NTNU
Norwegian University of
Science and Technology

Statistical analysis of factors influencing early neurological deterioration after acute ischemic stroke using sparse logistic regression

Marthe Larsen

Master of Science in Physics and Mathematics
Submission date: June 2016
Supervisor: Mette Langaas, MATH

Norwegian University of Science and Technology
Department of Mathematical Sciences

Problem description

The overall aim in the thesis is to analyze factors that influence early neurological deterioration after acute ischemic stroke based on data from the stroke unit at St. Olav's Hospital in Trondheim. The functional level of the patients is measured with the European Progressing Stroke Study scale and it is of interest to investigate how the scores are related to early neurological deterioration. We want to include both time-dependent variables and measurements on admission in a model where the response measures neurological deterioration, and we also want to use as much of the information from the stroke study in Trondheim as possible in a regression model. The data from the study in Trondheim contains both variables with missing values and variables with imputed values for missing data, and this problem must be considered prior to the statistical analyses.

Preface

This Master's thesis in Industrial Mathematics at the Department of Mathematical Sciences completes the Applied Physics and Mathematics Master's degree programme at the Norwegian University of Science and Technology (NTNU). The work has been carried out during the spring of 2016. Factors influencing early neurological deterioration after acute ischemic stroke are analyzed in this thesis. I want to thank my supervisor, Mette Langaas at the Department of Mathematical Sciences, for great guidance and feedback during the semester. I also want to thank Turid Follestad at the Department for Public Health and General Practice for being my co-supervisor, and Bernt-Harald Helleberg at the Department of Internal Medicine for guidance on medical terms and the study procedure.

Abstract

A total of 368 patients treated at the stroke unit at St. Olavs hospital in Trondheim were included in a study to analyze early neurological deterioration after acute ischemic stroke. Bad prognosis is associated with early neurological deterioration, and more research concerning the causes of early neurological deterioration is needed. In a preliminary analysis the time-dependent variables from the study were explored with unsupervised methods and quintile analysis. Principal component analysis and clustering were used to explore any possible groupings of the level of function measured with the European progression stroke study scale. The preliminary analysis led to using a binary response and different summary statistics for the time-dependent predictors. Potential factors influencing early neurological deterioration were analyzed with the lasso-penalized logistic regression method. Lasso regression uses a l_1-penalty to shrink parameters to zero, and preforms variable selection automatically. Also, the lasso has no limitations on the number of predictors and finds a sparse solution to complex problems. Lasso is a relatively new method and in particular developments towards statistical inference is still ongoing.

With the lasso method 22 predictors were included in the final model. Both variables measured on admission and time-dependent variables were estimated to be nonzero. The analysis also shows that both level and variability of the time-dependent predictors are important, so that monitoring patients closely the first few days after acute ischemic stroke is essential for the outcome. The penalty parameter controls the strength of the shrinkage and was chosen with cross-validation. Exact standard errors and confidence intervals for the regression parameters do not exist for the lasso method. Thus, inference about the regression parameters was performed using bootstrapping. The limitation of the lasso method is that it does not handle correlated variables very well, and the limitation is visualized and analyzed for the stroke data from Trondheim. In the field of medical statistics the lasso method has the potential to be very useful as it handles data with numerous predictors, and finds a good model for prediction.

Sammendrag

En analyse er gjort på 368 pasienter fra slagavdelingen på St. Olavs Hospital i Trondheim for å analysere nevrologisk forverring de første døgnene etter akutt hjerneinfarkt. Prognosene er dårligere for pasientene med forverring enn de som ikke opplever forverring de første døgnene, og det er nødvendig med mer omfattende analyse av faktorer som har innvirkning på nevrologisk forverring. Som en innledende analyse har vi gjort klyngeanalyse, prinsipal komponentanalyse og kvintilanalyse på de tidsavhengige variablene. Klyngeanalyse og prinsipal komponentanalyse er gjort for å se om det er en naturlig inndeling av pasientene basert på målingene av kroppens funksjonsevne. Den innledende analysen resulterte i å bruke en binær responsvariabel og ulike oppsummeringsstatistikker for de tidsavhengige prediktorene. Faktorer som potensielt påvirker nevrologisk forverring de første døgnene etter akutt hjerneinfarkt ble analysert med en lasso-straffet logistisk regresjonsmodell. Lasso-metoden bruker en l_1-straff for å forminske parametere til null, og utfører derfor variabelseleksjon automatisk. I tillegg takler metoden høydimensjonerte data og finner en modell med relativt få regresjonsparametere som ikke er estimert til null. Lasso er en relativt ny metode og utvikling spesielt innenfor statististik inferens pågår fortsatt.

Lasso-metoden inkluderte 22 prediktorer i den endelige modellen. Både variabler basert på målinger ved innleggelse og variabler basert på målinger over tid er med i modellen. Analysen viser også at både nivå og variabilitet i de tidsavhengige prediktorene er viktig, og derfor er det nødvendig med nøye oppfølging av pasientene de første døgnene etter akutt hjerneinfarkt. Straffeparametret kontrollerer graden av krymping og kryssvalidering er brukt for å finne den optimale verdien. Eksakte verdier for standardavviket og konfidensintervallet til de ulike regresjonsparametrene finnes ikke for lasso-metoden. Inferens av regresjonsparametrene er derfor basert på bootstrap-metoden. Ulempen med lasso-metoden er at den ikke inkluderer flere korrelerte variabler i den endelige modellen, og det er tilfeldig hvilken av de korrelerte variablene som blir inkludert i modellen. Denne ulempen er visualisert og analysert nærmere med slagdataene fra Trondheim. Lasso-metoden har potensialet til å bli nyttig innenfor medisinsk forskning, da metoden takler data med mange prediktorer og finner en god modell for prediksjon som er enkel å fortolke.

Contents

List of abbrevations

ASPECTS	Alberta Stroke Program Early CT Score
AIS	Acute ischemic stroke
BS	Blood sugar
CRP	C-Reactive protein
DBP	Diastolic blood pressure
EDE	Early deterioration episode
END	Early neurological deterioration
EPSS	European Progressing Stroke Study
EPV	Events per variable
LACI	Lacunar infarct
LASSO	Least Absolute Shrinkage and Selection Operator
LOCF	Last observation carried forward
MAR	Missing at random
MCAR	Missing completely at random
MNAR	Missing not at random
NIHSS	National Institutes of Health Stroke Scale
PACI	Partial anterior circulation infarct
POCI	Posterior infarct
SBP	Systolic blood pressure
SSS	Scandinavian stroke scale
TACI	Total anterior circulation infarct
TEMP	Body temperature
TIA	Transient ischemic attack

Chapter 1

Introduction

On a worldwide basis, 15 million people suffer a stroke every year. Almost six million of these people die and five million people are left disabled. Stroke is the second most common cause of death (Donnan et al., 2008). In Norway, 15 000 people suffer a stroke and almost 3100 die of stroke every year. Ischemic stroke and hemorrhagic stroke are the two main stroke types. Ishcemic stroke accounts for 85-90% of all stroke cases and occurs as a result of an obstruction within a blood vessel supplying blood to the brain. As a result, the blood flow to the brain is completely or partly blocked. Ischemic stroke can be divided into cerebral embolism and cerebral thrombosis. The prognosis of stroke depends on the stroke type, but for ischemic stroke one third of the patients will have the same body function as before the stroke. Both inheritance and lifestyle can contribute to the cause of stroke, and well-known exposure factors are high blood pressure, smoking, degree of alcohol consumption, high cholesterol, diabetic, inactiveness and obesity. Most patients gradually recover over days, weeks and months but patients can also deteriorate. The deterioration have different causes and it is incompletely understood.

Early neurological deterioration (END) is clinical worsening during the first 72 hours after an acute ischemic stroke. Despite the bad prognosis for patients with END, it is not until recently possible predictors of END have been studied. Many issues are unresolved and more research regarding predictors of END is needed. Another important aspect of the study of END after an acute ischemic stroke (AIS) is that available studies have used inconsistent definitions and time frames so that the findings are not easy to generalize when it comes to clinical guidance. In addition, due to the aging population in Norway the number of stroke incidences are expected to increase and it is more important than ever to optimize the treatment guidance after stroke. Factors influencing END after AIS is analyzed in this thesis and the data comes from a study of AIS patients conducted in Trondheim. The aim is to get a better understanding of END and to identify factors that are useful for predicting END.

Chapter 2 contains a presentation of the data and an explanation of the medical terms we encounter during the analyses. Understanding the study procedure and medical terms are important for the statistical analyses, and also useful when interpreting the result from the analyses. Chapter 2 also contains a discussion of certain aspects and limitations of the data. Chapter 3 contains a preliminary analysis of the time-dependent variables and a presentation of two unsupervised learning methods. The main statistical theory on sparse modeling is presented in Chapter 4. Both traditional statistical methods and also how it is adapted to a new and evolving method for model selection are presented in Chapter 4. Results from application of the statistical methods to the Trondheim early neurological deterioration study are presented and analyzed in Chapter 5, and Chapter 6 summarizes and discusses the medical and statistical findings in this thesis. The statistical analyses are done `R-Studio 0.98.1103` (R Core Team, 2015). The core part of the `R`-code used in the statistical analyses is found in Appendix A.

Chapter 2

The Trondheim early neurological deterioration study

The data comes from a study conducted within the stroke unit at St. Olav's Hospital in Trondheim, and the following presentation of the data is based on the study protocol of Helleberg et al. (2014). The stroke unit has a long experience of treating stroke patients in both the acute phase and early rehabilitation phase and has a personnel specialized in stroke therapy. On average, 325 patients per year have been discharged from the hospital with a diagnosis of ischemic stroke. A total of 368 patients from the time period May 2010 to December 2013 treated at St. Olav's Hospital are included in the study. Initially 401 patients were included, but 39 patients were excluded due to exclusion criteria and another 6 patients were added to the study with data from a pilot study performed in 2009. Follow-up for the last patient was complete in April 2014. The final inclusion criteria stated in the study protocol is that the patients had to be admitted to the stroke unit with acute stroke symptoms, admitted to the stroke unit within 24 after the stroke and previously living in their own home. The exclusion criteria were previously known preexisting condition which could confound follow-up, diagnosis other than acute ischemic stroke that could lead to the same symptoms, no capacity to follow the patient, consent could not be achieved and heamorrhage on native CT examination. The patients included in the study are managed according to current procedures and national guidelines as any other patient experiencing stroke, and the length of the stay and the treatment decisions were not affected by inclusion in the study.

The study design is a single-center prospective observational study. As opposed to an experimental study design where the researcher intervenes to change reality, the researcher studies what occurs and do not alter the study in an observational study. Every patient is exposed to the same treatment and measurements and the outcome is observed. In this setting, prospective means that the design of the study and the recruitment of patients are done before any of the patients have developed the outcome of interest. A single-center study is conducted at one location, and has some limitations compared to multi-center

studies. In multi-center studies data from different locations is used and better represents the general population. However, a single-center study is still very useful for clinical guidance at the specific location and comparing results from other studies are of interest.

People are affected by stroke in different ways. Both the symptom combination and effect of stroke differs from person to person. Common stroke symptoms are sudden loss in level of consciousness, facial droop, changes in hearing or taste, confusion or loss of memory, vertigo, loss of coordination, muscle weakness in arm or leg (usually on one side), emotional changes and trouble in speaking (Knator, 2015). The effects of a stroke depend on the location and the degree of affected brain tissue. For some people the effects are relatively minor while others are left with serious long term problems. The most noticeable effects are problems with movement and balance, problems with vision, problems controlling the bladder and bowels and excessive tiredness. However, stroke also causes hidden effects like problems with communication, problems with memory and changes of the behavior. The main outcome of interest in this analysis is the early neurological deterioration effect of stroke. Relevant definitions and related measurement scales for the level of function will be presented in the next sections.

2.1 Measurement scales

On admission to the stroke unit the level of function is measured to say something about the severity of the stroke. In addition, the level of function is measured frequently during hospitalization to say something about the neurological improvement or deterioration. Birschel et al. (2004) discuss the issue that several scales for measuring the level of function measurement exist and that there are different definitions of neurological deterioration. Stroke scales are useful when it comes to the diagnostic accuracy in the clinical routine settings. In the mid-1990s a collaboration was set up to standardize the terminology, classifications, clinical assessments and outcome measures of stroke (Birschel et al., 2004). The aim was to create a common clinical language to use in stroke studies. Different types of scales are of course needed to capture all the effects of stroke, but scales with the same purpose should be standardized and no single scale is suitable for all research situations.

2.1.1 Scandinavian Stroke Scale

The Scandinavian Stroke Scale (SSS) ranges from 0 to 58 points and measures a patient's condition after a stroke. The scale has nine items and quantifies the level of consciousness, eye movements, arm movements, hand movements, leg movements, language, orientation, gait and facial palsy. In general, higher score means higher level of function but the different items have different maximum score. The scale is a simple stroke scale and the rating can be performed in less

than 5 minutes (Christensen et al., 2005). This aspect of the scale is important in the acute phase of stroke.

2.1.2 European Progressing Stroke Study

The European Progressing Stroke Study (EPSS) group was a subgroup of the collaboration working with definitions of deterioration, improvement and progression based on clinical assessments. They decided to use five of the nine items from SSS. The exclusion of four of the categories was done to maximize the reliability and to make it easier to be repeated by nursing staff every few hours during the first three days. The EPSS scale includes the items level of consciousness, conjugate gaze, speech and motor function in the affected arm and leg (Birschel et al., 2004). The scale has been incorporated at several stroke units as the standard measurement scale. The EPSS scale ranges from 0 to 32 points and as for the SSS, higher score means higher level of function. Both the sum of the score from all the five items and the points in each separate item is of interest when analyzing the data, but only the sum is used in the statistical analysis of EPSS in Section 3.3.

2.2 Early neurological deterioration

Early neurological deterioration (END) is defined as clinical worsening during the first 72 hours after an ischemic stroke. The short term and long term consequences of END is associated with a worse functional outcome and higher mortality rate (Thanvi et al., 2008). Identifying predictors of END can help to prevent the condition because of early treatment. There are several causes of END and no single intervention benefits all patients. However, the treatment in a stroke unit is associated with reduced risk of END and recurrent stroke but it is not known if it also reduces the impact of END (Govan et al., 2007). This is why analyzing END is of great clinical importance, and there is still many unanswered questions related to acute ischemic stroke and END.

Due to different stroke scales during many years of medical research, END also has different definitions. In the Trondheim early neurological deterioration study, END is defined according to the EPSS scale. END is either a decrease during the first 72 hours of 2 or more SSS points in the conscious level, gaze or movement level, or a change of 3 or more SSS points in the language level. Consciousness was given precedence over the other measurements of functional level.

A table of a selection of baseline characteristics in terms of patients with END and patients without END (no END) can be seen in Table 2.1. From the table it can be seen that in the Trondheim early neurological deterioration study 13.8% of the patients experienced END. The number is in agreement with other studies, but the percentage is dependent on the definition used (Thanvi et al., 2008). Of the patients with END 24% died and of those with no END, only

4.3% died. The variable END is discussed and analyzed further in Section 3.1 and Chapter 5.

2.3 Early deterioration episode

Early deterioration episode (EDE) is defined in accordance with the EPSS definition of neurological deterioration. In contrast to END, EDE is only based on the change between two consecutive assessments. Birschel et al. (2004) say that the EPSS definition of EDE has a good prognostic validity, and that EDE happens more frequently than END. EDE and its relation to END will be discussed in more detail in Section 3.3.

Table 2.1: Baseline characteristics of patients with END and patients with no END. The numbers shown are either the mean with the corresponding standard deviation or the total number of patients with the given characteristic and the corresponding percentage.

	No END (n=317)	END (n=51)
Male	177 (55.8%)	25 (49.0%)
Female	140 (44.2%)	26 (51.0%)
Age(years), mean ± SD	76.01 ± 8.99	79.71 ± 7.94
History of hypertension	185 (58.4%)	17 (33.3%)
History of diabetes	45 (14.2%)	9 (17.6%)
History of stroke or TIA	98 (30.9%)	16 (31.8%)
History of atrial fibrillation	89 (28.1%)	24 (47.6%)
Initial SBP (mmHg), mean ± SD	145.73 ± 11.14	153 ± 15.36
Initial DBP (mmHg), mean ± SD	67.91 ± 10.88	86.27 ± 13.46
Thrombolytic treatment	84 (26.5%)	15 (29.4%)
Statins	234 (73.8%)	27 (52.9%)
Temperature (°C), mean ± SD	37.00 ± 0.53	37.28 ± 0.65
Blood sugar (mmol/l), mean ± SD	6.35 ± 1.65	7.02 ± 2.07
Kidney function (ml/min/1.73 m^2), mean ± SD	71.19 ± 18.83	65.28 ± 17.98
Potassium level (mmol/l), mean ± SD	4.00 ± 0.39	4.19 ± 0.41
CRP (mg/l), mean ± SD	11.28 ± 22.56	8.92 ± 15.65
Very severe stroke	24 (7.57%)	12 (23.5%)
Severe stroke	29 (9.15%)	15 (29.4%)
Moderate stroke	127 (40.1%)	17 (33.3%)
Mild stroke	84 (26.5%)	5 (9.80%)
Very mild stroke	53 (16.7%)	2 (3.92%)

2.4 Predictors of interest

Frequent neurological assessments, blood sample measurements, repeated imaging and continuous monitoring are performed in order to analyze early neurologi-

cal deterioration. The patients in the Trondheim early neurological deterioration study (Trondheim END study) were followed for 3 months, but measurements from 0-72 hours after being hospitalized are used in the analyses. A presentation of predictors that is of interest when analyzing early neurological deterioration is included in this section and is compared to results from other stroke studies.

2.4.1 Age

Age is the principal non-modifiable risk factor for stroke, and the stroke rate increases significantly with age for both men and women (Sacco et al., 1997). Half of all strokes occur in people over the age of 75, and one-third in the population over the age of 85 (Falcone and Chong, 2007). A histogram of the age distribution of patients in the Trondheim END study can be seen in Figure 2.1. The youngest person included in the study is 54 and the oldest is 95. In comparison to the percentages above, 56% of the patients in the Trondheim early neurological deterioration study is over age 75 and 17% over age 85. In addition, Table 2.1 shows that patients with END is on average older than patients with no END.

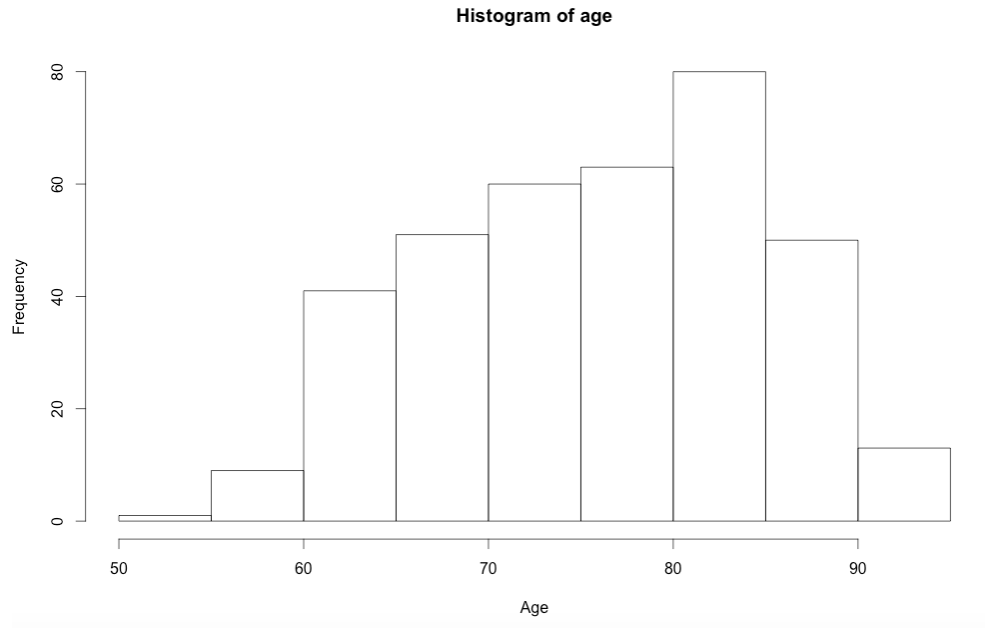

Figure 2.1: Histogram of the age of the patients in the Trondheim END study.

