure at HUNT3, we study their ability to classify the systolic hypertension status at HUNT3. The sensitivity score at approximately 15% shows that the models are quite poor at detecting the participants who become systolic hypertensive. Previously in this chapter we have discussed the large uncertainty in the prediction distributions, and in Figure 5.8 we observe that there are quite a few participants with just below 50% predicted probability of systolic hypertension. This explains why the Brier score is relatively reasonable, even though the sensitivity is poor. On the other hand, the C-statistic at approximately 78% for all the models is quite good, which means that most participants who become systolic hypertensive have a higher predicted probability of systolic hypertension than the participants who remain healthy.

We want to compare the performance of the prediction GLMs proposed by us with the performance of the Framingham model. Firstly, we compare the performance of our models to the performance of the modified Framingham model applied to our cohort. From Table 5.3 we see that the results are similar, but that the modified Framingham model estimates even lower probabilities of hypertension than the GLMs. A possible explanation for the lower estimates of the modified Framingham model is that the model was originally meant to predict the 4-year risk of hypertension, and we have only altered it slightly before using it to predict the 11-year risk of hypertension. On the other hand, the Framingham model, both the original and the modified version, predicts the probability of general hypertension, not systolic hypertension. Therefore, it would be reasonable to expect higher estimates, since systolic hypertension is a special case of general hypertension. The slightly worse results might also be caused by the fact that the Framingham model was not created for this population, and in previous external validations it has had poorer results for certain populations (Sun et al., 2017).

Secondly, we want to compare the performance of our models with the original Framingham model proposed by Parikh et al. (2008). In the original paper Parikh et al. (2008) reports a C-statistic of 78.8%, with the 95% confidence interval (73.3,80.3). This is just slightly higher than the C-statistics of the models we implement, which lie just below 78 % (see Table 5.3). According to Sun et al. (2017), the Framingham model has been externally verified by 7 studies from different countries. Consequently, the fact that our models produce similar C-statistics to this model is an indicator that, at least in this aspect, the GLMs and modified Framingham model perform well on the HUNT Study Data.

Parikh et al. (2008) also reported a Hosmer-Lemeshow chi-square statistic of 4.35 for the original Framingham model. Proposed by Hosmer and Lemeshow (1989), the Hosmer-Lemeshow chi-square statistic is a common evaluation method for blood pressure prediction models. However, we do not calculate the Hosmer-Lemeshow chi-square statistic for the prediction GLMs or the modified Framingham model. The reason for this is that Parikh et al. (2008) state that they use a modified version of the Hosmer-Lemeshow chi-square statistic, without specifying how they modified it. Consequently, we do not know how to calculate a Hosmer-Lemeshow chi-square statistic that would be comparable to the statistic given in the paper.

6.2 Conclusion

The goal of this thesis is to predict the systolic blood pressure at the time of HUNT3 for people with initially healthy blood pressure at HUNT2, based on HUNT2 Study data. We also evaluate the models' ability to classify the systolic hypertension status at HUNT3. In addition to this, we study the effect of the physical activity measurement PAI, proposed by Nes et al. (2017), on the predicted systolic blood pressure. To reach this goal we present, implement and evaluate a small and full version of a Gaussian GLM and a gamma GLM, and a modified version of the Framingham model. Ultimately, we compare the performances of the models to each other and to the performance of the original Framingham model. The performances of the GLMs are very similar, and none of them perform significantly better than the others. Comparing the GLMs to the Framingham model, they perform slightly better than the modified version applied to the same cohort, and has a marginally lower C-statistic than the C-statistic reported in the original Framingham paper (Parikh et al., 2008). Physical activity, measured in PAI, is observed to have a surprisingly small effect on the predicted systolic blood pressure.

