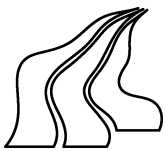
# Meteorological Influence on the Occurrence of Cardiovascular Diseases

Master of Science Thesis

June 1st, 2012

Charlotte Bisgaard and Janne Lund Tidselbak Larsen



Department of Mathematical Sciences, Aalborg University
Department of Cardiology, Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital

Department of Mathematical Sciences, Aalborg University
Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital

**TITLE:**
**Meteorological Influence on the Occurrence of Cardiovascular Diseases**

**SEMESTER:**
Master of Science Thesis
9th and 10th semester

**PROJECT PERIOD:**
September 1st, 2011 -
June 1st, 2012

**WRITTEN BY:**
Charlotte Bisgaard
Janne Lund Tidselbak Larsen

**SUPERVISOR:**
Poul Svante Eriksen

**Number of copies:** 8
**Number of pages:** 125
**Submitted:** June 1st, 2012

**SYNOPSIS:**
In this Master of Science Thesis we investigate a number of meteorological variables and their influence on the daily incidences of cardiovascular diseases (CVDs) in Denmark along with their lagged effect, since earlier studies have shown that the occurrence of incident CVDs varies according to time of year with the highest daily rates of incident CVDs during winter. The thesis is divided into three parts. The first part contains materials and methods, including a description of CVDs and data sources along with validation of diagnoses and an analysis strategy. The second part contains results of the analyses conducted by using generalized additive models (GAMs) and dynamic linear models (DLMs). The third part of the thesis consists of the mathematical theory used for the analyses. Results using GAMs show that temperature has a significant influence on CVDs, but the results are not unequivocally saying that temperature has the highest impact on the daily incidences of CVDs in the winter. Results using DLMs give a more consistent result which indicates that high temperatures have a negative effect on the daily counts of incident CVDs in Denmark.

# Preface

This project is written by Charlotte Bisgaard and Janne L. T. Larsen, group G4-102a in fall of 2011 and spring of 2012 during the 9th and 10th semester of Mathematics at Institute for Mathematical Sciences at Aalborg University in collaboration with Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital and supervisor Poul Svante Eriksen. It is assumed the reader possesses the mathematical qualifications corresponding to completion of the bachelor education of Mathematical Sciences at Aalborg University as a minimum.

We would like to thank Center for Cardiovascular Research for providing data and computer equipment and particularly Claus Dethlefsen, M.Sc., PhD, associate professor and Anette Luther Christensen, M.Sc., PhD student for help when needed.

## Reading instructions

References throughout the report will be presented according to the Vancouver method.
*We* always refers to the project group. Figures, tables, mathematical definitions, etc. are enumerated in reference to the chapter i.e. the first figure in chapter 7 has number 7.1, the second has number 7.2 etc.
The project is divided into three parts: a materials and methods, a results and a theory part. The materials and methods part contains a description of the materials and methods used in the thesis; here data sources, a short presentation of cardiovascular diseases, validation of the diagnoses used for the analysis, and an analysis strategy are given. The second part contains results of the analyses using generalized additive models and dynamic linear models and a synthesis containing a conclusion, a discussion of the materials, method and the results and further perspectives. The last part of the project contains the mathematical theory used through the project.
Mathematical notation and symbols along with further results are assembled in appendices in the end of the project. All illustrations are available at `http://homes.student.aau.dk/cbisga07/`, also the ones not shown in the thesis.

# Contents

# Dansk Resumé

I 2008 skyldtes ca. 30% af alle dødsfald i verden kardiovaskulære sygdomme. Kardiovaskulære sygdomme er en gruppe sygdomme, der omhandler hjertet eller blodkarrene og inkluderer bl.a. akut koronar syndrom, apopleksi, venøs tromboembolisme og atrieflimmer. I forbindelse med kardiovasklulære sygdomme findes der mange risikofaktorer. Disse kan være både genetiske, livsstilsrelaterede og socialt relaterede. Kardiovaskulære sygdomme kan behandles medicinsk og kirurgisk, og specielt den medicinske behandling er blevet mere udbredt gennem de sidste år. I Danmark er gennemsnitsomkostningerne af behandling steget fra ca. en milliard kr. i 1994 til ca. 2 milliarder kr. i 2005, og omkostningerne for indlæggelser ved kardiovaskulære sygdomme i 2005 var næsten fem milliarder kr. Tidligere studier har vist, at antallet af incidenser af kardiovaskulære sygdomme i Danmark er højest om vinteren, og det er derfor oplagt at undersøge, om dette skyldes vejret.