2.4.2 Gender

Similar to age, gender is an important non-modifiable risk factor for stroke and is reasonable variable to include in a statistical analysis of END. The gender

differences in stroke are complex, and there are differences in the incidence of both stroke and END in age subgroups (Falcone and Chong, 2007). The death rates are in general lower in women than in men, but the functional outcome is higher for women. However, women are also older when presenting with first stroke. From Table 2.1 it can be seen that 55% of the patients in the Trondheim END study were men, 45% were women and that 12% of the men experienced END and 19% of the women experienced END. In addition, the mean age for women were higher than the mean age for men and probably explains the higher rate of END.

2.4.3 Stroke severity

Stroke severity can be measured in different ways. One way is based on the measurement of the SSS on admission and the scale is divided into five categories. A very severe stroke has a SSS score on 0-14, a severe stroke has a SSS score on 15-29, a moderate stroke has a SSS score on 30-44, a mild stroke has a SSS score on 45-51 and a very mild stroke has a SSS score on 52-58. From Table 2.1 it can be seen that most of the patients in the Trondheim END study experienced a moderate stroke. Stroke severity seems to be related to END since 50% of the patients with a very severe stroke experienced END and only 4% of the patients with a very mild stroke experienced END. The study of Thanvi et al. (2008) is one of several studies that found that initial stroke severity increases the risk of END. Neither the continuous SSS measurements or the categorical variable with five levels are used in the statistical analysis in Chapter 4. In stead, the SSS score is divided into three categories is used due to the log linear assumptions of the predictors in logistic regression.

2.4.4 Blood sugar and body temperature

Hyperglycemia is defined as blood sugar level $>$ 6 mmol/L, and is common in the early phase of stroke. Two thirds of all ischemic stroke patients have hyperglycemia on admission and an increasing number of studies have found that blood sugar is associated with functional outcome (Lindsberg and Roine, 2011). Temperature is also a factor of interest when it comes to neurological outcome after acute ischemic stroke. Approximately one half of patients hospitalized for stroke develop fever, and clinical studies have found that high body temperature is associated with neurological outcome (Wrotek et al., 2014).

In the Trondheim END study, temperature and blood sugar were measured every 6^{th} hour during the first 48 hours, and after 60 and 72 hours. Temperature and blood sugar vary between these time points and both are time-dependent variables. The number of patients with hyperglycemia on admission in the Trondheim END study is 184, and from Table 2.1 it can be seen that the average blood sugar measurement on admission is higher in the patients with END. Table 2.1 also shows that the mean temperature on admission is higher for the patients

with END compared to the patients with no END. How temperature and blood sugar are related to END in the Trondheim early neurological deterioration study will be analyzed further in Section 3.1 and Chapter 5.

2.4.5 Blood pressure

High blood pressure is the most important risk factor for stroke, and it is of great interest to look more closely into blood pressure in the analysis of END after acute ischemic stroke. The role of long-term blood pressure control to improve the outcome in patients with stroke is undisputed, but the management of the blood pressure immediately after a stroke is controversial (Aiyagri and Gorelick, 2009). Several studies have looked at the effect of blood pressure level on the outcome after stroke and some of the results are inconsistent.

Similar to blood sugar and temperature, blood pressure is a time-dependent variable. In the Trondheim END study, the systolic blood pressure (SBP) and the diastolic blood pressure (DBP) are measured 11 times. Five times the first day, four times the second day and three times the third day. In Table 2.1 it can be seen that both the mean diastolic and the mean systolic initial blood pressure is higher for the patients with END than for the patients without END. In addition, 80% of the patients included in the study had high blood pressure (SBP $>$140 and DBP $>$90). Usually, the blood pressure decreases over the following days, and 63.9% of patients have lower blood pressure at 72 hours than at baseline. This is in fact what characterizes a stroke patient and it may be more interesting to look at blood pressure variability. Further analysis and graphical representation of different blood pressure parameters will be presented in Section 3.1 and Chapter 5. In addition to the systolic and diastolic blood pressure, pulse pressure on admission is also included in the statistical analysis in Chapter 5 and is the difference between the systolic and diastolic blood pressure.

2.4.6 Drugs

Thrombolytic drugs are used to dissolve blood clots and can be used in the immediate treatment of ischemic stroke and heart attack. This is called thrombolysis. Not all patients can get the treatment and the decision to give the drug is based upon a computerized tomography (CT) on admission to check for bleeding, degree of the stroke and medical history. If possible, thrombolytic drugs should be given within 3 hours of the stroke symptoms to help limit the possible disability, and a number of large trials have confirmed the benefits of the treatment in acute ischemic stroke (Bansal et al., 2004). However, the majority of the patients with acute ischemic stroke do not receive thrombolytic drugs due to late arrival to the emergency departments. From Table 2.1 it can be found that approximately 27% of the 368 patients in the Trondheim END study got the treatment. The small percentage is due to late arrival to the hospital and the extensive decision process that has to be done to be approved for the treatment.

From Table 2.1 it can also be seen that the percentage of patients receiving the treatment is only 3% higher for END than no END.

Statins is a group of drugs that are used to reduce cholesterol levels and have been found to decrease cardiovascular risk and to improve clinical outcome. In recent years, clinical trials looking at statins as a part of the treatment of acute ischemic stroke has increased (Zhao et al., 2014). From Table 2.1 it can be found that statins were given to 70% of the patients in the Trondheim END study, and the majority of the patients with no END were given statins.

2.4.7 Other predictors

In the Trondheim END study, both time from symptom onset to hospitalization and time from symptom onset to admission to the stroke unit are registered. This is called the prehospital delay time and would be interesting to include and explore in a statistical model. However, with approximately 25% missing values in the data and with no reasonable method to estimate the missing values this will not be done. However, it can be noted that with exclusion of the missing data, the mean prehospital delay time registered is approximately 4 hours. It can also be noted that the patients with END have 80 minutes shorter mean prehospital delay time compared to the mean time of the patients with no END. This may indicate that the prehospital delay time is related to the stroke severity and END.

Different blood sample measurements and medications given during the hospitalization are measured and could have been included in the statistical analysis. However, the majority of these variables are excluded in the analysis due to missing values. The blood sample measurements of potassium, glucose and C-reactive protein (CRP) on admission are included in the statistical analysis. Bazzano et al. (2001) suggest that low potassium intake is associated with an increased risk of stroke. The CRP level is a marker of inflammation in the body and a normal level is $<$10 mg/l. Data relating CRP to the prognosis after AIS are sparse, but Napoli et al. (2001) found that CRP is a marker of increased 1-year risk in ischemic stroke. Table 2.1 shows that the patients with END in the Trondheim END study had lower mean CRP level than the patients with no END. However, both have a high standard deviation. Kidney function on admission is also included in the analysis. Glomerular filtration rate (GFR) is a kidney function test. Normal levels ranges from 90-120 ml/min/1.73 m^2, but older people have lower GFR levels. A GFR lower than 15 ml/min/1.73 m^2 is a sign of kidney failure (Martin, 2015). From Table 2.1 it can be seen that the mean value of GFR in the Trondheim END study is lower for the patients with END than for the patients with no END.

The data also contains several binary variables with information about earlier or present conditions. History of stroke or transient ischemic attack (TIA), history of atrial fibrillation, history of ischemic heart disease, history of hypertension and history of diabetes mellitus are included in the statistical analysis.

Another binary variable included in the analysis is clinical/ASPECTS mismatch and is more complicated to classify. The Alberta Stroke Program Early CT Score (ASPECTS) is a measurement scale the radiologist uses to grade early CT-changes and ranges from 0 to 10. The National Institutes of Health Stroke Scale (NIHSS) is also used in the valuation. The scale ranges from 0 to 42 and has many of the same scoring categories as SSS and EPSS. A patient with an ASPECT score ≥ 8 combined with a NIHSS score ≥ 8 has clinical/ASPECTS mismatch. A NIHSS score ≥ 8 has been suggested to be used as a clinical indicator of large volume of ischemic brain tissue (Tei et al., 2007). Based on the ischemic stroke symptoms, the stroke episode can be classified as total anterior circulation infarct (TACI), partial anterior circulation infarct (PACI), lacunar infarct (LACI) or posterior infarct (POCI) (Tei et al., 2000). These variables are also included in the analysis in Chapter 5.

2.5 Quality of the data

An important part of the analysis of a data set is to investigate how the data was collected. When doing this, significant omissions or biases which may influence the analysis can be revealed. Procedures, definitions, measurements uncertainty etc. can differ from research location to research location, and it is important to keep in mind when comparing the result with published articles concerning the same field of interest. In general, anomalies should be investigated, but especially when doing statistical analysis in the field of medicine, anomalies should not be ignored. Often these anomalies provide useful information. Limitations of the data are important to state to use the result in a bigger context. The limitations and data quality of the Trondheim END study is presented below, and the theory is mainly taken from Little and Rubin (2002).

2.5.1 Missing data

Due to several different reasons some entries in a data set can be missing. In surveys the participants can for example refuse to answer some of the questions or they can be unable to choose between the given alternatives. In medical trials a patient can for example be too sick to go through with the planned measurements or the patient can refuse to continue in the study. Missing data can be handled by analyzing the available data and ignoring the missing values, by filling in the missing data with replacement values or by using statistical models to allow for the missing data and make assumptions about the relationship to the available data. Different methods exist to impute the values of the observations that are missing, and alternative procedures are constantly under development. Imputation can either be done by imputing one value for each missing item or by imputing more than one value to allow for uncertainty of the value. Multiple imputation is a risky procedure since it leads to a complete data set which in reality is not complete. How the missing data are handled can have a crucial influence on the final result and the certainty of the conclusion. There is no universal best approach and the method and assumptions should be connected to the nature and behavior of the variables in the study. On the contrary, simply removing the patients with missing data from the analysis will decrease the sample size and again result in a reduction of the statistical power, and useful measurements will be removed completely from the analysis. In addition, it is for example likely that excluding missing data would have been excluding patients that represents the healthier part of the stroke patients since missing data can occur when patients have left the hospital. This would lead to selection bias.

Both the pattern of missing data and the mechanisms that lead to missing data is important to consider prior to statistical analyses. If the complete data is defined to be $M = (m_{ij})$ and contains both the entries of the observed data, M_{obs}, and the entries of the missing components, M_{mis}, the missing data can depend on M, M_{obs} or none. When the missing data depends on the missing values in M_{mis}, the missing data is related to the data values and the mechanism

is called missing not at random (MNAR). On the other hand, if the missingness does not depend on M the data are called missing completely at random (MCAR). A less restrictive mechanism than MCAR is that the missing data does not depend on M_{mis} but only depends on M_{obs}, the data is called missing at random (MAR). Analyzing data missing at random as there were no missing data can give consistent and reliable results, but it is hard to obtain the same reliable results if the data are MNAR. This is due to the fact that the missing value contains important information and can not be ignored.

In a longitudinal study, each experimental or observational unit is measured at baseline and repeatedly over time. Incomplete data are not unusual under such designs, as many subjects are not available to be measured at all time points. In addition, a subject can be missing at one follow-up time and then measured again at one of the next, resulting in nonmonotone missing data patterns. Such data present a considerable modeling challenge for the statistician. It is also common that the subjects drop out prior to the final measurements and do not return which result in a monotone missing data pattern. For the Trondheim END study it is stated in the study protocol of Helleberg et al. (2014) that missing values can be retrieved or estimated from medical records and the former value is continued when estimation from clinical score sheet is not reliable. Scores may also be adjusted if there is inconsistency between the available clinical information and the value from the score sheet. In addition, patients discharged before the time limit on 72 hours were scored in accordance to their last measured values for the time-dependent variables. Often a data set is handed to a statistician with missing values and the statistician have to decide which imputation method to use. However, in the Trondheim END study the data set it complete and contains imputed values for missing measurements.

Last observation carried forward (LOCF) is a single imputation method and for each individual the missing values are replaced with the last observed value of that variable. As a result, a potential source of bias is introduced and variance in the data is most likely underestimated. In the Trondheim END study, LOCF is used when a measurement is missing but we have no information of the entries of the imputed values in the data set. For the majority of the time-dependent variables the LOCF is not so easy to justify. Blood sugar, temperature and blood pressure are expected to change over a six hour time interval and a measurement of these variables equal to the previous measurement is expected to be a LOCF-value. EPSS on the other hand, can be constant for stable patients and an imputed value is hard to distinguish from an observed value. Uncertainty about the score can also be a reason for LOCF-value in some of the EPSS entries. Especially for the high EPSS scores it is likely that some of the missing values are MNAR and a sign of improvement since it is likely that missing a measurement is due to a stable patient that is not bedridden or dischargement. In this case LOCF is reasonable. In addition, missing values are expected to some degree due to the fact that inclusion in the study should not affect the treatment given, and when the treatment provide no added benefit the patient is discharged.

Missing data is also a frequent problem in the variables not dependent on time. A reasonable value can probably be estimated based on other available information about the patient, but this is a comprehensive procedure depending on broad knowledge about medical conditions and association between clinical measurements.

It is assumed that most of the missing values are MNAR and excluding the patients with missing values will probably lead to selection bias. The sample size and the statistical power would also decrease dramatically. As an alternative solution to the problem, an algorithm to estimate the percentage of imputed values is made. The percentages is useful when discussing the strength of the statistical analysis. Values in possible unobserved entries are replaced with NA (Not available) and the EPSS values have the strictest NA-rule since it is possible that patients have a constant value over time. The algorithm is given by

- DBP - NA if the previous value is the same as the present value for the SBP and DBP
- SBP - NA if the previous value is the same as the present value for the SBP and DBP
- Blood sugar - NA if the previous value is the same as the present value for the blood sugar
- Temperature - NA if the previous value is the same as the present value and NA in blood sugar at the given position for the temperature
- EPSS - NA if the previous value is the same as the present for the EPSS and NA for the SBP and DBP at the given position for the EPSS.

Table 2.2: The estimated percentages of imputed values in the time-dependent variables based on the LOCF-algorithm above.

	LOCF
SBP	25.3%
DBP	25.3%
Temperature	11.4%
Blood sugar	16.0%
EPSS	28.8%

The percentage of the LOCF values for each variable from the algorithm is given in Table 2.2. This leads to a total of 21.4% LOCF values for the time-dependent variables. In the following statistical analysis the data with the imputed values will be used, and in Chapter 6 the issues with the LOCF data set will be discussed in a bigger context.

2.5.2 Limitations

The measured value of the blood pressure, blood sugar and temperature is sensitive to errors in the measurement tool and typing errors. Despite the fact that all relevant personnel responsible for scoring according to the different scales are experienced and trained, SSS and EPSS are to some degree a subjective value. It is common procedure, but it still is a potential source of bias in the data. Selection bias can also be suspected due to the fact that patients receiving thrombolytic treatment are always admitted to the stroke unit and may be more likely to be included in the study. On the other hand, patients with more subtle symptoms are less likely to be included in the study.

Chapter 3

Analysis of time-dependent variables

The time-dependent variables from the Trondheim END study that will be analyzed are systolic blood pressure, diastolic blood pressure, blood sugar, temperature and EPSS. The variables will be analyzed with two different strategies. The overall aim of this thesis is to develop a regression model to understand the neurological outcome after AIS, and in this model systolic blood pressure, diastolic blood pressure, blood sugar and temperature will be included as predictors. For these variables we will in Section 3.1 look at summary statistics and consider their marginal predictive potential in END as an alternative to modeling the variables over time. Variability parameters and other summary statistics can capture essential features of the response over time. Summary statistics are an approach that simplifies longitudinal data to a single value. When it comes to the analysis of the variable EPSS in time, the strategy is different. END is defined as in Section 2.2 and is a binary variable constructed only from the baseline measurement and 72hrs after stroke measurement of parts of the EPSS score. By analyzing EPSS it is of interest to investigate if it is possible to conceive more relevant information from EPSS that is not already contained in the END-variable. The motivation for this preliminary analysis with summary statistics and EPSS will be presented more in detail in Section 3.1 and 3.3

3.1 Time-dependent predictors

Due to the amount of imputed values, an alternative procedure than including the time points in a regression model can give a more realistic prediction. In addition, according to the results in Chung et al. (2015) it is more interesting to look at the variability in the blood pressure than the level at each measurement, and hopefully capture more information from different variability parameters than modeling blood pressure over time. According to Pezzini et al. (2011) the optimal management of blood pressure during acute ischemic stroke is controversial. It is of this reason important to capture as much information as possible

from the blood pressure measurements so that clinical guidance of blood pressure management can be improved. A study performed in Bergen found that low body temperature on admission were related to END (Nacu et al., 2016). By analyzing summary statistics that measures level and the variability, both of these findings will be explored further for the Trondheim END study. These facts are the motivation for analyzing different summary statistics for the time-dependent predictors.

The variability parameters calculated for each patient in Chung et al. (2015) are the range (*max-min*), the standard deviation (*sd*) and coefficient of variation (*cv*). The coefficient of variation is calculated as $sd \times 100/mean$. In addition, the mean, the minimum value (*min*) and the maximum value (*max*) are also calculated and represent different levels of the measurements. The same summary statistics are calculated for the blood pressure and the other time-dependent predictors in the Trondheim END study. The summary statistics based on the minimum, maximum and range are not affected by the problems with the imputed values. However, the standard error is underestimated and the average value of the measurements can either underestimate or overestimate the true mean.

3.1.1 Quintile analysis of binary outcome

When analyzing measurements of a continuous variable it is sometimes useful to group the subjects. The cut-off point for splitting the observations are called quantiles (Altman and Bland, 1994). Example of quantiles are tertiles which split the data in three and quintiles which split the data in five. To visualize and explore the behavior of the different summary statistics, the patients are divided into quintiles based on their value of the summary statistic and in each quintile the percentage of patients with END is calculated. This is done for all the time-dependent predictors. Dividing continuous variables into quantiles are often used in epidemiologic research to illustrate the relationship to a binary outcome (Bennette and Vickers, 2012). The calculation of the k-th quintile cut-off point is

$$q_i = \frac{k(n+1)}{5} \quad i = 1, 2, 3, 4 \tag{3.1}$$

where $k = 1, 2, 3, 4$ and n is the number of observations (Altman and Bland, 1994). If for example q_1=73.8 and $n = 368$, the first cut-off point is the 0.8 value of the way between the 73rd and 74th observation of the sorted observations in increasing order. If the value of the 73rd sorted observations is 131 and the value of the 74th sorted observations is 131.1, the 1st quintile is $0.8 * (131.1 - 131) + 131 = 131.1$.

In each quintile, the number of patients with END compared to the total number of patients will be treated as a binomial proportion. Often, confidence intervals for a binomial proportion is computed as a normal approximation interval, but there are other choices. Here, the confidence interval is calculated in

`R` with the `binom.test`-function which uses the Clopper-Pearson method and is based on the cumulative probabilities of the binomial distribution. The confidence interval is calculated by using the relationship between the binomial distribution and the beta distribution (Bilder and Loughin, 2015), and is given as

$$\text{Beta}(\frac{\alpha}{2}; n_{END}, n_q - n_{END} + 1) < \theta < \text{Beta}(1 - \frac{\alpha}{2}; n_{END} + 1, n_q - n_{END}) \quad (3.2)$$

where α is the confidence interval level, n_q is the number of patients in each quintile (trials) and n_{END} is the number of patients with END in each quintile (events).

3.1.2 χ^2-test for homogeneity and for trend

The χ^2-test for trend is will be used to investigate linearity between END and the different summary statistics. The test is closely related to the χ^2-test for homogeneity that will be presented first. It is often of interest to compare the distribution of a categorical variable in one sample with a categorical variable of another sample, and the χ^2-test can be used for this purpose. The null hypthesis is that the numbers in each cell are proportinately the same in both samples, and the alternative hypothesis is that there is a significant difference. The statistical theory in this section is from McHugh (2013).

The χ^2 statistics is given by

$$\chi^2 = \frac{\sum_{cells}(O_i - E_i)^2}{E_i} \quad (3.3)$$

where O_i is the observed value in each cell of the table and E_i is the expected value in cell i of the table (Example given in Table 3.1). The expected value is calculated as

$$E_i = \frac{n_{ri} \times n_{ci}}{n}$$

where n_r is the row total for cell i, n_c is the column total for cell i and n is the total sample size. Asymptotically χ^2 follows a χ^2-distribution with parameter df = (Number of rows-1) $\times$ (Number of columns-1). The underlying assumptions for using the test is that the data in the cells are frequencies or counts, the levels of the variables are mutually exclusive, the study groups must be independent, the value of E_i in each cell should be 5 or more in at least 80% of the cells and all cells should have $E_i \geq 1$. If the assumptions are met, the χ^2-statistic can be used to calculate a p-value and to reject or accept the null hypothesis.