We conclude that the prediction models we propose are able to identify some clear trends in the data, for instance, the importance of birth year and previous systolic and diastolic blood pressure. Furthermore, they generally predict a higher probability of systolic hypertension for the participants who become systolic hypertensive, and have a C-statistic similar to C-statistic of the original Framingham model by Parikh et al. (2008). However, the variances in the individual prediction distributions are large and the models are not able to accurately predict the systolic blood pressure at HUNT3.

6.3 Future work

We identify two possible directions for future work. Firstly, we observe that the variance of the residuals depends on the predicted systolic blood pressure and the explanatory variables in a different way than described by the model assumptions of either prediction GLM. A suggestion for future work is thus to choose a prediction model that models the variance. Another suggestion for future work is to include information about the lifestyle of the participants, for instance, physical activity level, from several time points between HUNT2 and HUNT3. This might help since it is 11 years between HUNT2 and HUNT3, which is a long time to assume a constant lifestyle. A simple version of this would be to include lifestyle variables from HUNT3, such as PAI and BMI, as explanatory variables. Another possibility is to include lifestyle information from wearable sensors, like they do in the MyMDT project (NTNU, 2020).

Bibliography

Angus, J.E., 1994. The probability integral transform and related results. SIAM Review 36, 652–654. doi:`10.1137/1036146`.

Ardern, C.I., Janssen, I., Ross, R., Katzmarzyk, P.T., 2004. Development of health-related waist circumference thresholds within bmi categories. Obesity Research 12, 1094–1103. doi:`10.1038/oby.2004.137`.

Bouttier, F., 1994. Sur la prévision de la qualité des prévisions météorologiques. Ph.D. thesis. Toulouse 3.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Monthly weather review 78, 1–3. URL: `https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2`.

Brown, T.A., 1974. Admissible scoring systems for continuous distributions. .

Chobanian, A.V., Bakris, G.L., Black, H.R., Cushman, W.C., Green, L.A., Izzo, J.L., Jones, D.W., Materson, B.J., Oparil, S., Wright, J.T., Roccella, E.J., null null, 2003. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. Hypertension 42, 1206–1252. doi:`10.1161/01.HYP.0000107251.49515.c2`.

Cornelissen, V.A., Smart, N.A., 2013. Exercise training for blood pressure: A systematic review and meta&#x2010;analysis. Journal of the American Heart Association 2, e004473. doi:`10.1161/JAHA.112.004473`.

Cui, B., 2020. DataExplorer: Automate Data Exploration and Treatment. URL: `https://CRAN.R-project.org/package=DataExplorer`. r package version 0.8.1.

Ernstsen, L., Rangul, V., Nauman, J., Nes, B.M., Dalen, H., Krokstad, S., Lavie, C.J., Blair, S.N., Wisløff, U., 2016. Protective effect of regular physical activity on depression after myocardial infarction: The hunt study. The American Journal of Medicine 129, 82 – 88.e1. doi:`https://doi.org/10.1016/j.amjmed.2015.08.012`.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Generalized Linear Models. Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 269–324. doi:10.1007/978-3-642-34333-9_5.

Franklin, S.S., Gustin, W., Wong, N.D., Larson, M.G., Weber, M.A., Kannel, W.B., Levy, D., 1997. Hemodynamic patterns of age-related changes in blood pressure. Circulation 96, 308–315. doi:10.1161/01.CIR.96.1.308.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. Radiology 143, 29–36. doi:10.1148/radiology.143.1.7063747. pMID: 7063747.

Harrell Jr., F.E., Lee, K.L., Mark, D.B., 1996. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine 15, 361–387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

He, J., Whelton, P., 1999. Elevated systolic blood pressure as a risk factor for cardiovascular and renal disease. Journal of hypertension. Supplement : official journal of the International Society of Hypertension 17, S7—13. URL: http://europepmc.org/abstract/MED/10465061.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting 15, 559–570. doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hosmer, D.W., Lemeshow, S., 1989. Applied Logistic Regression. John Wiley & Sons.