Dette speciale omhandler den meteorologiske indflydelse på de kardiovaskulære sygdomme akut koronar syndrom (blodprop i hjertet), apopleksi (blodprop i hjernen), venøs tromboembolisme (blodprop i benet) og atrieflimmer og er skrevet ved Institut for Matematiske Fag, Aalborg Universitet i samarbejde med Kardiovaskulært Forskningscenter, Aalborg Sygehus, Aarhus Universitet Hospital. Projektet er delt op i tre dele. Første del indeholder materialer og metoder, herunder beskrivelse af de kardiovaskulære sygdomme, datakilder, validering af diagnoser samt en analysestrategi. Anden del indeholder resultater af analyserne, og den sidste del indeholder den matematiske teori brugt til analyserne.

Data brugt i projeketet består af personer i Danmark, der er ældre end 20 år, med incidenser af ovenstående diagnoser i perioden 1.1.1995-31.12.2006 identificeret ved hjælp af det Danske Landspatientregister, samt meteorologiske variable fra DMI. For at undersøge om vejret og incidenser af kardiovaskulære sygdomme har en sammenhæng, laves der i projektet regressionsanalyser vha. hhv. generaliserede additive modeller (GAM) og dynamiske lineære modeller (DLM). I analyserne, der anvender GAM, er

der undersøgt, om alle de givne meteorologiske variable har en sammenhæng med incidenser af kardiovaskulære sygdomme, samt om disse har en lagged effekt herpå. På baggrund af resultaterne herfra bliver der i DLM analyserne kun testet, om temperaturen har en sammenhæng med incidenser af kardiovaskulære sygdomme. GAM analyserne viser, at temperaturen har en indflydelse på alle diagnoser i forskellig grad. De viser også, at indflydelsen er størst ved høje og lave temperature, men det er forskelligt fra diagnose til diagnose, om den er positiv eller negativ. Der er i GAM analyserne ikke noget entydigt svar på den laggede effekt af temperaturen på incidenser af kardiovaskulære sygdomme. DLM analyserne giver et mere realistisk og konsistent billede af, hvordan temperaturens indflydelse er på incidenser af kardiovaskulære sygdomme, også i forhold til lag. DLM analyserne giver et klart billede af, at antallet falder om sommeren, svarende til, at høje temperature har en negativ effekt på antallet af incidenser af kardiovaskulære sygdomme. Begge analyser viser, at ugedagene har en indflydelse på antallet af incidenser med kardiovaskulære sygdomme. Slutteligt beskrives den matematiske teori bag GAM og DLM, herunder maximum likelihood estimering, smoothing, filtrering og inferens.

Appendiks indeholder illustrationer af udvalgte resultater fra analyserne lavet i projektet. Disse og resterende illustrationer kan findes på `http://homes.student.aau.dk/cbisga07/`. Derudover indeholder appendiks en nomenklaturliste.

# Introduction

The number one cause of death globally is cardiovascular diseases (CVDs), which are responsible for approximately 30% of all deaths in 2008. CVDs are a class of disorders involving the heart or blood vessels and include among others acute coronary heart syndrome (ACS), apoplexy (APO), veneous thromboembolism (VTE) and atrial fibrillation (AF) [1]. There are many risk factors associated with CVDs and these can be both genetic and a matter of lifestyle and social relationships. Among them are unhealthy diet, smoking, too much alcohol consumption and to little physical activity. The exposure to risk factors is dependent on gender, age, educational level and economics. Treatment of a CVD consists of medical and surgical treatments and especially the extent of medical treatment has increased in recent years. In Denmark the average expenses of treatment of CVDs has doubled from approximately one billion Danish kroner in 1994 to approximately two billion Danish kroner in 2005 and the expenses for hospitalizations of CVDs in 2005 were almost five billion Danish kroner [2].