If there is a meaningful order of the groups, Armitage (1955) presented another test that can be used to perform a test for linear trend across the different groups. It is a modification of the χ^2-test to incorporate a suspected ordering and will have higher power than the test in Equation (3.3) if the trend is correct. The test can be used on a $k \times 2$ contingency table, and an example of a

5×2-table can be seen in Table 3.1. The test can be used when the response is a two-level variable and the other variable is ordinal in k groups. The null hypothesis is that the binomial proportion is the same for all levels and that there are no linear trend. The Cochran-Armitage trend statistic is given in Agresti (2002) and with the notation from Table 3.1, the test statics for trend is

$$z^2 = \left(\frac{\sum_{i=1}^{k}(w_i - \bar{w})n_i}{\frac{n_{c1}}{n}(1 - \frac{n_{c1}}{n})\sum_{i=1}^{k} n_{ri}(w_i - \bar{w})^2} \right)^2, \tag{3.4}$$

where $\mathbf{w} = (1, 2, 3, 4, 5)$ are weights and $\bar{w} = (\sum_{i=1}^{k} n_{ri}w_i)/n$. Asymptotically this test statistic also follows a χ^2-distribution, but now on 1 degree of freedom.

Table 3.1: An example of a 5×2-table of counts for the binary END-variable for *SBPmean* divided in fifths.

	END	No END	
Q1	n_1	$n_{r1} - n_1$	n_{r1}
Q2	n_2	$n_{r2} - n_2$	n_{r2}
Q3	n_3	$n_{r3} - n_3$	n_{r3}
Q4	n_4	$n_{r4} - n_4$	n_{r4}
Q5	n_5	$n_{r5} - n_5$	n_{r5}
	n_{c1}	n_{c2}	

3.2 Results of the time-dependent predictor analysis

The patients in the Trondheim END study is divided into fifths with the quintile cut-off point given in Equation 3.1 for the six summary statistics from each time-dependent predictors. A confidence interval for the probability of END in each quintile is calculated with the Clopper-Person method from Equation 3.2. The `CochranArmitageTest`-function from Signorell (2015) in `R` is used to explore significant trend between the quintile divided fifths for the summary statistics for each time-dependent predictor. The function uses Equation (3.4) and find the corresponding p-value. The result can be found in Table 3.2 and 3.3.

A plot of the percentage of END and the quintiles for the different blood pressure parameters can be seen in Figure 3.1 and 3.2. For the *SBPmax-min*-parameter there may be a trend, but the results from the trend test in Table 3.2 did not find any significant trends for the systolic blood pressure parameters. For the diastolic blood pressure, Figure 3.2 shows a possible trend in *DBPmax*, *DBPmax-min*, *DBPsd* and *DBPcv*, and using the trend test we found a significant linear trend for the same parameters. As a comparison, Chung et al. (2015) found significant p-values at a 0.05 level for all blood pressure quintiles expect for *SBPmin* and *DBPmean*.

How the percentage of END is associated with the quintiles of blood sugar (BS) parameters can be seen in Figure 3.3. There seems to be a possible increasing trend in *BSsd*, *BSmin* and *BSmean*, and using the trend test we found a significant linear trend for *BSmean*, *BSsd*, *BSmin* and *BSmax*. The results for the different temperature (TEMP) parameters can be seen in Figure 3.4 and all of the parameters show a possible linear trend. Also, using the trend test we found a p-value below 0.05 for all of the temperature parameters.

Table 3.2: The results from the Cochran-Armitage trend test for the blood pressure summary statistics. The *-marking indicates a significant trend at a 0.05 significance level.

SBPmean	0.2729	*DBPmean*	0.5734
SBPmax-min	0.1031	*DBPmax-min*	0.001238 *
SBPcv	0.5031	*DBPcv*	0.005058 *
SBPsd	0.3223	*DBPsd*	0.005058 *
SBPmin	0.7714	*DBPmin*	0.3521
SBPmax	0.1566	*DBPmax*	0.01302 *

Table 3.3: The results from the Cochran-Armitage trend test for the temperature and blood sugar summary statistics. The *-marking indicates a significant trend at a 0.05 significance level.

TEMPmean	$2.317 \cdot 10^{-6}$ *	*BSmean*	$4.435 \cdot 10^{-5}$ *
TEMPmax-min	0.007004 *	*BSmax-min*	0.05119
TEMPcv	0.009598 *	*BScv*	0.2729
TEMPsd	0.007004 *	*BSsd*	0.03052 *
TEMPmin	0.007004 *	*BSmin*	0.0001088 *
TEMPmax	$8.151 \cdot 10^{-8}$ *	*BSmax*	0.002555 *

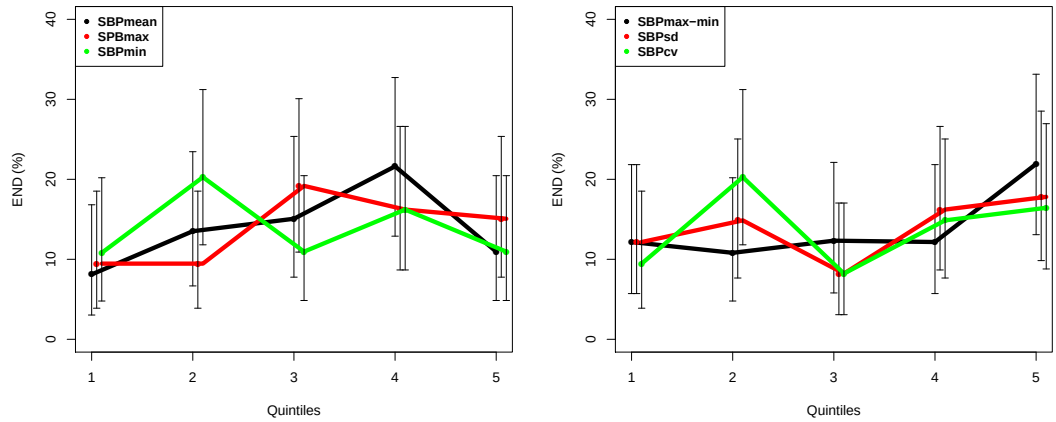

Figure 3.1: Proportions of patients developing END in the quintiles for the systolic blood pressure parameters together with the corresponding Clopper-Pearson confidence interval.

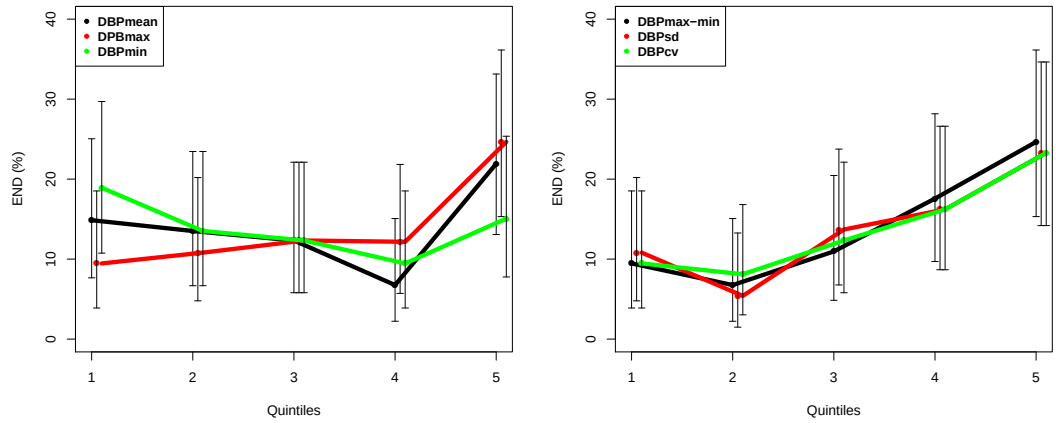

Figure 3.2: Proportions of patients developing END divided in the quintiles for the diastolic blood pressure parameters together with the corresponding Clopper-Pearson confidence interval.

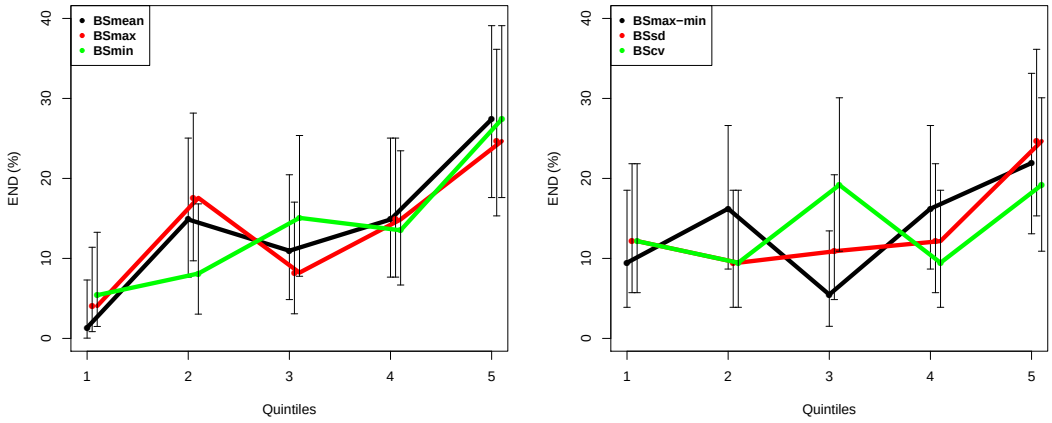

Figure 3.3: Proportions of patients developing END divided in the quintiles for the blood sugar parameters together with the corresponding Clopper-Pearson confidence interval.

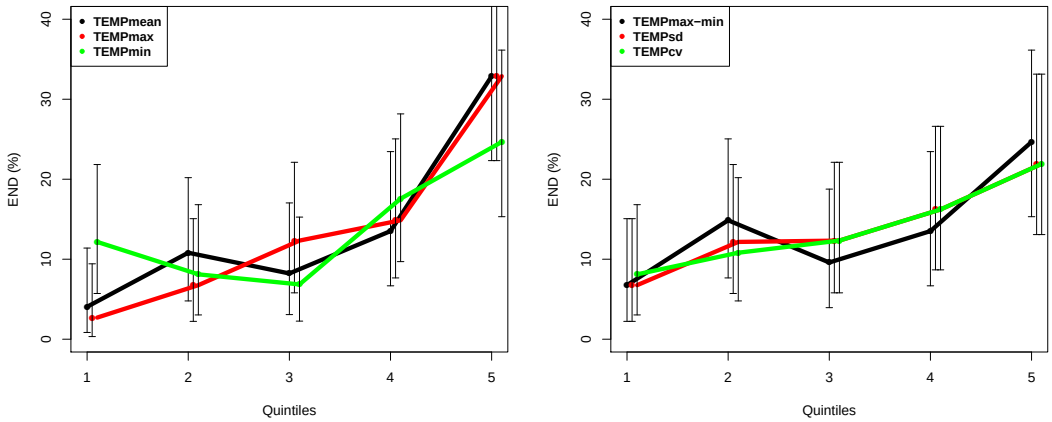

Figure 3.4: Proportions of patients developing END divided in the quintiles for the temperature parameters together with the corresponding Clopper-Pearson confidence interval.

3.3 Analysis of EPSS

Exploration of factors influencing early neurological deterioration is of overall interest. In our data a total of 11 measurements of the EPSS score have been collected at different time points during 72 hours after admittance to the hospital. Since END is defined based on EPSS, it is of interest to investigate the behavior of the EPSS values. The 11 EPSS values for a random selection of 9 patients are shown in Figure 3.5, and for most of the patients the values does not change drastically. A total of 42 of the 368 patients actually have the same value of EPSS for all 11 measurements and because of the imputation problem presented in Section 2.5 it is difficult to distinguish imputed values from actual measurements. Thus, analyzing the data over time with EPSS as response may not model the reality adequately. However, the classification between END or no END only uses the first and last measurement, and information between these time points are rejected. Of this reason a preliminary, unsupervised analysis of all of the EPSS values is done to explore all information in the measurements. Unsupervised means that there are no known answer, no quantitative response variable and no direct measure of success. Prediction of a response is not the goal of unsupervised analysis. Exploration of possible trend or groupings in the data is often a good place to start. In addition, and on the contrary to experimental studies, observational studies often rely on statistical techniques to account for differences that result from lack of randomization and external variations.

There are several classification possibilities based on the EPSS values. One option is END vs no END, and other options are EDE or a combination of EDE and END. A variable with three levels is already made. The patient is classified as 0 if he or she did not experience END or any EDE, 1 if the patient has experienced at least one EDE but no END and 2 if the patient has END. A total of 13.9% of the patients are in group 2, 28.3% of the patients are in group 1 and 57.8% of the patients are in group 0. The aim of this section is to explore possible groups based on the EPSS values and compare the groups to the three level classification rule and the binary END/no END variable. As a result we want to find the dependent variable of primary interest when it comes to modeling early neurological deterioration.

The focus in this chapter is not to present statistical methods in detail, but rather to explore and visualize the EPSS values and the classification rules. However, the fundamental idea and statistic behind each method is presented. To do this and to see if there are patterns in the EPSS values that is in coordination with one of the three classifications above, a linear model, principal component analysis and clustering are used. The theory will be presented first and then the results from the Trondheim END study follows.

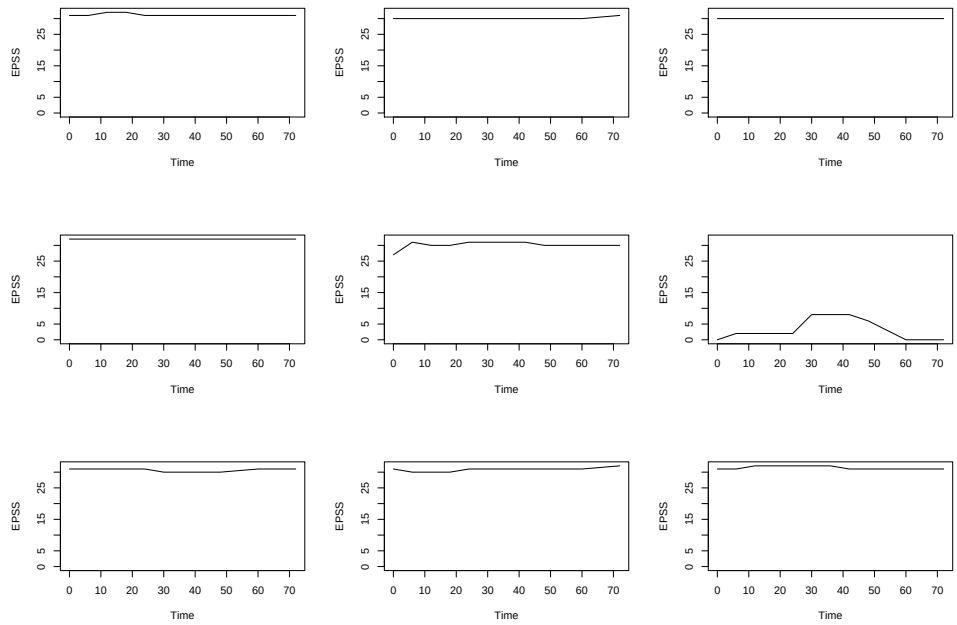

Figure 3.5: The development in the EPSS values over time for 9 randomly chosen patients. The EPSS score ranges from 0-32, where 32 indicates that the patient has the highest level of function.

3.3.1 Linear model

In general, the linear model is given as

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i, \quad (i = 1, ..., n) \tag{3.5}$$

where Y_i is the response, n is the number of observations, p is the number of predictor variables, $\mathbf{x_i}$ is the value of the p predictors for the i-th observation, $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)$ are unknown parameters and $e_i \sim N(0, \sigma^2)$. With matrix notation, minimizing the sum of the squared errors with respect to the model parameters gives that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{3.6}$$

and is derived in Chapter 3 in Bingham and Fry (2010). The linear model is not directly a part of unsupervised analysis, but will be used to look for a linear trend and to explore the behavior of the response. EPSS will be used as the response and time will be used as the predictor variable.

3.3.2 Principal component analysis

Principal components analysis (PCA) is a tool for exploratory data analysis and can be used to give a low-dimensional representation of the data. PCA summa-

rizes the correlated variables with a smaller number of representative variables that explains most of the variability. Graphical representation is difficult with data of high dimension and PCA is a powerful tool when it comes to visualizing high dimensional data. PCA highlights similarities and differences in the data and can also be used to explore hidden structures in the observations and to find outliers, and is because of this a part of unsupervised learning. PCA uses eigenvectors and eigenvalues of the covariance matrix to reduce the dimensions without loosing too much information, and the eigenvector with the highest eigenvalue is the principle component of the data. Singular value decomposition (SVD) is directly related to PCA in the case where principal components are calculated from the covariance matrix (Wall et al., 2003, Chapter 5). The SVD of a matrix $\mathbf{X}$ is to express each x_{ij} as

$$x_{ij} = \sum_{k=1}^{r} \theta_k u_{ki} v_{kj} \tag{3.7}$$

where $\theta_1 \geq \theta_2 \geq ... \geq \theta_r$ and r is the rank of $\mathbf{X}$. In matrix notation the expression is given as

$$\underset{n\times p}{\mathbf{X}} = \underset{n\times p}{\mathbf{U}} \; \underset{p\times p}{\theta} \; \underset{p\times p}{\mathbf{V}^T}, \tag{3.8}$$

where the matrix $\boldsymbol{\theta}$ is a $p \times p$ diagonal matrix with positive or zero elements called singular values, $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^\theta\mathbf{V} = \mathbf{I}$. The columns of $\mathbf{U}$ are called left singular values and the rows of $\mathbf{V}^T$ is called the right singular vectors. The principal components are the eigenvectors of the covariance matrix and with the SVD of $\mathbf{X}$ the matrix can be written as

$$\mathbf{C} = \frac{1}{n}\mathbf{X}\mathbf{X}^T = \frac{1}{n}\mathbf{U}\boldsymbol{\theta}^2\mathbf{U}^T \tag{3.9}$$

which is often easier to work with than the covariance matrix itself (Madsen et al., 2004) and

Interpretation of the principal components is based on finding which of the original variables that are correlated with each component. The correlation values that are farthest from zero in either positive or negative direction is of most importance. The first principal component is a linear combination of the original variables which captures the maximum variance in the data set and determines the direction of the highest variability in the data. The second component is uncorrelated to the first one and the directions between the components are orthogonal. Principal components can also be used as predictors in a regression model and to prepare the data for further analysis with other statistical techniques.

3.3.3 Clustering

Cluster analysis is also a part of unsupervised learning and can be used to investigate if the observations can be grouped in clusters, such that the objects

within each cluster are more closely related to one another than objects in a different cluster. There exist different clustering methods, but all methods attempt to group the objects based on a measure of similarity supplied to it. The clusters are believed to reflect the underlying structure of the data. Clustering can for example be based on the correlation or the Euclidean distance between the observations, and the choice of similarity measure must be made based on prior knowledge of the data (Tibshirani, 2013). The k-means algorithm is a popular clustering method and partitions the observations into a pre-specified number of clusters, where k is the number of clusters. Hierarchical clustering is a popular method when the number of clusters is not known in advance. The different approaches both have their advantages and disadvantages, and the clustering method used in this section is the k-mediods method which is more robust and more computationally intensive than k-means. A mediod is a data point where the average dissimilarity to all the other data points is minimal.

3.4 Results of the analysis of EPSS

Visualization of the behavior of the EPSS values over time is done with the linear model presented in Section 3.3.1. The function `lm` in `R` is used to estimate the intercept $\hat{\beta}_0$ and the slope $\hat{\beta}$ for each patient and the result is given in Figure 3.6. The color coding is based on the three level classification presented in Section 3.3. The spread of points around the constant slope in the figure indicates that little change is observed in the EPSS values over time. Most of the patients with high EPSS values experience little change, and this is the clustering in the right half in the figure. Often, the trend of the regression line is of interest and deviating points can be treated as outliers from the trend. In this case, on the other hand, the patients that deviate from the regression line are the one of interest.