Jordan, A., Krüger, F., Lerch, S., 2019. Evaluating probabilistic forecasts with scoringRules. Journal of Statistical Software 90, 1–37. doi:10.18637/jss.v090.i12.

Keys, A., Fidanza, F., Karvonen, M.J., Kimura, N., Taylor, H.L., 1972. Indices of relative weight and obesity. Journal of Chronic Diseases 25, 329 – 343. doi:https://doi.org/10.1016/0021-9681(72)90027-6.

Kieffer, S.K., Croci, I., Wisløff, U., Nauman, J., 2019. Temporal changes in a novel metric of physical activity tracking (personal activity intelligence) and mortality: The hunt study, norway. Progress in Cardiovascular Diseases 62, 186 – 192. doi:https://doi.org/10.1016/j.pcad.2018.09.002.

Lalkhen, A.G., McCluskey, A., 2008. Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia Critical Care Pain 8, 221–223. doi:10.1093/bjaceaccp/mkn041.

Leemis, L.M., McQueston, J.T., 2008. Univariate distribution relationships. The American Statistician 62, 45–53. doi:10.1198/000313008X270448.

Matheson, J.E., Winkler, R.L., 1976. Scoring rules for continuous probability distributions. Management Science 22, 1087–1096. URL: http://www.jstor.org/stable/2629907.

Nes, B.M., Gutvik, C.R., Lavie, C.J., Nauman, J., Wisløff, U., 2017. Personalized activity intelligence (pai) for prevention of cardiovascular disease and promotion of physical activity. The American Journal of Medicine 130, 328 – 336. doi:https://doi.org/10.1016/j.amjmed.2016.09.031.

NTNU, 2020. My Medical Digital Twin. https://www.ntnu.no/cerg/mymdt. Accessed on 08.05.2020.

Parikh, N., Pencina, M., Wang, T., Benjamin, E., Lanier, K., Levy, D., D'Agostino, R., Kannel, W., Vasan, R., 2008. A risk score for predicting near-term incidence of hypertension: The framingham heart study. Annals of internal medicine 148, 102–110. doi:10.7326/0003-4819-148-2-200801150-00005.

Paz, M.A., de La-Sierra, A., Sáez, M., Barceló, M.A., Rodríguez, J.J., Castro, S., Lagarón, C., Garrido, J.M., Vera, P., de Tuero, G.C., 2016. Paz ma, de-la-sierra a, sáez m, et al. treatment efficacy of anti-hypertensive drugs in monotherapy or combination: Atom systematic review and meta-analysis of randomized clinical trials according to prisma statement. Medicine 95, e4071. doi:10.1097/MD.0000000000004071.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

RStudio Team, 2016. RStudio: Integrated Development Environment for R. RStudio, Inc.. Boston, MA. URL: http://www.rstudio.com/.

Signorell, A., et al., 2020. DescTools: Tools for Descriptive Statistics. URL: https://cran.r-project.org/package=DescTools. r package version 0.99.32.

Statistics Norway, S., 2018. Women and men in norway 2018. URL: https://www.ssb.no/en/befolkning/artikler-og-publikasjoner/_attachment/347081?_ts=1632b8bcba0. Accessed on 27.04.2020.

Sullivan, L.M., Massaro, J.M., D'Agostino Sr., R.B., 2004. Presentation of multivariate data for clinical use: The framingham study risk score functions. Statistics in Medicine 23, 1631–1660. doi:10.1002/sim.1742.

Sun, D., Liu, J., Xiao, L., Liu, Y., Wang, Z., Li, C., Jin, Y., Zhao, Q., Wen, S., 2017. Recent development of risk-prediction models for incident hypertension: An updated systematic review. PLOS ONE 12, e0187240. doi:10.1371/journal.pone.0187240.

UIO, 2014. Lecture in STK3100/4100, "Exponential family". URL: https://www.uio.no/studier/emner/matnat/math/STK3100/h14/lectures/lecture2.pdf. Accessed on 15.04.2020.