Earlier studies have shown that the first time occurrence of a CVD varies according to time of year [3][4] and that the daily rates of incident CVDs are highest during winter in Denmark [5][6]. However, there are conflicting results from country to country of when the daily rate of incident CVDs is at its highest and this could indicate that it varies between climatic areas [7]. Temperature has been recognized as being able to induce health effects [8][9]; therefore it is desirable to achieve a better understanding of the direct and lagged effects of changes in the daily counts of incident CVDs associated with the weather and especially with the temperature, as it could help improve treatment and prevention of CVDs. Therefore the purpose of this study is to investigate a number of meteorological variables and their influence on the daily incidences of CVDs in Denmark.

This study has taken its point of reference in a 2002 study of Braga et al. [10]. Braga et al. conducted a time-series analysis estimating both the acute and lagged effect of weather on respiratory

and cardiovascular deaths in 12 US cities. In epidemiological studies the relationship between the outcome and some variables are expected to be nonlinear; therefore they used additive Poisson regressions for each city, which fit non-parametric smooth functions for these variables. A smooth function of time was used to capture the long-term time-trend and to capture the lagged effect, Braga et al. used a distributed polynomial lag model (PDLM), with the motivation that the weather today can have an influence on deaths not only occurring today, but also on several subsequent days. The basis for the PDLM used was found in a study conducted by Schwartz in 2000 [11], who examined the distributed lag between air pollution and daily deaths, but the model was originally developed by Almon in 1965 in a study examining the distributed lag between the capital appropriations and expenditures [12]. Based on an earlier study [13], Braga et al. chose the lag time to be three weeks and estimated the effect of temperature and humidity. They found that in cold cities both high and low temperatures were associated with an increase in deaths caused by CVDs; furthermore the effect of the low temperatures stayed for days, whereas the effect of high temperatures only was on the same day or the day before. In hot cities neither high nor low temperatures showed significant effect. Nothing consistent showed for humidity, neither in lag nor associated with high or low temperatures. In this study we followed the basic ideas in the study of Braga et al. using additive Poisson regressions with a LOESS smooth function of time and a factor indicating the day of the week and holidays. The lagged effect of the meteorological variables was captured using a PDLM with a third degree polynomial and a three week lag.

Another method of analyzing the daily counts of incident CVDs is by use of dynamic linear models (DLMs). DLMs are a special case of state space models (SSMs), introduced by Harrison and Stevens in 1976, that are both linear and Gaussian. SSMs are a statistical tool appropriate to analyze multivariate time-series and longitudinal data and a great advantage is that they allow the parameters in the model to change over time; this could provide a more clarifying description when analyzing the temperatures impact on incident CVDs. Therefore the data in this thesis is also analyzed using DLMs. The model includes two processes, a latent and an observation process, where the observation process is considered to be indirect observations of the latent process. Assessment of the

latent process is performed by filtering the observations by Kalman filtering techniques and indicates the potential impact on incident CVDs in the population due to the temperature [14]. The lagged effect of the meteorological variables is as in the GAM case captured using a PDLM, but this time with a third degree polynomial, but only 7 days of lag.

The data used in this study was constructed in Denmark from January 1st, 1977 through November 10th, 2011 using the Danish national registry of patients from incident cases of acute coronary syndrome, apoplexy, venous thromboembolism and atrial fibrillation. Furthermore, meteorological data was available on a daily basis from the Danish Meteorological Institute in the same time period.

## Structure of the thesis

The thesis is divided in three parts. The first part contains materials and methods, the second part contains results of the analyses conducted by regressions using GAMs and DLMs and the third part of the thesis consists of the mathematical theory used for the analyses. The chapters in the thesis contain the following:

**Chapter 1** Description of data sources and the cardiovascular diagnoses used for analysis along with validation of the diagnoses. The chapter also includes data preparation and description of data and study design.