When defining $\mathbf{X}$ presented in Section 3.3.2 to be the 368×11-matrix of the EPSS values, the result from the PCA on the EPSS values from the Trondheim END study can be seen in Figure 3.7. The data are projected onto the first two principal components. The color coding in the plot is based on the same classification as above, and it does not seems to highlight any structure of the data that we didn't see in Figure 3.6. Also, the first principal component explains 91.1% of the total variance. The value of the four first principal components for each time point can be seen in the `R`-output below. PC1 is just a linear combination of the average values of EPSS which is not very informative when it comes to significance of the variables. James et al. (2013) say that if no interesting patterns are found in the first few principal components, then it is unlikely that other principal components are of interest.

```
> epss.pca$rotation
                PC1        PC2         PC3         PC4
epss1   -0.2744192  0.5330892 -0.75648373 -0.13040522
epss2   -0.2944197  0.4118888  0.23147770  0.17063754
```

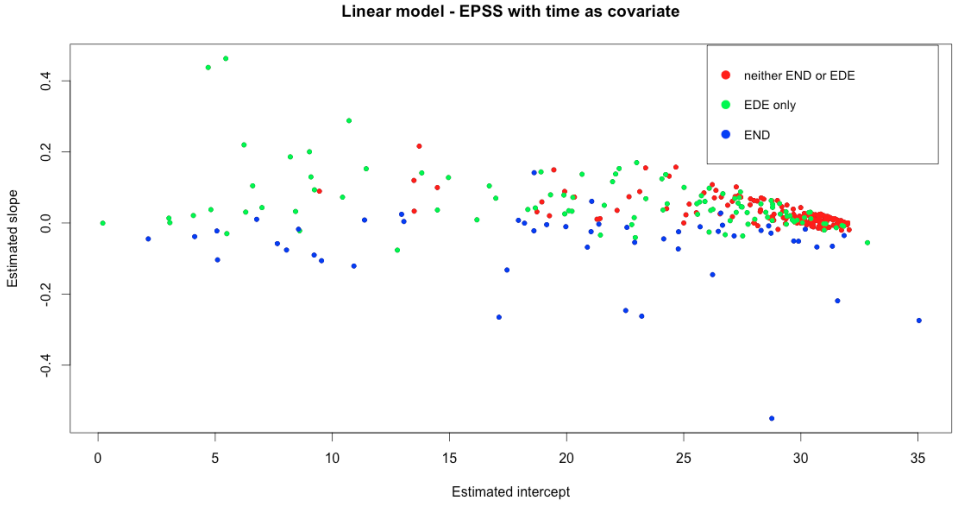

Figure 3.6: Linear model with EPSS as response and time as predictor. The points represents the $n = 368$ patients and is colored according to the three level classificiation.

```
epss3   −0.2977366   0.3842571   0.26701083   0.16324996
epss4   −0.3009821   0.2420529   0.47251699   0.07700169
epss5   −0.3095214  −0.1281256   0.05406679  −0.32120095
epss6   −0.3089475  −0.1480964  −0.03519635  −0.40300270
epss7   −0.3088352  −0.1942365   0.04775780  −0.31423267
epss8   −0.3096448  −0.1913885   0.06499454  −0.24624445
epss9   −0.3066234  −0.2619321  −0.04957589   0.06112270
epss10  −0.3044499  −0.2711949  −0.14380508   0.37968161
epss11  −0.2992574  −0.2916492  −0.21395930   0.59029117
```

The function `pamk` (partitioning around mediods) from the library `cluster` in R is used to assign the observations to different clusters. A pre-specified number of cluster must also be done in the k-mediods method, and the `pamk`-function also solves the problem of finding k. Regarding the EPSS values in the Trondheim early neurological deterioration study, both level differences measured by Euclidean distance and shape of the observation profiles measured by correlation is of interest. The correlation can not be calculated for patients with the same value over the 11 measurements of EPSS, and the cluster analysis with correlation as similarity measure uses 326 of the 368 patients. The result of the cluster analysis can be seen in Figure 3.4 and 3.9, and we see that both methods chose two clusters. In spite of the two clusters, it is hard to see very clear differences between the observations in the different clusters and the cluster regions are overlapping. Interpreting the results according to the END/no END variable, we find that 270 patients with no END are the cluster group 1 based on the Euclidean distance, and that 214 patients with no END are in cluster group 1 based on the correlation. All in all, this is a total of 70% of the patients and

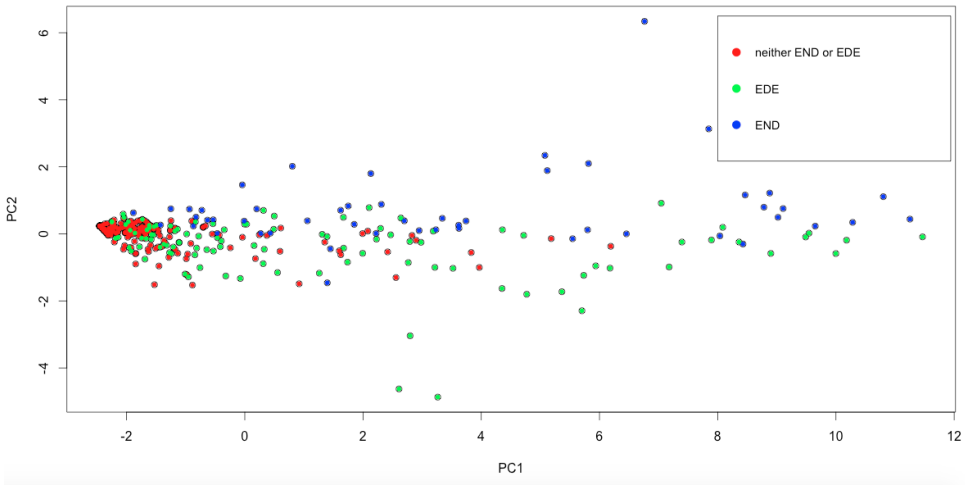

Figure 3.7: Plot of the EPSS values over time projected onto the first two principal components.The points represents the $n = 368$ patients and is colored according to the three level classificiation

can indicate that a binary representation of the EPSS values is adequate.

3.5 Conclusion of the analysis of the time-dependent variables

For the blood sugar, temperature and diastolic blood pressure both summary statistics measuring level and variability were significant. The mean, minimum, maximum, range and standard deviation are included in the statistical analysis in Chapter 5. It especially of interest to look at the behavior of the highly significant parameters from Table 3.2 and 3.3 when other predictors are also included in the analysis, and this will be done in Chapter 6. The coefficient of variation is not included in further analysis since it is based on the value of the mean and the standard deviation.

The aim of the EPSS section was twofold: 1) To explore possible groups based on the 11 EPSS values for each patient 2) comparison of the groups found in 1) to the classification rules based on END and EDE. We have seen good correspondence with the linear regression slope and intercept and the classification rules with 3 levels, that the PCA result didn't show any new information and that the clustering analysis and the END/no END variable were in agreement for the majority of the patients. After doing the unsupervised analysis of EPSS, we choose to use the binary classification between END or no END as dependent variable of primary interest in further statistical analysis

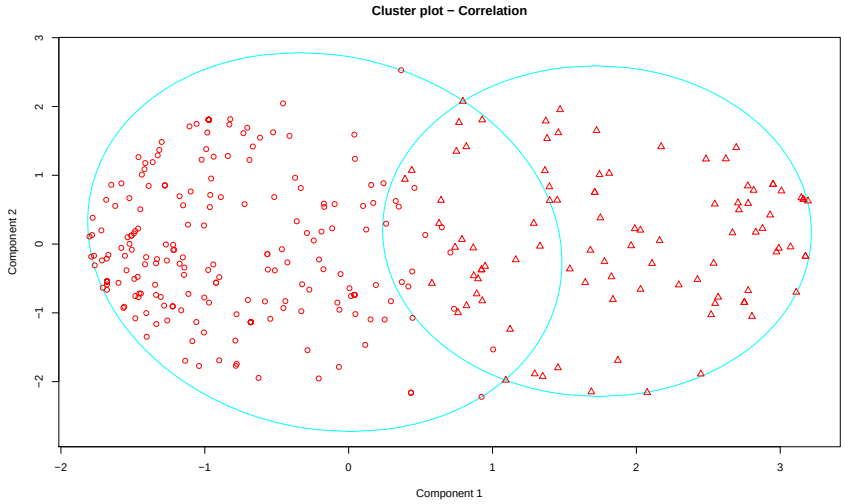

Figure 3.8: The patients divided into $k = 2$ clusters with correlation as similarity measure.

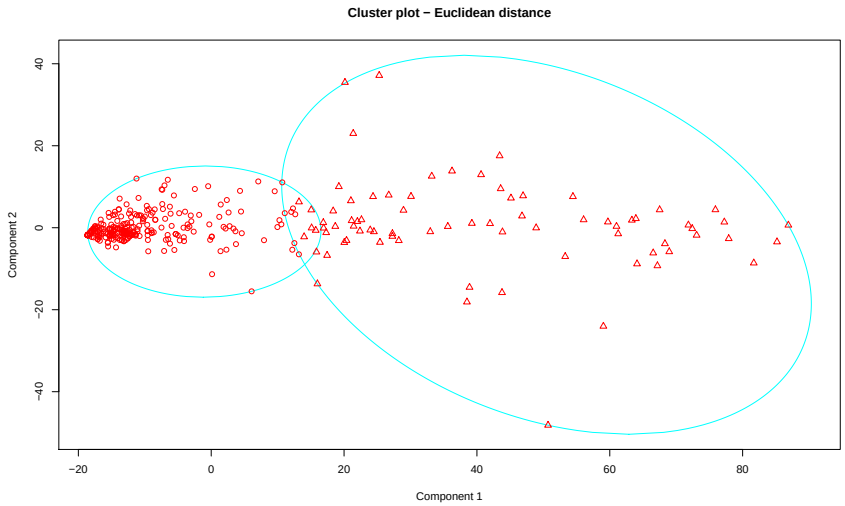

Figure 3.9: The patients divided into $k = 2$ clusters with Euclidean distance as similarity measure.

Chapter 4

Sparse modeling in logistic regression

Logistic regression is and has for a long time been popular in medical and epidemiological research since binary variables such as alive/dead, diseased/healthy or medication/placebo are of general interest. The goal of logistic regression is to find the best model to describe the relationship between the binary data and independent predictors. Logistic regression permits the use of both continuous and categorical predictors. The motivation for using a logistic regression model in this analysis is the binary variable END/no END. In medical research and trials there are often numerous predictors that is interesting to include in a statistical model, and the best ratio between number of predictors and number of events for a stable model is not known. Including many perdictors in a logistic model may lead to a overfitted model. The outcome in logistic regression is usually coded as 0 or 1, and 1 is often referred to as "event" and is usually the least common of the two. Different simulation studies have evaluated the effect of the number of events per variable (EPV) in logistic regression. Peduzzi et al. (1996) found that when EPV > 10, no problems with overestimation and underestimation of the variance and biased regression coefficients occurred. In the Trondheim END study we have 51 events (END) and 40 possible predictors. The corresponding EPV-value is 1.3, and including all predictors may lead to imprecise coefficient estimates. In the high-dimensional setting when $p > N$ the logistic regression model cannot be used at all without regularization. In addition, a common problem in statistical analysis is which selection procedure to use for finding variables that might influence the outcome variable. Forward selection, backward selection and stepwise selection can be used to select variables where each variable is evaluated individually, and each method have their has advantages and disadvantages. To overcome the issues with the number of predictors and the choice of the variable selection procedure an alternative to traditional logistic regression is presented in this chapter.

The Least Absolute Shrinkage and Selection Operator (lasso) is a shrinkage and selection method for regression models. We have chosen to use this method

instead of traditional logistic regression due to our low EPV-value and the model selection property. The method was originally applied to ordinary least squares (OLS) regression, but has in the recent years been extended to logistic, multinomial, Poisson and Cox regression models. Interpretability of the final model and accuracy of prediciton is the motivation for finding alternatives to OLS. The method was first introduced in Tibshirani (1996) and is becoming more and more popular as the method has been expanded and improved upon. The lasso is a central part of the rapidly evolving field of sparse statistical modeling which in our setting means that only a small number of predictors are included in the final model. Variable selection becomes increasingly important in modern data analysis as we want to find the predictors giving the best prediction model among numerous possible predictors in big data sets. The lasso estimation of the parameters is done using `R` and the package `glmnet` from Friedman et al. (2010), and the underlying mathematical and statistical theory is presented in this chapter. Concepts and algorithms from optimization theory is also an important part of the lasso and some of the most fundamental optimization theory used in the lasso methods are also presented in this chapter. In the field of medical statistics lasso has the potential to be very useful as it handles data with numerous predictors.

This chapter begins with a presentation of generalized linear models, the logistic regression model and then the theory behind the lasso regression and lasso-penalized logistic regression follows. Statistical terms and methods for understanding the lasso and for presenting the result are also presented in this chapter. The generalized linear model theory is taken from Rodríguez (2007) and the main reference for the statistical methods presented in Section 4.2 and 4.3 is Hastie et al. (2015). In Chapter 5, the theory presenten in this chapter is applied to the data from the Trondheim early neurological deterioration study.

4.1 Generalized linear models

Let $Y_1, ...Y_n$ be random variables in a sample of size n and let $\mathbf{x}_i' = (x_{i1}, ..., x_{ip})$ be the value of the predictors for the i-th observation where p is the number of the predictors. The linear model presented in Section 3.3.1 assumes that the random variable Y_i has a normal distribution with mean μ_i and variance σ^2,

$$Y_i \sim N(\mu_i, \sigma^2), \tag{4.1}$$

and that the expected value μ_i is a linear function of p predictors and a vector of unknown parameters $\boldsymbol{\beta}$, such that

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

The normality assumption and the assumption of a linear relationship between the response and the predictors are not always reasonable. Two generalization of the linear model can be done to obtain a more applicable model

and this is what is called generalized linear model (GLM). The first part of the generalization is that it is assumed that observations can follow any distribution belonging to the exponential family. A probability distribution function that can be written as

$$f(y_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\}$$

where θ_i and ϕ are parameters and $a_i(\phi)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known functions is a member of the exponential family. Examples of exponential families are the normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions. The second part of the generalization is the focus of modeling a transformed mean η_i instead of the mean μ_i, that is,

$$\eta_i = g(\mu_i). \tag{4.2}$$

The function $g(\mu_i)$ is called a link function and needs to be one-to-one continuous differentiable. The transformed mean is assumed to have a linear relationship to the predictors, so that

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta}, \tag{4.3}$$

and

$$\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}).$$

Maximum likelihood estimation is used to estimate the parameters rather than ordinary least squares, and an iterative computational procedure is used in the estimation.

4.1.1 Logistic regression

The binomial distribution belongs to the exponential family and the logistic regression model is a generalized linear model with a so called logit link function. The logistic regression model is used when the response variable is binary, and is the most popular model for binary data. The logit link function is given as

$$\eta_i = \text{logit}(\pi_i) = \log\frac{\pi_i}{1-\pi_i}, \tag{4.4}$$

where π_i is the probability that Y_i takes the value 1 and $1-\pi_i$ is the probability that Y_i takes the value 0. The distribution of the random variable Y_i then follows a binomial distribution with size 1. This special case of the binomial distribution is the same the Bernoulli distribution given as

$$\Pr\{Y_i = y_i\} = \pi_i^{y_i}(1-\pi)^{1-y_i}. \tag{4.5}$$

The expected value of a Bernoulli distributed random variable is

$$E[Y_i] = \Pr\{Y_i = 1\} \cdot 1 + \Pr\{Y_i = 0\} \cdot 0 = \pi_i \cdot 1 + (1-\pi_i) \cdot 0 = \pi_i$$

and the variance is

$$\begin{aligned}\text{Var}[Y_i] &= E[Y_i^2] - E[Y_i]^2 = \Pr\{Y_i = 1\} \cdot 1^2 + \Pr\{Y_i = 0) \cdot 0^2 - \pi_i^2 \\ &= \pi_i - \pi_i^2 = \pi_i(1 - \pi_i).\end{aligned}$$

We observe that the variance of Y_i is dependent on the parameter π_i. This motivates that linear model in Equation (3.5) that assumes constant variance across Y_i will not be adequate when analyzing binary data.

From the relationship in Equation (4.3) and (4.4) we find that

$$\eta_i = \text{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta} \tag{4.6}$$

and

$$\pi_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

The unknown parameters ($\boldsymbol{\beta}$) are estimated with maximum likelihood estimation and the estimated value of the j-th predictor β_j represents the change in the logit of the probability with a unit change in the predictor when the other predictors are constant. The exponentiated coefficient e^{β_j} represents an odds ratio and is often a useful representation of the result. The likelihood function for n independent Bernoulli observations is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{4.7}$$

and the log-likelihood is given as

$$\text{log}L(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i\text{log}(\pi_i) + (1 - y_i)\text{log}(1 - \pi_i)). \tag{4.8}$$

Based on the relationships in Equation (4.3) and (4.6), the log-likelihood can be expressed as

$$\begin{aligned}\text{log}L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i\text{log}(\pi_i) + \text{log}(1 - \pi_i) - y_i\text{log}(1 - \pi_i) \\ &= \sum_{i=1}^{n}(y_i(\text{log}(\pi_i) - \text{log}(1 - \pi_i)) + \text{log}(1 - \pi_i)) \\ &= \sum_{i=1}^{n}(y_i\text{logit}(\pi_i) + \text{log}(1 - \pi_i)) \\ &= \sum_{i=1}^{n}(y_i\mathbf{x}_i'\boldsymbol{\beta} + \text{log}(\frac{1}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}})) \\ &= \sum_{i=1}^{n}(y_i\mathbf{x}_i'\boldsymbol{\beta} - \text{log}(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}})).\end{aligned} \tag{4.9}$$

From this equation it can be seen that the log-likelihood function is a concave function. Due to the fact that e^x and $\log(x)$ are convex, $\log(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}})$ is also convex and then $-\log(1 + e^{\mathbf{x}_i'\boldsymbol{\beta}})$ is a concave function. This is a motivation for using the negative log-likelihood in the lasso-penalized logistic model and will be discussed further in Section 4.3.

4.1.2 Deviance

If the number of parameters equals the number of observations, the fitted model is called a saturated model and in a saturated model the log-likelihood achieves its maximum value. It is useful to compare any proposed model to the saturated model. The saturated model is the most complex model possible and provides a perfect fit. The deviance D is defined as

$$D = -2(\text{logL(proposed model)-logL(saturated model)}) \tag{4.10}$$

and should be small if the proposed model is a good approximation to the true model. The asymptotic sampling distribution of the deviance follows a χ^2-distribution with $n - p$ degrees of freedom. The binomial deviance is

$$D = 2\sum_{i=1}^{n}\{y_i\log(\frac{y_i}{\hat{\mu}_i}) + (1 - y_i)\log(\frac{1 - y_i}{1 - \hat{\mu}_i})\} \tag{4.11}$$

where y_i is the observed value and $\hat{\mu}_i$ is the fitted value for the i-th observation. The derivation of the deviance is based on the log-likelihood functions of the saturated model and the fitted model and can be seen in Rodríguez (2007). From Equation (4.11) it is clear that zero deviance is a perfect fit due to the fact that $\hat{\mu}_i = y_i$. The deviance can also be used as a test statistic for the hypothesis that all parameters that are in the saturated model but not in the fitted model are zero (Agresti, 1966).

4.2 Lasso regression

To present the lasso regression model, we start with the idea of parameter estimation by the ordinary least squares (OLS) method. It is now convenient to include the intercept in the notation to follow the notation in Hastie et al. (2015), and the linear model can be written as

$$Y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$$

where Y_i is the response, p is the number of predictor variables, x_{ij} is the value of the j-th predictor for the i-th observation, n is the number of observations

and β_0 and $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)$ are unknown parameters as before. The ordinary least squares problem is given as

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \tag{4.12}$$

The estimated parameters from Equation (4.12) will typically be nonzero with OLS. When $p > n$ the least-squares estimates are not unique and the solutions will almost surely overfit the data. A constraint is needed to overcome this problem. In the lasso method all p predictors are initially included in the model and the parameters are estimated by solving the constrained problem

$$\text{minimize} \quad \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j_1}^{p} x_{ij}\beta_j)^2 \quad \text{subject to} \quad ||\boldsymbol{\beta}||_1 \leq t \tag{4.13}$$

where $||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|$ is the l_1 norm of β and is called the lasso penalty. The factor $1/2n$ is useful for unequal sized splitting of the data in the cross-validation. The factor makes no difference in the optimization result and the factor is replaced by $1/2$ and 1 in many formulations of the lasso. The maximum number of coefficient selected by the lasso is $\min(N, p)$. Using the l_1 instead of for example the l_2 norm has the advantage of forcing some of the coefficients to be zero and this is the most important property of the l_1-constraint. As a result, lasso does variable selection automatically. The final model is then easier to interpret and the computational advantages are also an important aspect of the l_1-penalty. The parameter t is user-specified and limits the sum of the absolute values of the parameter estimates. Smaller values of t mean a stricter bound and leads to a sparser model than larger values of t. A sparser model means that more β_j's are set to zero. Of course, a too small value of t can prevent the inclusion of important predictors and too large values can lead to overfitting. Cross-validation can be used to find the best value of t and is discussed in Section 4.4.