Unger, D., 1985. A method to estimate the continuous ranked probability score. preprints, in: Ninth Conf. on Probability and Statistics in Atmospheric Sciences, pp. 206–213.

Weisstein, E.W., 2020a. Gamma Distribution from mathworld –a wolfram web resource. URL: https://mathworld.wolfram.com/GammaDistribution.html. Accessed on 15.04.2020.

Weisstein, E.W., 2020b. Normal Distribution from mathworld –a wolfram web resource. URL: https://mathworld.wolfram.com/NormalDistribution.html. Accessed on 15.04.2020.

WHO, 2019. World Health Organization's webpage about hypertension. URL: https://www.who.int/news-room/fact-sheets/detail/hypertension. Accessed on 03.04.2020.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. URL: https://ggplot2.tidyverse.org.

Worldometer, 2020. Norway population (2020). URL: https://www.worldometers.info/world-population/norway-population/. Accessed on 27.04.2020.

Appendix

A1. Additional figures from EDA

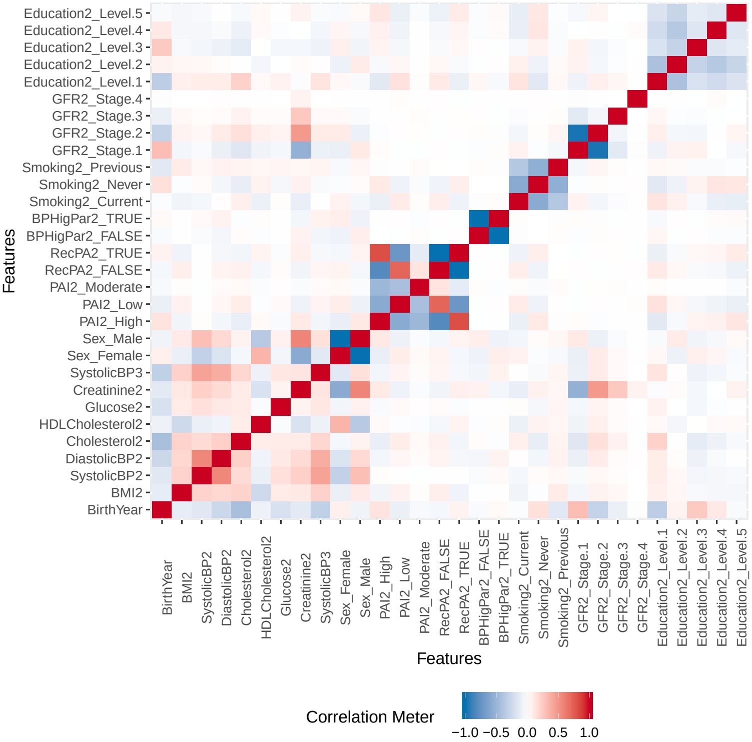


Figure 6.1: Correlation between all the explanatory variables from HUNT2, both categorical and continuous, and the response, ie. the systolic blood pressure from HUNT3.