**Chapter 2** Analysis strategy for the statistical models.

**Chapter 3** Results of the analysis using generalized additive models.

**Chapter 4** Results of the analysis using dynamic linear models.

**Chapter 5** Synthesis including conclusion, discussion and perspectives.

**Chapter 6** Theory of generalized linear model including estimations of parameters and models for count data.

**Chapter 7** Theory of generalized additive models including smoothing and estimation of parameters via local scoring.

# Part I

# Materials and Methods

# Materials

In this chapter, materials and methods used for analysis in the thesis are presented. Initially, a description of the three data sources, the Central Person Registry, the Danish National Registry of Patients and the Danish Meteorological Institute, is given followed by a description of the chosen cardiovascular diagnoses along with methods of validation of these diagnoses. Lastly, a data description, an outline for data preparation, study design and analysis strategies are presented.

## 1.1 Data Sources

### 1.1.1 The Danish National Registry of Patients

Since 1977, the Danish National Registry of Patients (LPR) has kept information for every contact made with a hospital, and since 1995, outpatient and emergency room contacts have been registered as well. Furthermore, data were added from the psychiatric hospital departments and other separate registries and since 1995, LPR has included information of all contacts made with clinical hospital departments in the entire country. LPR records a lot of information every time a contact is made to the Danish hospital service. It includes, among other things, the civil registry number, hospital and ward name, dates of admission and discharge, surgical procedures and diagnosis, which is specified according to the Danish version of the international classification of diseases, revision 8 (ICD8) and from the beginning of 1995 revision 10 (ICD10). [15][16][17].

### 1.1.2 The Central Person Registry

The Central Person Registry (CPR) contains information about every resident in Denmark since 1968 and on Greenland since 1972. Residents of Faroe Islands are not recorded. All residents are recorded in the CPR from birth in Denmark or by settlement from the Faroe Islands or foreign countries. In CPR all residents have a unique iden-

tification number. Also the CPR contains certain groups in other countries, e.g. persons who are liable to pay taxes. When a person dies or emigrates, their record is kept in the CPR. In 2011, the CPR contained informations about approximately 8.4 million people. The CPR includes, among other things, the date of birth, vital status, civil registry number, civil status, residence and possible emigration of every person registered. [18].

### 1.1.3 Danish Meteorological Institute

The Danish Meteorological Institute (DMI) was established in 1872 and handles the meteorological service of Denmark, the Faroe Islands and Greenland including the surrounding waters and airspaces. This includes surveillance, measuring and collecting information of the weather, climate and other related environmental variables in the atmosphere, on ground and in the ocean. DMI also does research and innovation work in this field. For information, see `www.DMI.dk`. [19].

## 1.2 Cardiovascular Diagnoses

Cardiovascular diseases (CVDs) are a class of disorders involving the heart or blood vessels. It is the number one cause of death globally; The World Health Organization estimates that 17.3 million people died from CVDs worldwide in 2008, corresponding to 30% of all deaths. [1]. Among others, CVDs include atherosclerosis, atrial fibrillation, atrial flutter, coronary artery disease, apoplexy, acute myocardial infarction and thromboembolism.

Atherosclerosis is a condition in which deposits of lipids such as cholesterol cause a thickening of the arterial walls impeding blood flow and reducing elasticity. The result of this is an atherosclerotic plaque, a fatty mass formation in the lumen of the blood vessel. If the condition persists, the inner wall of the vessel becomes swollen with lipids and gaps can appear. To rectify the damage, platelets will begin to adhere to the exposed collagen fibers. The combined accumulation of lipids and platelets form a localized thrombus.[20, pp. 713-714].

Coronary artery disease (CAD) refers to areas of blockage, partial

or complete, of the coronary circulation. A reduction in blood flow to the muscles of the heart will result in a reduction of cardiac performance and thereby reduced circulatory supply, called coronary ischemia. A consequence of CAD can be that cardiac muscle cells die as a result of the lack of oxygen, referred to as acute myocardial infarction (AMI) or heart attack. When the condition is caused by thrombus formation, it is referred to as coronary thrombosis. If the blockage is situated near the start of a coronary artery, the consequences will be comprehensive and the heart might stop beating. [20, p. 682].