To understand the model selection property of the lasso, a geometric comparison with the ridge regression is useful. Ridge regression is similar to the lasso method, but instead of using the l_1-norm as constraint, the l_2-norm is used. The minimization problem is given as

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}} \quad \Big\{\frac{1}{2n}\sum_{i=1}^{N}(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\Big\} \quad \text{subject to} \quad ||\boldsymbol{\beta}||_2 \leq t_{ridge}^2 \tag{4.14}$$

where $||\boldsymbol{\beta}||_2 = \sum_{i=1}^{p} \beta_j^2$ and t_{ridge} is a user-specified parameter for the ridge regression model. On the contrary to lasso, ridge regression only shrinks the coefficients and none of the coefficients are shrunken to zero. When $p = 2$ the constraint region for the ridge regression and lasso is $\beta_1^2 + \beta_2^2 \leq t_{ridge}^2$ and $|\beta_1| + |\beta_2| \leq t$, respectively. Figure 4.1 contrasts the difference in the constraints

in the two methods. The solution to both methods is the point where the elliptical contours hit the constraint region. For the l_1-penalty the solution often occurs in a corner, and β_j is equal to zero. This can only happen for lasso regression since the ridge regression constraint is a disk because of the quadratic terms.

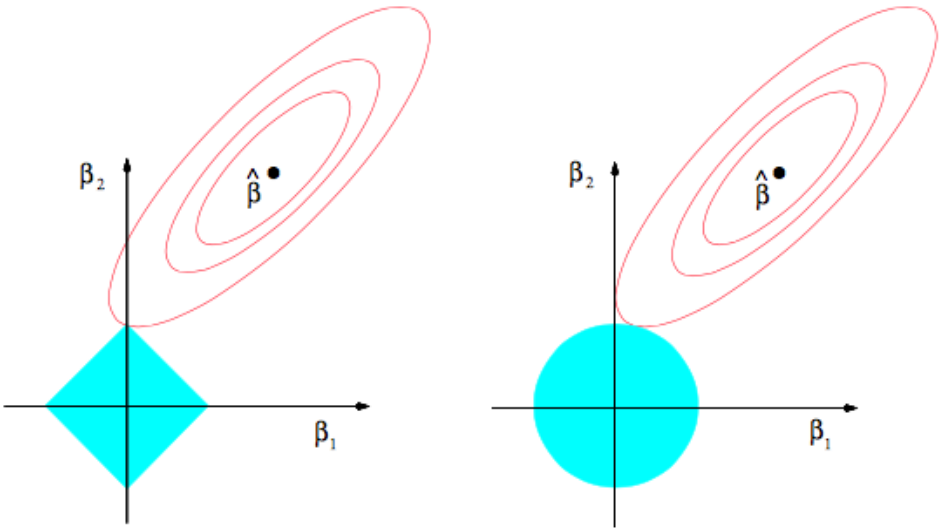

Figure 4.1: Estimation picture for the lasso and ridge regression. The blue areas are the constraint regions and the ellipses are the contour of the residual sum of squares. The picture is taken from (Hastie et al., 2015, Ch. 2, p. 11)

Now, let $\mathbf{Y} = (Y_1, ...Y_n)$ denote the n-vector of responses and let $\mathbf{X}$ be an $n \times p$ matrix of predictors. The minimization problem in Equation 4.13 can then be re-expressed using matrix-vector notation as

$$\underset{\beta_0, \boldsymbol{\beta}}{\text{minimize}} \quad \Big\{ \frac{1}{2n} ||\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta}||_2^2 \Big\} \quad \text{subject to} \quad ||\boldsymbol{\beta}||_1 \leq t \tag{4.15}$$

where $\mathbf{1}$ is a $n \times 1$ vector of ones, and $||\cdot||_2$ is the usual Euclidean norm on vectors. To make the lasso solutions independent on the units of the predictors, the predictors are standardized so that each column is centered ($\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$) and has unit variance ($\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$). The centering and standardization is needed since the constraint on the size of the coefficient associated with each predictor will depend on the magnitude of each predictor. This is also an advantage in the optimization part of the lasso method. In addition, the response value Y_i is centered ($\frac{1}{n}\sum_{i=1}^{n} Y_i = 0$) for convenience. Since the data is centered the intercept can be omitted in the following lasso presentation and can found by calculating

$$\hat{\beta}_0 = \bar{Y} - \sum_{j=1}^{p} \bar{x}_j \hat{\beta}_j$$

where $\bar{Y}$ and $\bar{x}_1, ...\bar{x}_p$ are the means and $\hat{\beta}_j$ is the optimal lasso solution for the j-th predictor.

On Lagrangian form[1] the problem from Equation 4.15 is rewritten as

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\Big\{\frac{1}{2n}||\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}||_2^2+\lambda||\boldsymbol{\beta}||_1\Big\} \tag{4.16}$$

for some $\lambda \geq 0$ and when β_0 is omitted. This formulation of the problem is convenient for numerical computation of the solution. There is a one-to-one correspondence between the optimization problem given in Equation (4.15) and the Lagrangian form given in Equation (4.16). In general this means that for each t where the constraint in Equation (4.13) is active, there exist a corresponding λ that solves the Lagrangian form of the problem. Estimation of the lasso method relies on numerical optimization approaches and a brief overview of possible procedures for solving the minimization problem is presented in Section 4.6.

Another important aspect of the lasso method is the bias-variance trade-off. The prediction error contains both a bias-part and a variance-part. Ideally, both bias and variance should be reduced as much as possible simultaneously. Traditionally, the focus have been on unbiased estimators and the least square estimate that has the minimum variance among all linear unbiased estimators. However, there can be a lot of variability in the least squares fit and the predictions may be poor. The lasso method trades off an increase in the bias with a decrease in the variance and will be discussed later in the chapter.

4.3 Lasso-penalized logistic regression

As discussed in Section 4.1, maximum likelihood is used to estimate the unknown parameters $\boldsymbol{\beta}$. Fitting a generalized linear model based on maximizing the log-likelihood is the same as minimizing the negative log-likelihood, and the lasso method for a generalized linear model is given as

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}} \quad \Big\{-\frac{1}{n}\text{log}L(\beta_0,\boldsymbol{\beta};\mathbf{Y},\mathbf{X})+\lambda||\boldsymbol{\beta}||_1\Big\}.$$

The log-likelihood for the logistic model is given in Equation (4.9). The resulting negative log-likelihood with l_1-penalty is

$$-\frac{1}{n}\sum_{i=1}^{n}(Y_i(\beta_0+\mathbf{x}_i'\boldsymbol{\beta})-\log(1+e^{\beta_0+\mathbf{x}_i'\boldsymbol{\beta}}))+\lambda||\boldsymbol{\beta}||_1.$$

Lasso for logistic models are similar to lasso for linear models expect that the response variable Y_i can only take two possible values and that the negative log-likelihood is minimized instead of error sum of squares. Finding a model with

[1]Nocedal and Wright (2006) say that the Lagrangian form of the minimization problem $\min \quad f(x) \quad \text{subject to} \quad c_1(x) \leq t$ is $\mathcal{L}(x,\lambda) = f(x)+\lambda c_1(x)$

few important predictors among numerous possible predictors is also possible for logistic regression models and this is known as sparse logistic regression. By introducing a penalty on the maximum likelihood estimation, the logistic regression model can be used in the high-dimensional setting when $p > n$. The log-likelihood given in Equation (4.9) was shown to be a concave function, so then the minimization problem

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n}(-\frac{1}{n}\sum_{i=1}^{n} Y_i(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) - \log(1 + e^{\beta_0+\mathbf{x}_i'\boldsymbol{\beta}})) \quad \text{subject to} \quad ||\boldsymbol{\beta}||_1 \leq t$$

has the benefits of being a convex optimization problem. The Lagrangian form of the lasso-penalized logistic model is

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}} \quad \big\{ -\frac{1}{n}\sum_{i=1}^{n}(Y_i(\beta_0 + \boldsymbol{\beta}^T x_i) - \log(1 + e^{\beta_0+\boldsymbol{\beta}^T x_i})) + \lambda||\boldsymbol{\beta}||_1 \big\}.$$

4.4 Cross-validation

A crucial part of the lasso is to choose the regularization parameter λ since it controls the strength of shrinkage and thereby the model complexity. This can done with the cross-validation method. The method is basically to use one part of the available data to fit the model and the rest of the data to evaluate the model fit on the new data and then repeat the procedure. Analysis is performed on the train subsets and the remaining subset is the validation subset. There are different ways to divide the the data. K-fold cross-validation for example divides the data into K roughly equal-sized parts. A model is fitted to $K - 1$ parts of the data, and one part is used to evaluate the fitted model. The procedure is repeated for $k = 1, 2, ..., K$ and for each k a prediction error is calculated. In this way, K different estimates of the prediction error is obtained and is then averaged for each value of λ. Often $K = 5$ or $K = 10$ is used, but choosing K involves different considerations. Visualization of the scenario when $K = 5$ can be seen in Figure 4.2. With $K = N$ the method is called leave-one-out cross-validation and the chosen λ is approximately unbiased but the variance can be high (Hastie et al., 2008). For large data sets, the computational burden of choosing $K = N$ is considerable. On the other hand, choosing $K = 5$ or $K = 10$, the cross-validation has lower variance but bias can be a problem.

The function `cv.glmnet` from the `glmnet`-package returns two values of λ, both the value for that minimizes the deviance ($\hat{\lambda}_{min}$) and the largest value of λ such that the error is one standard error from the minimum ($\hat{\lambda}_{1se}$). The latter includes the fewest predictors in the final model of the two and with numerous predictors $\hat{\lambda}_{1se}$ will find the most interpretable model. In the case of a binary response and lasso-penalized logistic regression, the deviance is used as prediction error measure and to find $\hat{\lambda}_{min}$ and $\hat{\lambda}_{1se}$. The formula for the deviance

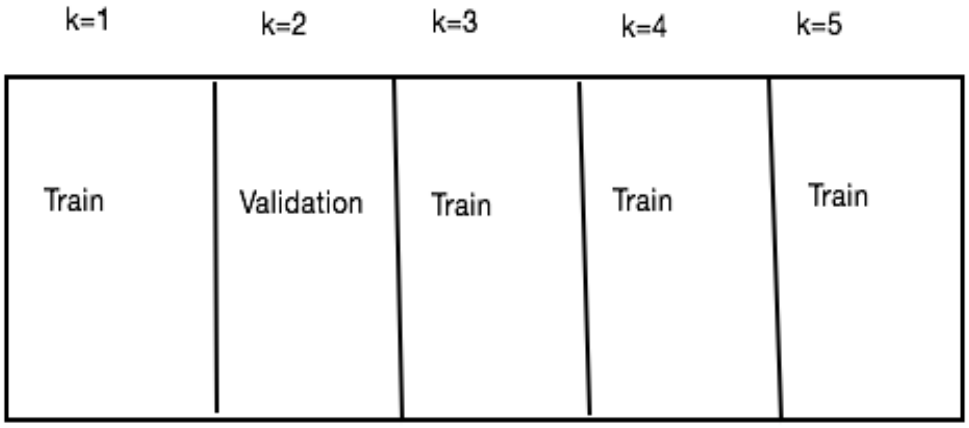

Figure 4.2: Example of cross-validation when $K = 5$.

was given in Equation (4.11) so in this is case

$$\hat{D}(\lambda) = \sum_{k=1}^{K} \hat{D}^{-k}(\lambda) \tag{4.17}$$

where $\hat{D}^{-k}$ is the deviance of the training set where set k is left out.

4.5 Bootstrap

Bootstrapping is a technique that uses the observed data for making inference on parameters, and the goal of the bootstrap method is computer-based implementation of basic statistical concepts (Sartori and Miliani, 2010). The basic idea behind the bootstrap method is to draw $b = 1, ..., B$ samples with size n by random sampling with replacements from the original data set. Because of the replacement an observation can be chosen several times. In each bootstrap sample an observation weight $w_i^* = k/n$ is assigned to each observation, where k is the number of times the observation is chosen in the given bootstrap sample. For each bootstrap sample the quantity of interest is calculated, and based on the B replicates of the quantity it is possible to assess aspects of the distribution. The bootstrap replicates can for example be used to estimate the standard deviation

$$\widehat{SD}_{bootstrap} = \Big(\frac{\sum_{b=1}^{B}(s(x^{*b}) - \frac{\sum_{b=1}^{B} s(x^{*b})}{B})^2}{B-1}\Big)^{1/2} \tag{4.18}$$

where $s(x^{*b})$ is the statistic calculated for each bootstrap sample. The bootstrap replicates can also be used to find a 95% confidence interval for each parameter by sorting the B values for each coefficient and cutting off the lowest 2.5% and the highest 2.5% of the values. The remaining smallest and largest value are the 95% confidence limits. This is called a bootstrap percentile confidence interval.

Three classes of bootstrap methods exist for GLMs. The first class is parametric bootstrap which involves simulating from a fitted parametric model. The second class is semiparametric and involves sampling of model error. The third class and the one that is used in the following analysis is nonparametric bootstrap which involves simulating new data without any distribution assumptions of the original data, and is used in this analysis.

The cross-validation method presented in the previous section is run in each bootstrap sample when doing inference on estimated parameters from lasso-penalized logistic regression. In the cross-validation the n observation is used together with the observation weights and the result is B λ-values that minimizes the deviance for each b. It is also possible to find B values of λ such that the error is 1 standard error from the minimum. A pseudo-code of the bootstrap method with cross-validation is given below.

1: **for** $b = 1 : B$ **do**
2: Draw a sample from the data with replacements of the same size as the data
3: Run cross-validation to choose λ
4: Find the regression coefficients for the given λ
5: **end for**
6: Compute the standard deviation of the parameters
7: Sort and find the 0.025 and 0.975 percentile of the parameters

4.6 Convex optimization

In general a convex optimization problem is

$$\underset{x \in \mathcal{R}^n}{\text{minimize}} \quad f(x) = g(x) + h(x),$$

where g is convex and continuously differentiable function and h is a convex, continuous, but not necessarily differentiable function (Lee et al., 2014). Convex optimization problems have the advantages of being efficiently and reliable solved. Both the lasso regression for linear models and the logistic lasso regression are convex optimization problems since the above requirements holds and the penalty function is not differentiable. Since the lasso method can be used in domains with very large data set, the algorithm for solving the optimization problem needs to be fast and efficient. Implementation of new methods and methods adapted for different l_1-penalized regression models are currently evolving. Least angle regression was earlier the most popular algorithm, but

the pathwise coordinate descent is a new and more efficient method. The pathwise coordinate descent for the lasso start with a large value for λ and slowly decrease the value. The coordinate descent method is defined in Nocedal and Wright (2006) to be a method that cycles through n coordinate directions and performs line search along each direction to obtain new iterates. The method is useful because it does not require the calculation of the gradient of the objective function. Each step is fast and there is an explicit formula for each minimization step. The `glmnet`-package in `R` uses a proximal-Newton iterative approach. The method is also a line search method and computes search directions by minimizing local models $\hat{f}_k$ where the subscript denotes the k-th step. The advantage of the proximal-Newton method is that it converges rapidly near the optimal solution (Lee et al., 2014).

4.7 Limitations of the lasso method

Strong theoretical backing and fast algorithms is the reason for the popularity of the lasso, but there are also major gaps when it comes to the estimation procedure and significance testing (Lockhart et al., 2013). Exact p-values, standard deviations and confidence intervals for the lasso is difficult to find because of the nature of the estimation procedure and since the l_1-penalty is not differentiable. In addition, cross-validation is a random procedure and the value of the chosen λ will vary every time and also the predictors selected by the lasso. The majority of the inference of the regression coefficient is based on resampling of the original data and data splitting. Standard errors and confidence intervals for the parameters from the lasso method can be found using the bootstrap method presented in Chapter 4.5. According to Goeman (2010) the standard errors are not very meaningful since the estimated coefficients are not unbiased, and the standard errors will not tell the whole story. The bias introduced by the lasso itself is therefore a major component of the squared error, and it is impossible to precisely estimate the bias.

Lockhart et al. (2013) propose a significance test of the coefficient in the lasso model that does not employ resampling of the data. The test statistic is based on the fitted values from the lasso method and follows an Exp(1) asymptotic distribution under the null hypothesis that all variables are included in the current lasso model. As no published rigorous theory exist for the logistic models, this will not be explored further in this thesis.

Another shortcoming of the lasso method is that it does not handle highly correlated predictors well. The true model can be recovered if there are no high correlations between relevant predictors and irrelevant predictors. If there is group of highly correlated predictors, the lasso tends to only choose one of the predictors (Zou and Hastie, 2005). It is also arbitrary which of the correlated predictors that is selected. The elastic net penalty can be used as an alternative, and uses a compromise between the ridge penalty from Equation (4.14) and the lasso penalty from Equation (4.13). The elastic net minimization problem is

given as

$$\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}} \quad \Big\{\frac{1}{2}\sum_{i=1}^{n}(Y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda(\frac{1}{2}(1-\alpha)||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1)\Big\}, \qquad (4.19)$$

where $\alpha \in [0, 1]$ and has to be determined in addition to λ. When $\alpha = 1$ the minimization problem corresponds to the lasso method and when $\alpha = 0$ the minimization problem corresponds to the ridge method. The elastic net tends to select the whole group of highly correlated predictors if one predictor in the group is selected, and is particularly useful when $p >> n$ (Zou and Hastie, 2005). The limitations of the lasso method will be discussed further when presenting the results from the Trondheim END study and in Chapter 6.

Chapter 5

Analysis of the Trondheim early neurological deterioration study with the lasso-penalized logistic regression model

In this chapter, the lasso-penalized logistic regression model and related methods presented in Chapter 4 are used to analyze 368 binary responses and 40 centered and standardized predictors from the Trondheim END study. The response vector $\mathbf{Y}$ is a binary variable of length $n = 368$ with $Y_i = 1$ if the i-th patient has END and with $Y_i = 0$ if the i-th patient does not have END. The predictors can be seen in Table 5.1. Figure 5.1 shows the correlation between the predictors. For each summary statistic it can be seen that the standard deviation is correlated to the range. Also, *BSmax-min* is correlated to *BSmax*, and *BSsd* is correlated to *BSmax*. As discussed in Section 4.7, if two predictor are highly correlated, the lasso method only tends to pick one of the correlated predictors. This will be discussed further in Section 5.4 and Chapter 6.

5.1 Fitted model

In Figure 5.2, the coefficient path for each predictor as a function of $\log(\lambda)$ can be seen. The regularization parameter is uniformly spaced on the log scale. The number of nonzero coefficients in each model can be seen on along the top in the figure, and as $\log(\lambda)$ increases the number of predictors in the model decrease. The coefficients can also be plotted against the fraction of deviance explained and the result can be seen in Figure 5.3. The fraction of deviance explained is calculated as

$$D_\lambda^2 = \frac{D_{null} - D}{D_{null}} \tag{5.1}$$

where D is given in Equation (4.10) and D_{null} is the deviance computed when only the intercept is included in the proposed model. The fraction of deviance

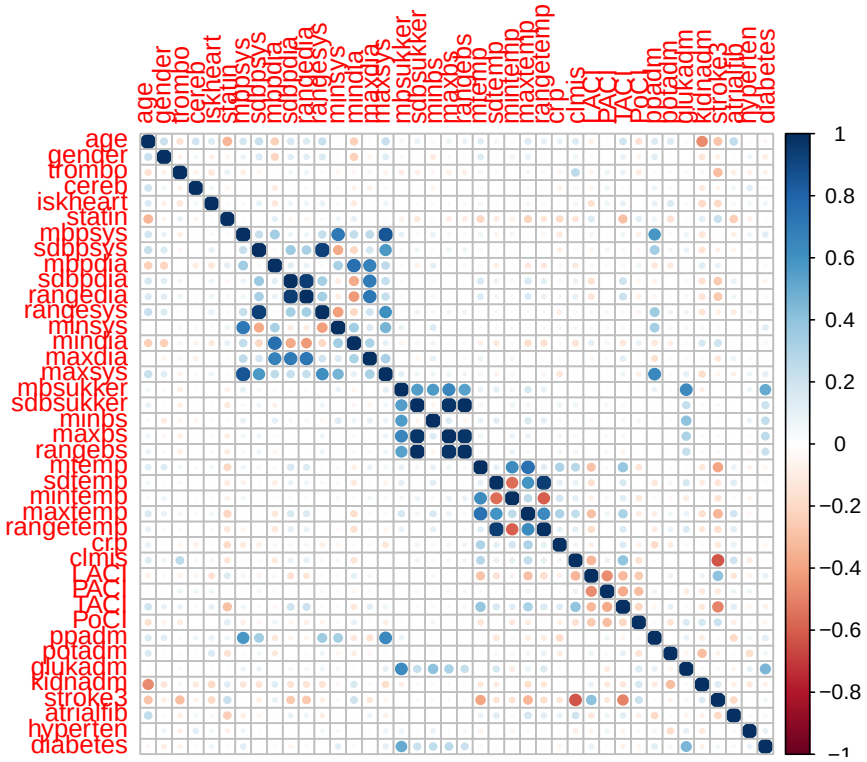

Figure 5.1: Correlation plot of the 40 predictors. The figure is symmetric, dark blue dots or dark red dots correspond to highly correlated predictors and larger dots also correspond to higher correlation.

explained increases as the number of predictors in the model increases.