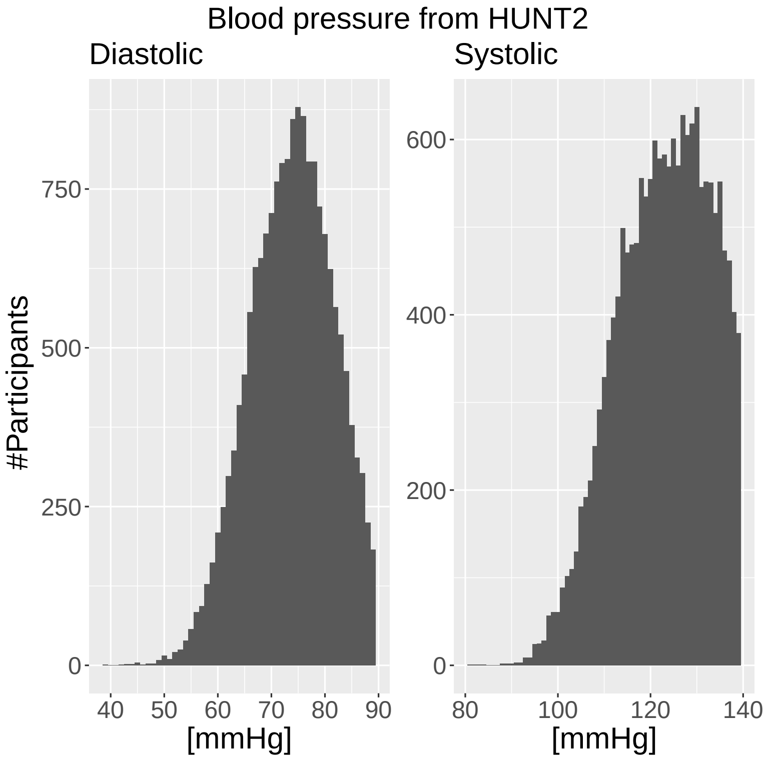


Figure 6.2: Distribution of diastolic (left) and systolic (right) blood pressure from HUNT2. Notice that we have removed all participants who were hypertensive at time of HUNT2, ie. diastolic blood pressure $>=$ 90 mmHg and/or systolic blood pressure $>=$ 140 mmHg.

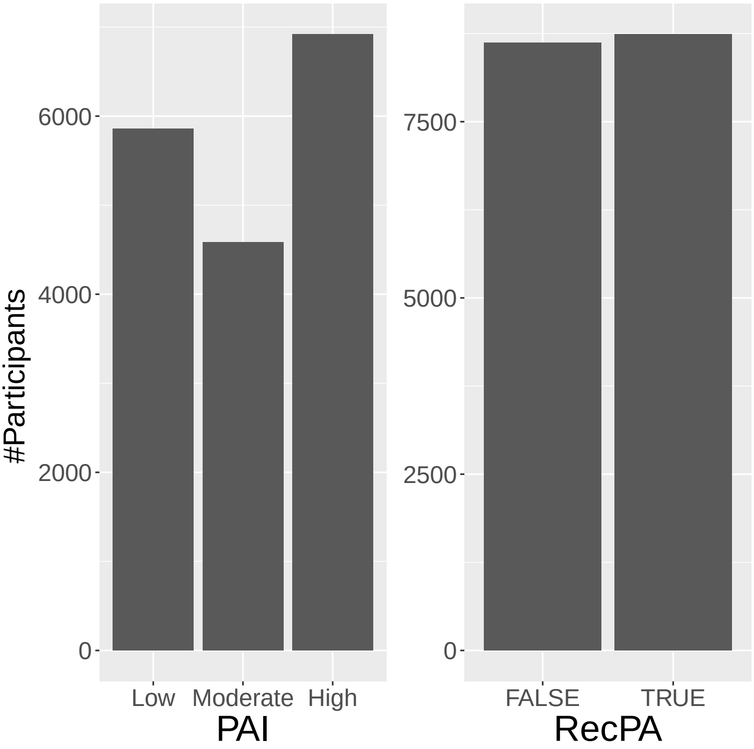


Figure 6.3: The distributions of physical activity of the participants at the time of HUNT2. To the left the PAI distribution is shown, and the RecPa distribution is shown to the right. See Section 2.1.3 for more details.