A plot of the deviance for different values of $\log(\lambda)$ can be seen in Figure 5.4. The red dots in the figure corresponds to the deviance averaged over the 10 folds in the cross-validation at the given value of $\log(\lambda)$, and the error bars are plus and minus one empirical standard deviation of the cross-validated estimates of the deviance. The 10-fold cross-validation method presented in Section 4.4 is used to find the value of $\hat{\lambda}_{min}$ that minimizes the deviance over a sequence of different λ-values and the value of $\hat{\lambda}_{1se}$ such that the error is 1 standard error of the minimum value. The blue vertical line in Figure 5.4 corresponds to $\log(\hat{\lambda}_{min})$ and the green vertical line to the right corresponds to $\log(\hat{\lambda}_{1se})$. The numbers along the top is the number of nonzero coefficient. Choosing the penalty $\hat{\lambda}_{min}$ results in a model with 22 nonzero coefficients and choosing $\hat{\lambda}_{1se}$ results in a model with 8 nonzero coefficients.

The 8 predictors that are included in the model from $\hat{\lambda}_{1se}$ can be seen in Table 5.1. A total of 18 coefficients are estimated to be zero with $\hat{\lambda}_{min}$, and the 22 remaining predictors in the model can be also be seen in the table. The model with the 1 standard deviation error rule is the sparsest of the two and is the easiest to interpret. However, since the analysis of END is relatively new and

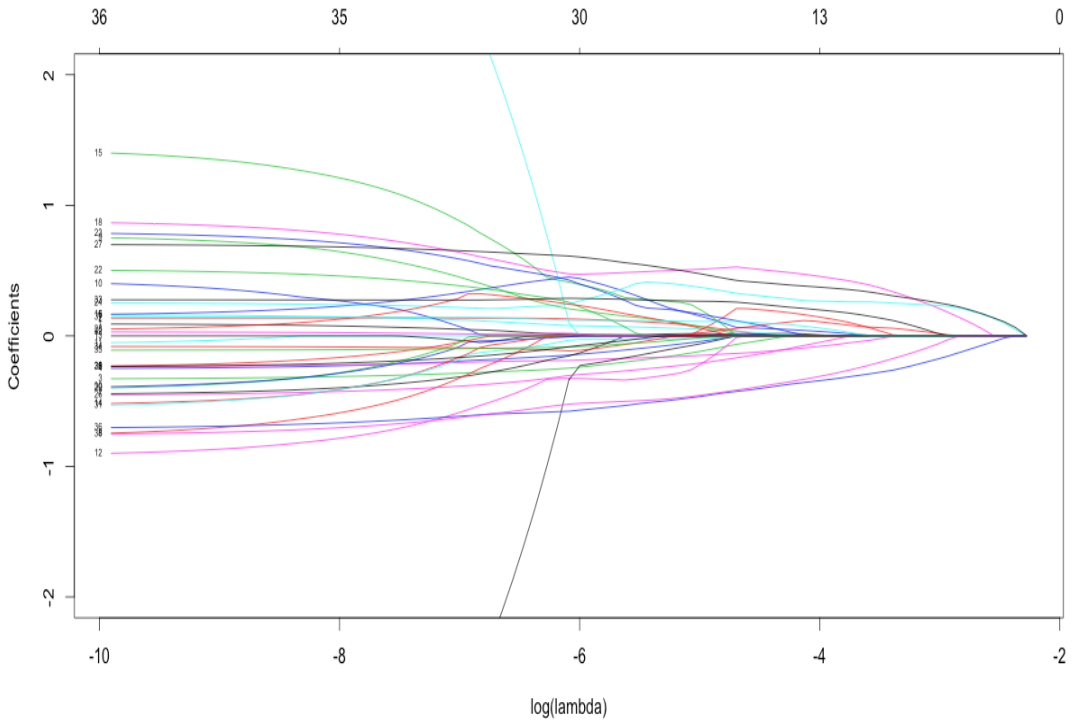

Figure 5.2: Coefficient path for different values of $\log(\lambda)$. To the left there is no penalty on the negative log-likelihood and all covariates are included, and the model complexity decreases as $\log(\lambda)$ increases.

not to exclude any possible important predictors or new predictors, inference is done for the regression coefficients estimated by the less sparse model using $\hat{\lambda}_{min}$. The value of the estimated coefficients based on $\hat{\lambda}_{min}$ can be seen in Table 5.2 and is termed $\hat{\boldsymbol{\beta}}(\hat{\lambda}_{min})$. Note that this is the chosen model, but due to the randomness in the cross-validation using the same strategy a second time could result in another choice of $\hat{\lambda}_{min}$ and thus a different final model. This is what is exploited in the analysis presented in the next section.

Table 5.1: A summary table of the predictor variables included in the model when $\lambda = \hat{\lambda}_{min}$ and when $\lambda = \hat{\lambda}_{1se}$. The penalty parameter $\lambda = \hat{\lambda}_{min}$ estimates 8 regression parameters to be nonzero and $\lambda = \hat{\lambda}_{1se}$ estimates 22 regression parameters to be nonzero.

Predictor variable	$\hat{\lambda}_{min}$	$\hat{\lambda}_{1se}$	Predictor variable	$\hat{\lambda}_{min}$	$\hat{\lambda}_{1se}$
Age			TEMPmean	×	×
Gender	×		TEMPsd		
Thrombolysis	×		TEMPmax-min		
Cerebr.			TEMPmax	×	×
Iskheart.	×		TEMPmin	×	
Statins	×		CPR	×	
SBPmean			Clinical mismatch	×	×
SBPsd			LACI		
SBPmax-min	×		PACI	×	
SBPmax			TACI	×	×
SBPmin			POCI		
DBPmean	×		Pulse pressure		
DBPsd	×		Potassium	×	×
DBPmax-min			Glucose		
DBPmax			Kidney function	×	
DBPmin			Stroke severity	×	×
BSmean	×		Atrial fib.	×	
BSsd			Hypertension	×	×
BSmax-min	×		Diabetes		
BSmax			BSmin	×	×

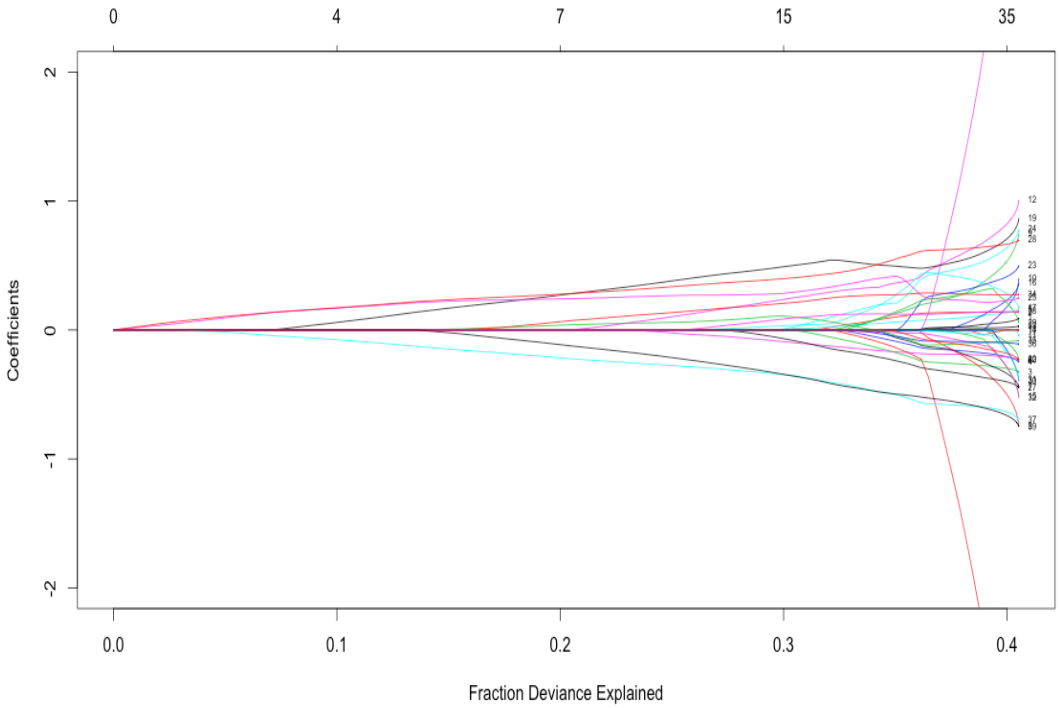

Figure 5.3: Coefficient path where the coefficients are plotted as a function of the fraction of deviance explained.

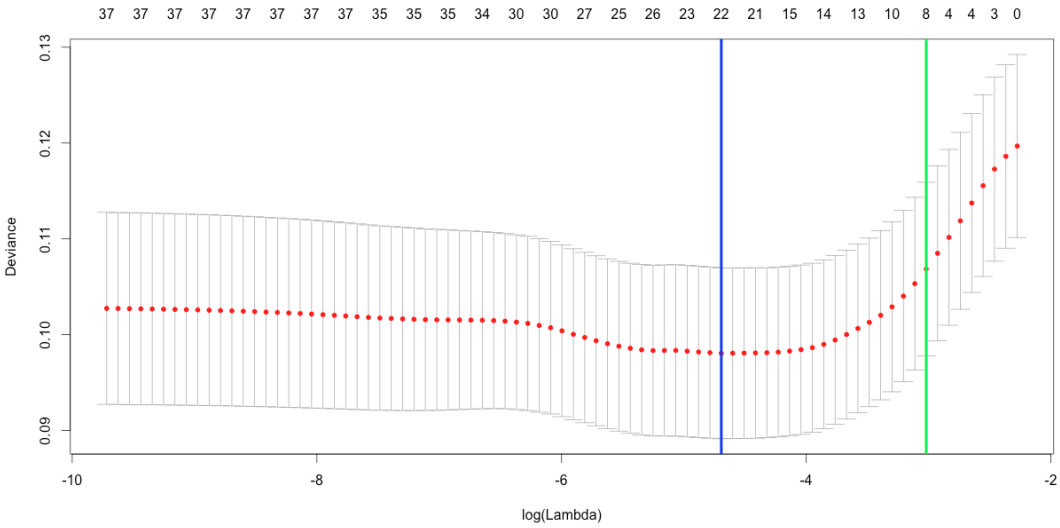

Figure 5.4: Binomial deviance as a function of different values for $\log(\lambda)$. The blue line is the value of λ that minimizes the deviance and the green line is the value of λ such that the error is 1 standard error of the minimum.

5.2 Post-selection inference for the regression parameters

As pointed out earlier, the lasso method is attractive due to its ability to combine variable selection with parameter fitting. Research on approaches for inference for the regression coefficients are currently ongoing and there is not one standard procedure for presenting the result. Thus, the presentation of the result is not as straight forward as for logistic regression and other traditional statistical methods. Since the cross-validation is a random procedure, both the number of selected predictors and the selected predictors will vary if the procedure is repeated. Inference for the estimated parameters is therefore an important part after choosing the model. The behavior of the estimated parameters are explored with the bootstrap method presented in Section 4.5.

For each bootstrap sample the full model fitting procedure is performed, including the cross-validation to choose λ, so we have $\hat{\lambda}^b_{min}$ and regression parameters $\hat{\boldsymbol{\beta}}_b(\hat{\lambda}^b_{min})$. We have saved the results of the B bootstrap samples in a $B \times p$-matrix where each column corresponds to the estimated value of a specific parameter for all the bootstrap samples and each row corresponds to one bootstrap replicate. We have also saved B values for λ that minimizes the deviance. Visualization of the result for all predictors from $B = 1000$ bootstrap realizations of $\hat{\boldsymbol{\beta}}_b(\hat{\lambda}^b_{min})$ can be seen in Figure 5.5. Estimated parameters with boxes away from zero represents the predictors that most often are estimated not to be zero by the lasso-penalized logistic regression method, and boxes with the median at approximately zero are most often not selected by the method. The number of times each parameter is estimated to zero can also be used to analyze the importance of each predictor. Figure 5.6 shows the proportion of times each of the 40 regression parameters were shrunken to zero in the bootstrap samples and Figure 5.7 shows the same, but only for the 22 nonzero parameters of the final model. Table 5.2 shows the percentage each of the parameter estimated to be nonzero, and the intuition is that high percentage means association with END.

From Figure 5.6 it is seen that the maximum blood sugar predictor is almost always estimated to be zero, and that the clinical mismatch predictor is estimated to be nonzero for almost all bootstrap simulations. In the rest of the section the focus will be on $\hat{\boldsymbol{\beta}}(\hat{\lambda}_{min})$, but Figure 5.6 is useful for comparison. For example, the range of the blood sugar is included in $\hat{\boldsymbol{\beta}}(\hat{\lambda}_{min})$, but the predictor is only estimated to be nonzero in 9.1% of the 1000 bootstrap simulations. Also, from Figure 5.6 it can be seen that there are other predictors estimated to be zero more seldom than the range of the blood sugar and is not included in $\hat{\beta}(\hat{\lambda}_{min})$. Because of this, it is expected that the range of the blood sugar is one of the predictors that will shift between inclusion and exclusion of the model for different values of λ. The same holds for the upper predictors in Figure 5.7. To summarize, the similarity between Figure 5.6 and 5.7 is that the predictors at the bottom in the figure is almost always among the predictors selected by

Table 5.2: The first column contains the regression parameters not estimated to be zero, and the other columns represent results based on the 1000 bootstrap samples. The standard deviation is calculated from Equation (4.18) and the CI is the 95% percentile interval presented in Section 4.5.

	$\hat{\beta}(\hat{\lambda}_{min})$	% not 0	$\widehat{SD}$	95% percentile CI
Clinical mismatch	0.063	97.3%	0.025	(0, 0.10)
BSmin	0.057	96.8%	0.021	(0, 0.087)
Hypertension	-0.047	93.6%	0.019	(-0.0723, 0)
Potassium - admission	0.026	82.2%	0.015	(0, 0.050)
TEMPmax	0.044	81.2%	0.024	(0, 0.078)
Stroke severity	-0.0205	70.8%	0.016	(-0.053, 0)
CRP	-0.027	67.8%	0.015	(-0.048, 0)
Statins	-0.020	64.2%	0.015	(-0.049, 0)
BSmean	0.018	63.2%	0.016	(0, 0.052)
SBPmax-min	0.027	60.2%	0.019	(0, 0.055)
TEMPmean	0.012	55.0%	0.020	(0, 0.065)
Atrial fibrillation	0.013	53.2%	0.012	(0, 0.040)
TACI	0.0022	47.5%	0.013	(0, 0.044)
Ischemic heart disease	0.0060	40.7%	0.0095	(0, 0.031)
DBPsd	0.0065	34.9%	0.012	(0, 0.033)
Kidney function - admission	-0.0047	30.7%	0.0073	(-0.025, 0)
TEMPmin	0.0070	29.5%	0.0098	(0, 0.032)
PACI	-0.0093	28.2%	0.0092	(-0.031, 0)
Gender	0.0033	25.5%	0.0070	(-0.0049, 0.023)
DBPmean	0.0076	24.9%	0.012	(0, 0.040)
Thrombolysis	-0.0056	23.4%	0.0077	(-0.028, 0)
BSmax-min	-0.0045	9.1%	0.048	(-0.0178,0)

the lasso to be included in the final model. The difference is that the upper predictors in Figure 5.6 will almost always be excluded from the final model, but the upper predictors in Figure 5.7 will fluctuate between inclusion in the model and being estimated to zero.

The percentile confidence intervals and standard deviations based on the bootstrap samples can be seen in Table 5.2. As discussed in the previous chapter, the standard deviations must be used with caution. The standard deviation can give an impression of great precision, but the bias introduced by the penalization must also be accounted for.

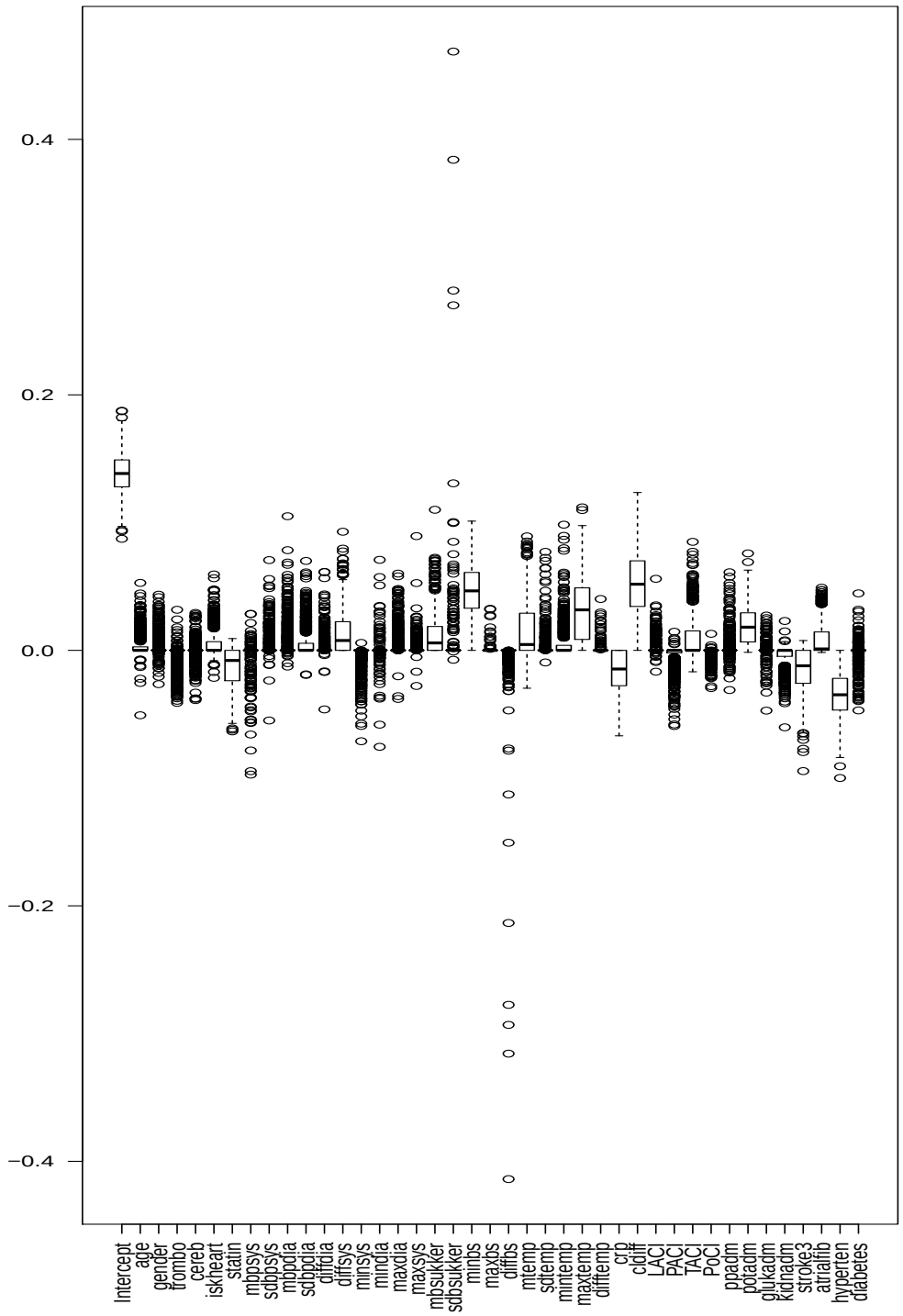

Figure 5.5: Boxplots of 1000 nonparametric bootstrap simulations. The plot visualizes the distribution of the estimated parameters for each predictor variable.

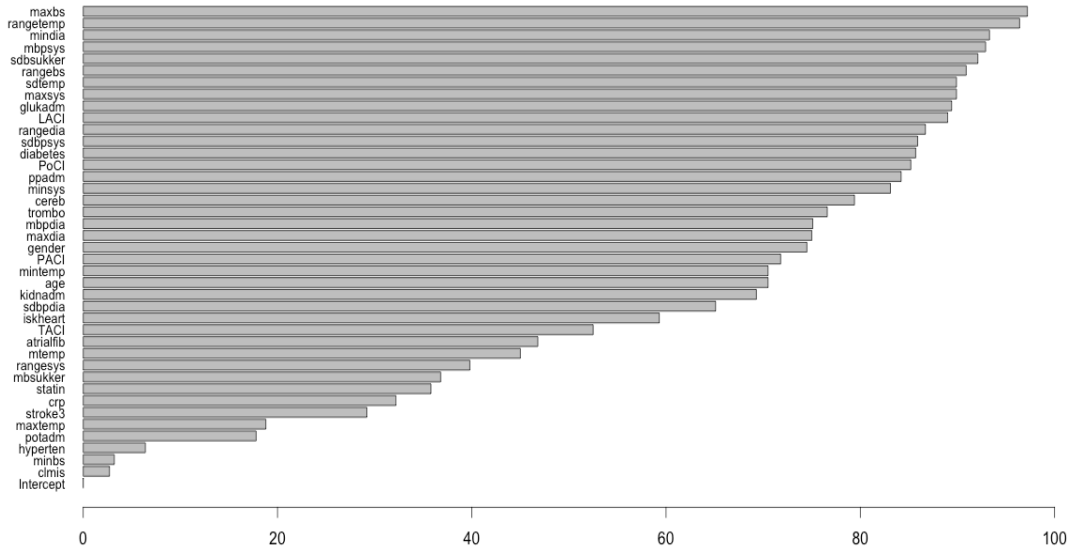

Figure 5.6: The percentage of 1000 bootstrap samples each of the 40 regression parameters are estimated to be zero.

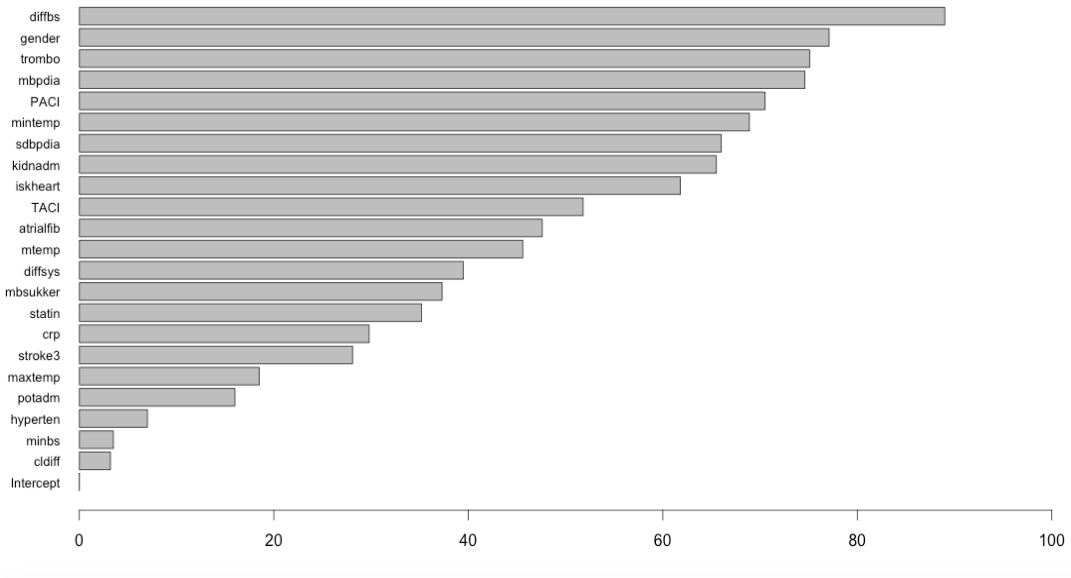

Figure 5.7: The percentage of 1000 bootstrap samples each of the regression parameters selected by the lasso when $\lambda = \hat{\lambda}_{min}$ is estimated to be zero.

5.3 The shrinkage parameter

It is also of interest to explore how the optimal λ is chosen by the 10-fold cross-validation for each bootstrap sample due to the fact that this is the parameter that controls the sparsity of the model. A boxplot and a density plot of $\log(\hat{\lambda}^b_{min})$

that is based on the bootstrap samples can be seen in Figure 5.8. The boxplot shows the median of $\log(\hat{\lambda}^b_{min})$ is approximately -4, and both figures show that $\log(\hat{\lambda}^b_{min})$ is estimated to be around -4 in most cases. However, the minimum value of $\log(\hat{\lambda}^b_{min}) = -6$ and from Figure 5.4 it can be seen that 29 regression parameters are estimated to nonzero and the maximum value of $\log(\hat{\lambda}^b_{min}) = -2.26$ only includes the intercept. Running the cross-validation procedure and ending up with a value of λ near these extreme cases do not predict an adequate model.

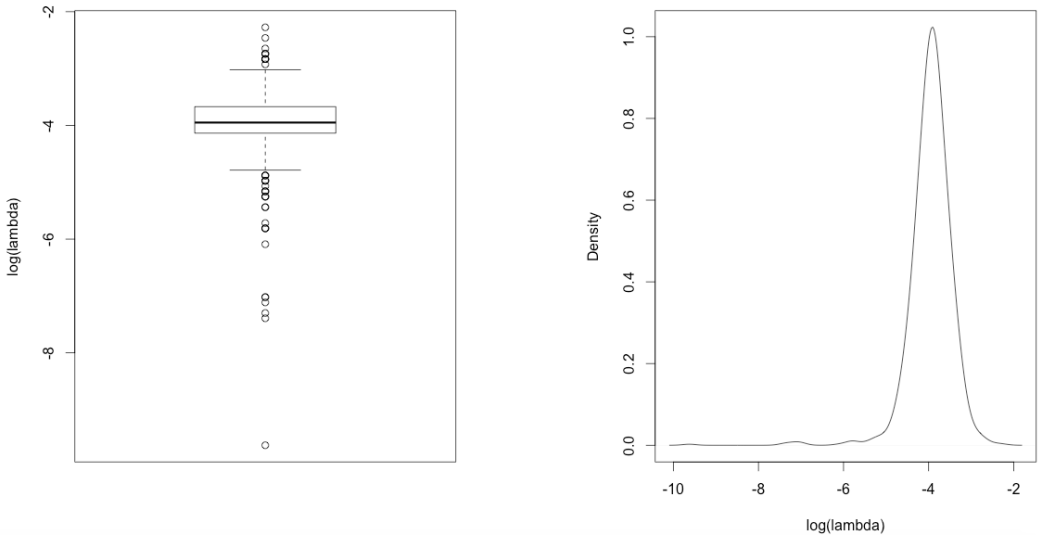

Figure 5.8: Left: A boxplot based on the $B = 1000$ different $\hat{\lambda}^b_{min}$-values. Right: The density distribution of $\hat{\lambda}^b_{min}$ from the 1000 bootstrap samples.

5.4 The correlation problem

A correlation plot of the predictors included in the final model when $\lambda = \hat{\lambda}_{min}$ can be seen in Figure 5.10. Comparing this correlation plot with the correlation plot in Figure 5.1, we see that there are no longer any predictor highly correlated predictors. Thus, figure 5.10 visualizes the fact that the lasso method includes predictors that have low pairwise correlation in the final model. The regression coefficient for *BSmax-min* for example is estimated to be nonzero. As mentioned earlier, this predictor is correlated with *BSmax* and *BSsd*, and as expected these coefficients are estimated to be zero. In comparison, the results from Table 3.3 showed that *BSmax* and *BSsd* were linearly related to END, and it is a possibility that these covariates should also have been included. However, the predictor *BSmax-min* was not shown to be significant. Both *DBPmax-min* and *DBPsd* showed a significant linear trend in Table 3.2. They are correlated

predictors, and only the coefficient for *DBPsd* is estimated to be nonzero in the lasso regression. For the systolic blood pressure, the coefficient *SBPmax-min* is estimated not to be zero, but in this case Table 3.2 showed that the correlated predictor *SBPsd* had a nonsignificant p-value.

To explore the correlation limitation further, *TEMPsd* and *TEMPmax-min* can be used to illustrate this. A plot of the estimated parameters for *TEMPsd* against the estimated parameters for *TEMPmax-min* can be seen in Figure 5.9. Both coefficients are often estimated to be zero at the same time, but when one coefficient is not estimated to be zero, the other coefficient is almost always zero.

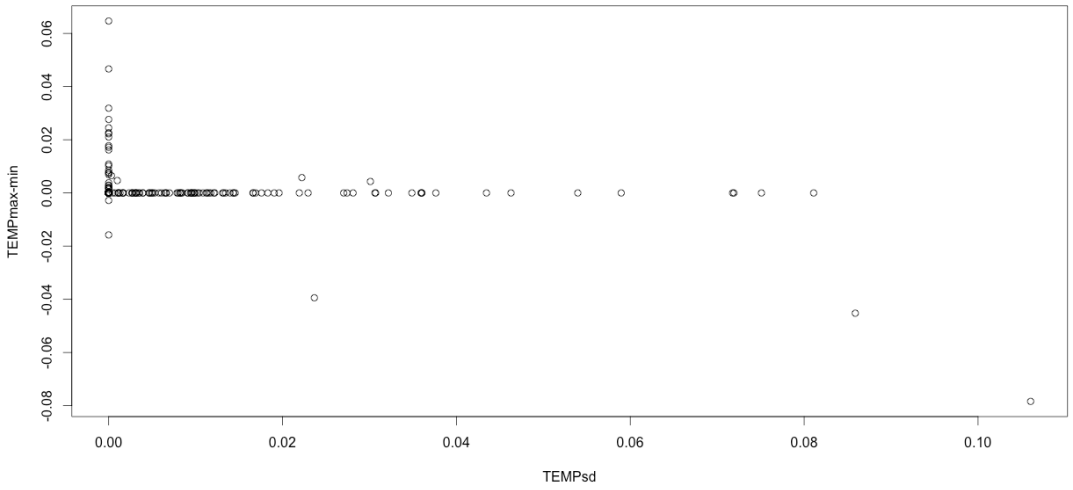

Figure 5.9: The estimated parameters from the bootstrap simulations for *TEMPsd* against the estimated parameters for *TEMPmax-min*.

Another interesting part of the correlation problem is how much correlation the lasso method handles before it ignores a correlated predictor. The correlation between *TEMPmax* and *TEMPmean* is 0.75 and both coefficients are estimated to be nonzero when $\lambda = \hat{\lambda}_{min}$. This is also the highest correlation between to predictors in the chosen model, and may indicate that correlations higher than this will produce problems for the lasso method. Investigation of the correlation between predictors prior to regression coefficient estimation can be used to choose the best method for prediction.

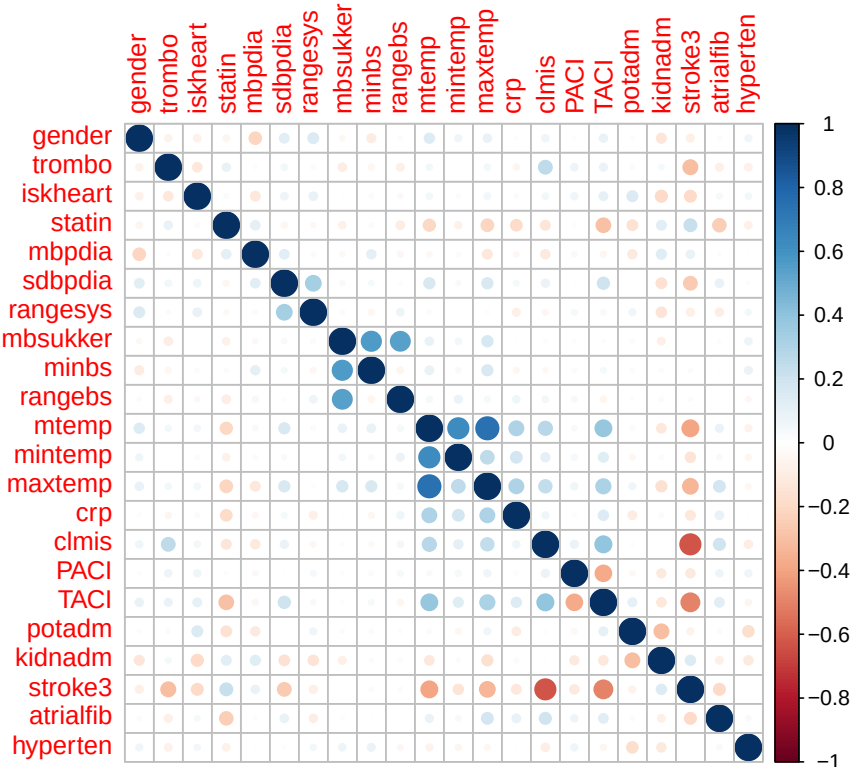

Figure 5.10: Correlation plot of the predictors in the chosen model. The figure is symmetric, dark blue dots or dark red dots correspond to highly correlated predictors and larger dots correspond to higher absolute correlation.

Chapter 6

Discussion and conclusions

In this final chapter we focus on the statistical issues and the medical results. The issue with missing data, the correlation and issues concerning presentation of the result are summarized and discussed in this chapter. In spite of these statistical issues, the medical result and applicability is discussed in the next section.

6.1 Statistical issues

We start by adressing the LOCF imputation issue presented in Section 2.5. The issue is that missing values are imputed with the former value for the time-dependent variables in the Trondheim END study. The LOCF-imputation problem must also be taken into account when discussion the results since our algorithm estimated approximately 1/4 of imputed values of the time-dependent measurements. Because of the imputation there are less variability in the data compared to the reality. Assuming that the LOCF values were removed correctly, the problem is not solved. Logistic lasso regression do not handle data with missing values, and an estimation of the missing values had to be done. Data NMAR is difficult to deal with analytically and the results are often sensitive to the model chosen to impute the missing value. The summary statistics presented in Section 3.1 can be calculated by omitting missing values, but this will not reflect the reality. Another way of dealing with missing data is to remove the patients with missing values would lead to biased estimates and misleading results since the date is not assumed to be MCAR.

When analyzing high dimensional data, the focus is more on prediction than addressing the cause. In predictive studies the power of the predictive model is more important than accurate coefficient estimates. As a result, high correlation between predictors is not problematic if the predictive model as a whole explains the reality in a good way. The elastic net could have been used to overcome the problem with correlated predictors, but then an estimation of the parameter α had to be performed. However, in this analysis the correlation problem is not that crucial since the majority of the highly correlated predictors are measuring

the variability in the same predictor (e.g. blood sugar, temperature and blood pressure). In addition, the preliminary analysis of the summary statistics give an indication of the association to END for the predictors that is not chosen due to high correlation with another predictor.

Only the 4 of the 22 nonzero parameters in the lasso-penalized logistic regression model from Chapter 5 have issues with imputed values, and it is tempting to draw the conclusion that the issues is not as extensive in this analysis. However, the correlation problem restricts the exploration of the impact of the variables based on imputed values since a variable based on imputed values can be excluded from the model due to high correlation with another variable. In Chapter 3, there are no correlation problems and the imputed values reduce the statistical validity of the result for the variables based on imputed values.

Since there are no single correct way to present the result from the logistic lasso regression, different approaches is discussed here. The mean value over all bootstrap simulations $\overline{\hat{\beta}}_{boot} = \frac{1}{B}\sum_{i=1}^{B}\hat{\beta}_b(\lambda^b_{min})$ can also be used to estimate the regression parameters and to overcome the randomness of one cross-validation simulation. This is problematic in two ways. The first problem is that variable selection is not done automatically. The second problem is that the predictors are underestimated compared to the parameters in $\hat{\beta}(\hat{\lambda}_{min})$ since $\hat{\beta}$ is included in the mean. Counting the nonzero regression parameters based on all bootstrap simulations can also be done to measure the importance of the different predictors. The problem is that there are cut-off rule to distinguish between predictors that should be included in the final model and predictors not to be included in the final model.

6.2 Medical results

Since the selection of the shrinkage parameter involves a random procedure, the parameters in $\hat{\beta}(\hat{\lambda}_{min})$ do not represent the whole picture. Some of the nonzero parameters will probably be estimated to zero if we have repeated the cross-validation procedure. However, the bootstrapping say something about the probability of being included in the final model and the parameters included in most of the simulations are associated with END. The parameters included in over 900 of the 1000 bootstrap simulation are clinical mismatch, the minimum value of the blood sugar over 11 measurements, history of hypertension and potassium level measured on admission, and are variables that most definitely should be included in a predicative model for END.

As mentioned in Section 3.1, a study in Bergen found that early neurological deterioration (defined according to NIHSS) and serious consequences for the short-term outcome were associated with low body temperature on admission. Other studies have also found association between measurements on admission and END. In this anslysis, the *BSmin* predictor is the minimum value of 11 measurements and was included in the final model, but only 12% of the patient experienced the minimum value at the first measurement. The *TEMPmax*

predictor was also included in the final model, and only 9.5% of the patient experienced the maximum body temperature on admission. The *TEMPmin* was also included in the final model, and only 27% of the patients had the lowest body temperature on admission. The results from Chapter 5 also shows the importance of monitoring patients closely the first few days after acute ischemic stroke, and that both level and variability of the time-dependent predictors are important. The variability in the blood pressure was of great interest due to the findings in Chung et al. (2015). However, due to the imputed values, the variability in the blood pressure is underestimated and therefore not as important in the analysis as expected. The medical analysis of the Trondheim END study can be extended by looking at different definitions for early neurological deterioration and compare regression parameters from this analysis and results from other studies.

To conclude, Section 3.1 and Chapter 5 show both new results and results in accordance with other studies. Due the amount of missing data the result has limited applicability, but the same important parameters are interesting to look at when planning and analyzing future stroke studies.

Bibliography

Agresti, A. (1966), *An Introduction to Categorical Data Analysis*, first edn, John Wiley and Sons, INC.

Agresti, A. (2002), *Categorical Data Analysis*, second edn, John Wiley and Sons, INC.

Aiyagri, V. and Gorelick, P. B. (2009), 'Management of blood pressure for acute and recurrent stroke', *Stroke* **40**, 2251–2256.

Altman, D. G. and Bland, J. M. (1994), 'Statistics notes: Quartiles, quintiles, centiles and other quantiles', *BMJ* .

Armitage, P. (1955), 'Tests for linear trends in proportions and frequencies', *Biomatrics* **11**(3), 375–386.

Bansal, S., Sangha, K. S. and Khatri, P. (2004), 'Drug treatment of acute ischemic stroke', *Stroke* **35**, 1085–1091.

Bazzano, L. A., He, J., Ogden, L. G., Loria, C., Vupputuri, S., Myers, L. and Whelton, P. K. (2001), 'Dietary potassium intake and risk of stroke in us men and women', *Stroke* **32**, 1473–1480.

Bennette, C. and Vickers, A. (2012), 'Against quantiles: categorization of continous variables in epidemiologic research, and its discontents', *Medical research methodology* .

Bilder, C. R. and Loughin, T. M. (2015), *Analysis of categorical data with R*, CRC Press.

Bingham, N. H. and Fry, J. M. (2010), *Regression - Linear models in statistics*, Springer.

Birschel, P., Ellul, J. and Barer, D. (2004), 'Progressing stroke: Towards and internationally agreed definition', *Cerebrovascular Diseases* **17**(2-3), 242–252.

Christensen, H., Boysen, G. and Truelsen, T. (2005), 'The scandinavian stroke scale predicts outcome in patients with mild ischemic stroke', *Cerebrovascular Diseases* **20**(1), 46–48.

Chung, J.-W., Kim, N., Kang, J., Park, S. H. and Wook-JooKim (2015), 'Blood pressure variability and the development og early neurological deterioration following acute ischemic stroke', *Journal of Hypertension* **33**(1).