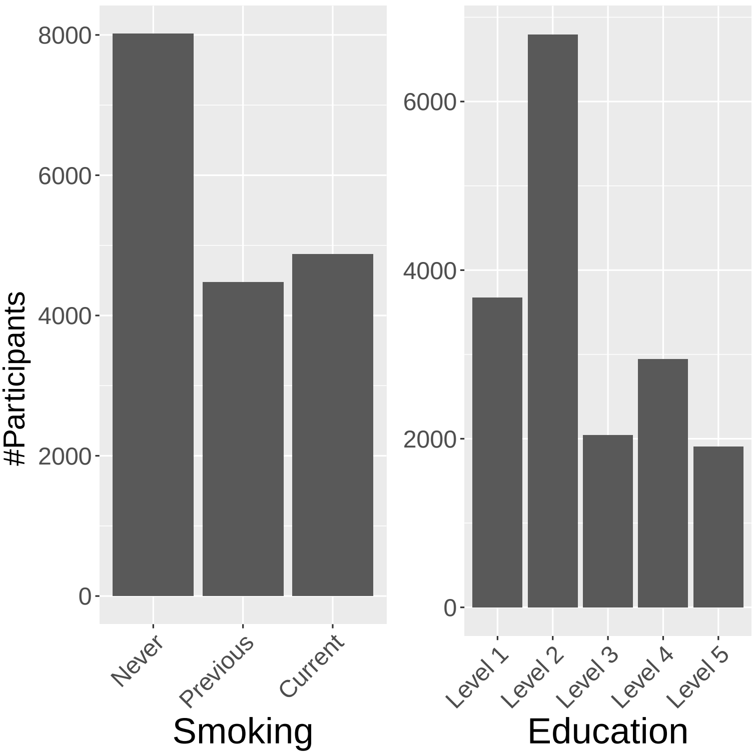


Figure 6.4: The distributions of the daily smoking habits (left plot) and highest education level (right plot) of the participants at the time of HUNT2. The levels of Smoking and Education are described in further detail in Section 2.1.3

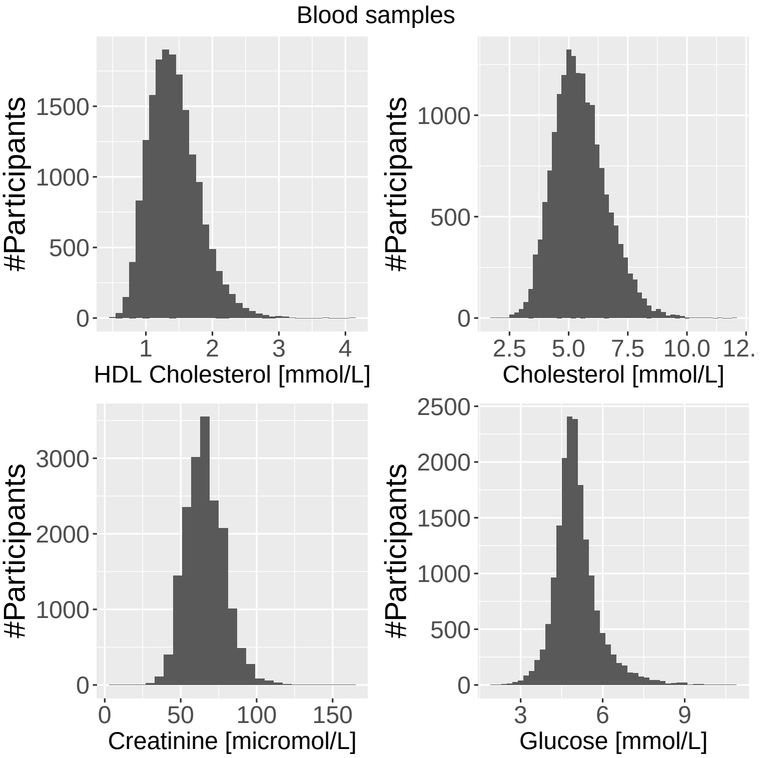


Figure 6.5: The distribution of the continuous explanatory variables from blood samples in HUNT2. Top left: HDL Cholesterol; Top right: Cholesterol; Bottom left: Creatinine; Bottom right: Glucose. See Section 2.1.3 for details.

NTNU

Norwegian University of
Science and Technology