Donnan, G. A., Fisher, M., Macleod, M. and Davis, S. M. (2008), 'Stroke', *The Lanclet* **371**(9624), 1612–1623.

Falcone, G. and Chong, J. Y. (2007), 'Gender differences in stroke among older adults', *Geriatrics and Aging* **10**(8), 497–500.

Friedman, J., Hastie, T. and Tibshirani, R. (2010), *Regularization paths for generalized linear models via coordinate descent.*
URL: *URL http://www.jstatsoft.org/v33/i01/*

Goeman, J. (2010), 'L1 penalized estimatin in the Cox proportional hazards model', *Biomatrical Journal* (52), 70–84.

Govan, L., Langhorne, P. and Weir, C. J. (2007), 'Does the prevention of complications explain the survival benefit of organized inpatient(stroke unit) care?', *Stroke* **38**, 2536–2540.

Hastie, T., Tibshirani, R. and Friedman, J. (2008), *The Elements of Statistical Learning*, second edn, Springer.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015), *Statistical Learning with Sparsity, The Lasso and Generalizations*, CRC Press.

Helleberg, B. H., Ellekjær, H., Rohweder, G. and Indredavik, B. (2014), 'Mechanisms, predictors and clinical impact of early neurological deterioration: the protocol of the trondheim early neurological deterioration study', *BMC Neurology* **14**(201).

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Springer.

Knator, D. (2015), 'Stroke', `https://www.nlm.nih.gov/medlineplus/ency/article/000726.htm`. [Online; accessed 22-May-2016].

Lee, J. D., Sun, Y. and Saunders, M. A. (2014), 'Proximal newton-type methods for minimizing composite functions'.
URL: *https://arxiv.org/pdf/1206.1623v13.pdf*

Lindsberg, P. J. and Roine, R. O. (2011), 'Hyperglycemia in acute stroke', *Pharmacotherapy* **31**(11), 363–364.

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, second edn, John Wiley and Sons, INC.

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2013), 'A significance test for the lasso'.

Madsen, R. E., Hansen, L. K. and Winther, O. (2004), 'Singular value decomposition and principal component analysis'.

Martin, L. J. (2015), 'Glomerular filtration rate'.
URL: *https://www.nlm.nih.gov/medlineplus/ency/article/007305.htm*

McHugh, M. L. (2013), 'The chi-square test of independence', *Biochemia medica* **23**(2), 143–149.

Nacu, A., Bringeland, G., Khanevski, A., Thomassen, L., Waje-Andreassen, U. and Naess, H. (2016), 'Early neurological worsening in acute ischaemic stroke patients', *Acta Neurologica Scandinavia* **133**, 25–29.

Napoli, M. D., Papa, F. and Bocola, V. (2001), 'C-reactive protein in ischemic stroke', *Stroke* **32**, 917–924.

Nocedal, J. and Wright, S. J. (2006), *Numerical Optimization*, second edn, Springer.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. and Feinstein, A. R. (1996), 'A simulation study of the number of event per variable in logistic regression analysis', *Journal of Clinical Epidemoiology* **49**(12), 1373–1379.

Pezzini, A., Grassi, M., Zotto, E. D., Volonghi, I., Giossi, A., Costa, P., Cappellari, M., Magoni, M. and Padovani, A. (2011), 'Influence of acute blood pressure on short- and mid-term outcome of ischemic and hemorrhagic stroke', *Journal of Neurology* **258**(4), 634–640.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Rodríguez, G. (2007), 'Lecture notes on generalized linear models'.
URL: *http://data.princeton.edu/wws509/notes/*

Sacco, R. L., Benjamin, E. J., Broderick, J. P., Dyken, M., Easton, D., Feinberg, W. M., Goldstein, L. B., Gorelick, P. B., Howard, G., Kittner, S. J., Manolio, T. A., Whisnat, J. P. and Wolf, P. A. (1997), 'Risk factors', *Stroke* **28**, 1507–1517.

Sartori, S. and Miliani, S. (2010), 'Penalized regression: Bootstrap confidence intervals and variable selection for high dimensional data sets'.

Signorell, A. (2015), *DescTools: Tools for Descriptive Statistics*. R package version 0.99.11.
URL: *http://CRAN.R-project.org/package=DescTools*

Tei, H., Uchiyama, S., Ohara, K., Kobayashi, M., Uchiyama, Y. and Fukuzawa, M. (2000), 'Deteriorating ischemic stroke in 4 clinical categories classified by the oxfordshire community stroke project', *Stroke* **31**, 2049–2054.

Tei, H., Uchiyama, S. and Usui, T. (2007), 'Clinical-diffusion mismatch defined by nihss and aspects in non-lacunar anterior circualtion infraction', *Journal of Neurology* **254**, 340–346.

Thanvi, B., Treadwell, S. and Robinson, T. (2008), 'Early neurological deterioration in acute ischaemic stroke: predictors, mechanisms and management', *Postgraduate Medical Journal* **84**, 412–417.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society* **58**(1), 267–268.

Tibshirani, R. (2013), 'Clustering 1: K-means, k-mediods'.
URL: *http://www.stat.cmu.edu/ ryantibs/datamining/lectures/04-clus1-marked.pdf*

Wall, M. E., Rechtsteiner, A. and Rocha, L. M. (2003), *A practical approach to microarray data analysis.*

Wrotek, S. E., Kozak, W. E., Hess, D. C. and Fagan, S. C. (2014), 'Treatment of fever after stroke: Conflictiong evidence', *Am J Cardiovasc Drugs* **13**(1).

Zhao, J., Zhang, X., Dong, L., Wen, Y. and Cui, L. (2014), 'The many roles of statins in ishcemic stroke', *Current neuropharmacolocy* **12**(6), 564–574.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elstic net', *Journal of the Royal Statistical Society* **67**(2), 301–320.

Appendix A

R-Code

```
1
2  #Generation of the data
3  ds <- read.csv("/Users/MARTHE/Desktop/10.semester/slagdata
       -01022016.csv",sep=";", dec=",")
4  bakgr=read.csv("/Users/MARTHE/Desktop/10.semester/slagdata-
       bakgr.csv",sep=";", dec=",")
5
6
7  age=ds[,4]
8  gender=ds[,3]
9  bsukker=ds[,269:279] #Blood sugar
10 mbsukker=apply(bsukker,1,mean)
11 sdbsukker=apply(bsukker,1,sd)
12 maxbs=apply(bsukker,1,max)
13 minbs=apply(bsukker,1,min)
14 diffbs=maxbs-minbs
15 coeffvarbs=(sdbsukker*100)/mbsukker
16 temp <- ds[,280:290] #Body temperature
17 temp[81,1]=37.1
18 temp[74,5]=36.6
19 temp[290,8]=37.1
20 temp[314,8]=37.4
21 temp=as.matrix(temp)
22 mtemp=apply(temp,1,mean)
23 sdtemp=apply(temp,1,sd)
24 maxtemp=apply(temp,1,max)
25 mintemp=apply(temp,1,min)
26 difftemp=maxtemp-mintemp
27 coeffvartemp=(sdtemp*100)/mtemp
28 bpsys=ds[,seq(93,115,by=2)][,-2] #Systolic blood pressure
29 bpdia=ds[,seq(94,116,by=2)][,-2] #Diastolic blood pressure
30 bpdia[35,11]=bpdia[35,10]
31 mbpsys=apply(bpsys,1,mean)
```

```
sdbpsys=apply(bpsys,1,sd)
mbpdia=apply(bpdia,1,mean)
sdbpdia=apply(bpdia,1,sd)
maxsys=apply(bpsys,1,max)
minsys=apply(bpsys,1,min)
maxdia=apply(bpdia,1,max)
mindia=apply(bpdia,1,min)
diffsys=maxsys-minsys
diffdia=maxdia-mindia
coeffvarsys=(sdbpsys*100)/mbpsys
coeffvardia=(sdbpdia*100)/mbpdia
trombo=ds[,25] #Thrombolysis
stroke5=ds[,333] #Stroke severity, five categories
statin=ds[,33] #Statins
cereb=ds[,369] #Cerebrovascular disease
iskheart=ds[,371] #Ischemic heart disease
pacilaci=as.matrix(ds[,383:386]) #Classification of stroke
    symptoms
clmis=ds[,390] #Clinical mismatch
ppadm=ds[,381] #Pulse pressure
potadm=ds[,37] #Potassium level on admission
potadm[279]=4.1
glukadm=ds[,45] #Glucose level on admission
kidnadm=ds[,417] #Kidney function on admission
sssadm=ds[,291] #SSS level on admission
stroke3=ntile(sssadm,3) #SSS on admission in three categories
atrialfib=bakgr[,20] #Atrial fibrillation
hyperten=bakgr[,21] #History of hypertension
diabetes=bakgr[,25] #History of diabetes
crp=ds[,46] #CRP level on admission

y=ds[,401] #end
x=cbind(age,gender,trombo,cereb,iskheart,statin,mbpsys,sdbpsys
    ,mbpdia,sdbpdia,diffdia,diffsys,minsys,mindia,maxdia,
    maxsys,mbsukker,sdbsukker,minbs,maxbs,diffbs,mtemp,
    sdtemp,mintemp,maxtemp,difftemp,crp,cldiff,pacilaci,
    ppadm,potadm,glukadm,kidnadm,stroke3,atrialfib,
    hyperten,diabetes)
xstand=scale(x,center=TRUE,scale=TRUE)

#LOCF-algorithm

bpdiadiff=matrix(NA,ncol=10,nrow=n)
bpsysdiff=matrix(NA,ncol=10,nrow=n)
bsukkerdiff=matrix(NA,ncol=10,nrow=n)
tempdiff=matrix(NA,ncol=10,nrow=n)
epssdiff <-matrix(NA,ncol=10,nrow=n)
```

```
for (i in 1:10) {
  bpsysdiff[,i]=bpsys[,i+1]-bpsys[,i]
  bpdiadiff[,i]=bpdia[,i+1]-bpdia[,i]
  bsukkerdiff[,i]=bsukker[,i+1]-bsukker[,i]
  tempdiff[,i]=temp[,i+1]-temp[,i]
  epssdiff[,i]=epssmat[,i+1]-epssmat[,i]
}

agreebp <- (bpdiadiff==0)*(bpsysdiff==0)
agreesukkertemp <- (bsukkerdiff==0)*(tempdiff==0)
agreediff5 <- (bpdiadiff==0)*(bpsysdiff==0)

nbpdia <- bpdia[,2:11]
nbpsys <- bpsys[,2:11]
nbsukker <- bsukker[,2:11]
ntemp <- temp[,2:11]
nepss <- epssmat[,2:11]

id0bpdia <- (agreebp==TRUE)
id0bpsys <- (agreebp==TRUE)
id0bsukker <- ((bsukkerdiff==0)==TRUE)
id0temp <- (agreesukkertemp==TRUE)
id0epss <- (agreediff5==TRUE)

nbpdia[id0bpdia==TRUE]=NA
nbpsys[id0bpsys==TRUE]=NA
nbsukker[id0bsukker==TRUE]=NA
ntemp[id0temp==TRUE]=NA
nepss[id0epss==TRUE]=NA

nbpdia <- cbind(bpdia[,1],nbpdia)
nbpsys <- cbind(bpsys[,1],nbpsys)
nbsukker <- cbind(bsukker[,1],nbsukker)
ntemp <- cbind(temp[,1],ntemp)
nepss <- cbind(epssmat[,1],nepss)

sum(is.na(nbpdia))
sum(is.na(nbpsys))
sum(is.na(nbsukker))
sum(is.na(ntemp))
sum(is.na(nepss))

#Quintile analysis
library(dplyr)
library(Hmisc)
library(binom)

```

```
120 plottrend <- function(var1, var2, var3) {
121 gruppe=ntile(var1,5)
122 gruppe2=ntile(var2,5)
123 gruppe3=ntile(var3,5)
124 pros=c()
125 pros2=c()
126 pros3=c()
127 confupp=c()
128 conflow=c()
129 confupp2=c()
130 conflow2=c()
131 confupp3=c()
132 conflow3=c()
133 for (i in 1:5) {
134 prosent=sum(subset(end, gruppe==i))/length(subset(end, gruppe
    ==i))*100
135 prosent2=sum(subset(end, gruppe2==i))/length(subset(end,
    gruppe2==i))*100
136 prosent3=sum(subset(end, gruppe3==i))/length(subset(end,
    gruppe3==i))*100
137 confupp[i]=binom.test(sum(subset(end,gruppe==i)),length(subset
    (end, gruppe==i)), prosent/100)$conf.int[2]
138 conflow[i]=binom.test(sum(subset(end,gruppe==i)),length(subset
    (end, gruppe==i)),prosent/100)$conf.int[1]
139 confupp2[i]=binom.test(sum(subset(end,gruppe2==i)),length(
    subset(end, gruppe2==i)), prosent2/100)$conf.int[2]
140 conflow2[i]=binom.test(sum(subset(end,gruppe2==i)),length(
    subset(end, gruppe2==i)),prosent2/100)$conf.int[1]
141 confupp3[i]=binom.test(sum(subset(end,gruppe3==i)),length(
    subset(end, gruppe3==i)), prosent3/100)$conf.int[2]
142 conflow3[i]=binom.test(sum(subset(end,gruppe3==i)),length(
    subset(end, gruppe3==i)), prosent3/100)$conf.int[1]
143 pros[i]=prosent
144 pros2[i]=prosent2
145 pros3[i]=prosent3
146 }
147 plot(seq(1,5), pros, type="l", ylim=c(0,40), ylab="END (%)",
    xlab="Quintiles", lwd=5)
148 errbar(seq(1,5),pros, confupp*100,conflow*100, add=TRUE)
149 lines(seq(1.1,5.1), pros2, type="l", ylim=c(0,40), col="red",
    lwd=5)
150 errbar(seq(1.05,5.05),pros2, confupp2*100,conflow2*100, add=
    TRUE, col="red")
151 lines(seq(1.1,5.1), pros3, type="l", ylim=c(0,40), col="green"
    , lwd=5)
152 errbar(seq(1.1,5.1), pros3, confupp3*100, conflow3*100, add=
    TRUE, col="green")
```

```
153 }
154
155 par(mfrow=c(1,2))
156 plottrend(meansys, maxsys, minsys)
157 legend("topleft", c("SBPmean","SPBmax", "SBPmin"),
158        col=c("black","red", "green"), cex=0.9, pch=20, pt.cex
    =1, text.font=2)
159 plottrend(diffsys, sdsys, coeffvarsys)
160 legend("topleft", c("SBPmax-min","SBPsd", "SBPcv"),
161  col=c("black","red", "green"), cex=0.9, pch=20, pt.cex=1,
    text.font=2)
162
163 par(mfrow=c(1,2))
164 plottrend(meandia, maxdia, mindia)
165 legend("topleft", c("DBPmean","DPBmax", "DBPmin"),
166        col=c("black","red", "green"), cex=0.9, pch=20, pt.cex
    =1, text.font=2)
167 plottrend(diffdia, sddia, coeffvardia)
168 legend("topleft", c("DBPmax-min","DBPsd", "DBPcv"),
169        col=c("black","red", "green"), cex=0.9, pch=20, pt.cex
    =1, text.font=2)
170
171 par(mfrow=c(1,2))
172 plottrend(meanbs, maxbs, minbs)
173 legend("topleft", c("BSmean","BSmax", "BSmin"),
174        col=c("black","red", "green"), cex=0.9, pch=20, pt.cex
    =1, text.font=2)
175 plottrend(diffbs, sdbs, coeffvarbs)
176 legend("topleft", c("BSmax-min","BSsd", "BScv"),
177        col=c("black","red", "green"), cex=0.9, pch=20, pt.cex
    =1, text.font=2)
178
179 par(mfrow=c(1,2))
180 plottrend(meantemp, maxtemp, mintemp)
181 legend("topleft", c("TEMPmean","TEMPmax", "TEMPmin"),
182        col=c("black","red", "green"), cex=0.9, pch=20, pt.cex
    =1, text.font=2)
183 plottrend(difftemp, sdtemp, coeffvartemp)
184 legend("topleft", c("TEMPmax-min","TEMPsd", "TEMPcv"),
185        col=c("black","red", "green"), cex=0.9, pch=20, pt.cex
    =1, text.font=2)
186
187
188
189 #Trend test
190 trend <- function(summarystat) {
191 matrise=matrix(NA,nrow=5,ncol=2)
```

```
192 gruppe=ntile(summarystat,5)
193 for (i in 1:5){
194   a=sum(subset(end, gruppe==i))
195   b=length(subset(end,gruppe==i))-a
196 matrise[i,1]=a
197 matrise[i,2]=b
198
199 }
200 CochranArmitageTest(x=matrise)
201 }
202
203 #Linear model
204 tepss=c(0,6,12,18,24,30,36,42,48,60,72)
205 epssmat<- as.matrix(ds[,c(187,189:198)])
206
207 regrmat=matrix(ncol=2,nrow=nn)
208 for (i in 1:n)
209 {
210   thisepss=unlist(epssmat[i,])
211   thisds=data.frame("epss"=thisepss,"time"=tepss)
212   thisres=lm(epss~time,thisds)
213   regrmat[i,]=thisres$coeff
214 }
215
216 #PCA
217
218
219 epss.pca=prcomp(scale(epssmat))
220 summary(epss.pca)
221 pca=cbind(epss.pca$x[,1], epss.pca$x[,2])
222 plot(pca, xlab="PC1", ylab="PC2")
223 points(pca[mdet==0,],col="red",pch=20)
224 points(pca[mdet==1,],col="green",pch=20)
225 points(pca[mdet==2,],col="blue",pch=20)
226 legend(8,6.5, c("neither END or EDE","EDE only", "END"),
227        col=c("red","green", "blue"), cex=0.9, pch=20, pt.cex
      =2)
228
229 #Clustering
230 library(cluster)
231 library(fpc)
232
233 rows = apply(epssmat, 1, function(i) length(unique(i)) > 1)
234 epssmat_2=subset(epssmat,rows==TRUE) #Removing patients with
      11 equal measurements
235 epssmat_2=as.matrix(epssmat_2)
236
```

```
237 distcor <- as.dist(1-cor(t(epssmat_2),method="pearson"))
238 nc=pamk(distcor)$nc
239 clusplot(pam(distcor, nc), main="Cluster plot - correlation",
      sub="", col.p="red", plotchar=TRUE)
240
241 disteuc=dist(epssmat, method="euclidean")
242 nc=pamk(disteuc)$nc
243 clusplot(pam(disteuc, nc), main="Cluster plot - euclidean
      distance", sub="", col.p="red")
244
245 #Lasso
246 library(glmnet)
247
248 cvres=cv.glmnet(xstand,y,nfolds=10)
249 coeff=coef(cvres,s="lambda.min")
250 names=rownames((coeff))[which(coeff!=0)]
251
252 #Bootstrapping
253
254 npar=1+dim(xstand)[2]
255 B=1000
256 lambdaminres=rep(NA,B)
257 betamat=matrix(NA,ncol=npar,nrow=B)
258 for (b in 1:B)
259 {
260   this=sample(x=1:n,size=n,replace=TRUE)
261   thisw=rep(0,n)
262   for (i in 1:n) thisw[i]=sum(this==i)
263   thiscvres=cv.glmnet(xstand,y,nfolds=10,weights=thisw,lambda=
      lambdas)
264   lambdaminres[b]=thiscvres$lambda.min
265   betamat[b,]=(coef(thiscvres,s=thiscvres$lambda.min))[1:npar]
266 }
267 colnames(betamat)=c("Intercept",colnames(xstand))
268
269 boxplot(betamat,las=2, horizontal=TRUE)
270 perc0=apply(betamat==0,2,mean)
271 barplot(sort(perc0,decreasing=FALSE),las=1,horiz=TRUE,cex.
      names=0.8, xlim=c(0,1))
272
273 #Percentile bootstrap confidence interval
274 p=length(which(coeff!=0))
275 ul=rep(NA,p)
276 ll=rep(NA,p)
277 for (i in 1:p){
278   ul[i]=quantile(betamat[,which(coeff!=0)[i]], probs=c
      (0.025,0.975))[2]
```

```
279     ll[i]=quantile(betamat[,which(coeff!=0)[i]], probs=c
        (0.025,0.975))[1]
280   }
281   cbind(names, ll, ul)
282
283   #Inference of the shrinkage parameter
284   boxplot(log(lambdaminres), ylab="log(lambda)")
285   plot(density(log(lambdaminres), adjust=2), main="", xlab="log(
        lambda)")
```