Apoplexy (APO) or cerebrovascular accidents are interruptions in the blood supply to a portion of the brain. APO can result in aphasia, sensory and motor paralysis and can affect the ability to draw and interpret spatial relations. [20, pp. 741-742].

Venous thrombosis is the formation of thrombi within a vein, generally in the leg and often due to decreased rate of blood flow e.g. in immobilized patients, damage to the walls of the blood, e.g. after leg fractures, and hypercoagulability, an increased tendency of the blood to clot. When a thrombi is formed in the deep veins of the leg, it is called a deep venous thromboses (DVT). As the veins return blood to the heart, a dislodged thrombus can be transported to the heart and from there to the pulmonary vessel and cause the pulmonary circuit to become blocked. A thrombus that has been dislodged and is transported to another location of the body is called an embolism and the process of forming a thrombus that becomes embolic is called thromboembolism. An embolism that locates in the lungs is a pulmonary embolism (PE). Venous thromboembolism (VTE) refers to both DVT and PE. [20, p. 829].

Fibrillation is a condition of the heart where individual muscle fibers contract independently. Fibrillation can occur in the muscles of both the atria and the ventricles.
Atrial fibrillation (AF) is one of the most common arrhythmias and during AF, the normal electrical impulses that prompt the contraction of the heart muscles are interrupted by disorganized electrical impulses originating in the atria and pulmonary veins. The coordinated contraction is replaced by fibrillation or quivering of the atria. As a result a large amount of the blood in the atria is not moved to

the ventricles; this can increase the predisposition for thrombus formation. [21]. Atrial flutter (AFL) is periodic or permanent presence of rapid regular contractions of the atria of the heart. [22]. Patients with AFL often have AF and vice versa; the two have a close clinical interrelationship between them. [23].

In this thesis we focus on **acute coronary syndrome** (ACS), symptoms related to the heart, **apoplexy**, symptoms related to the brain, **venous thromboembolism**, thromboembolism in the veins and **atrial fibrillation** and therefore also atrial flutter, fibrillation or quivering of the atria. ACS, APO and VTE can be divided into subdiagnoses. ACS is divided into AMI, unstable angina pectoris and cardiac arrest as proposed by Joensen et al. (2009) [24]. Unstable angina pectoris is a condition of chest pain and pressure occurring when the heart does not receive enough blood and oxygen and cardiac arrest is a condition where the heart ceases to contract effectively and thereby ceases to circulate the blood normally. APO is divided into subarachnoid hemorrhage (SAH), intracerebral hemorrhage (ICH), ischemic stroke and unspecified stroke as proposed by Johnsen et al. (2002) [25]. The subarachnoid hemorrhage is located outside the brain in the subarachnoid space between the arachnoid membrane and pia matter, and an intracerebral hemorrhage is located inside the brain. [22]. VTEs are divided into DVT and PE as proposed by Severinsen et el. (2008) [26].

## 1.3 Validation of the Diagnoses

In four articles, Joensen et al. (2009) [24], Johnsen et al. (2002) [25], Severinsen et al. (2008) [26] and Rix et al. (2011) [27], the authors investigated the positive predictive values (PPVs) of ACS, APO, VTE and AF diagnoses, respectively, in the LPR. The studies in the articles are based on the Danish cohort Diet, Cancer and Health (DCH) described by Tjønneland et al. (2007) [28]. In order to make the DCH cohort, 160,725 subjects, 80,996 male and 79,729 female, were invited to participate between December 1993 and May 1997. All subjects were in the age group of 50-64 years and lived in urban areas of Copenhagen or Aarhus, Denmark. None of the subjects had a registered diagnosis of cancer at the time of invitation. In total, 57,053 subjects, 27,179 male and 29,876 female, agreed to participate. For each of these, information about lifestyle and diet was received. [28] [24]. The PPVs were calculated as proportions, i.e. the numer-

ator contained the number of patients with a verified diagnosis after review and the denominator contained the total number of patients registered in the LPR with this specific diagnosis. For high reliability of the diagnoses for this project we have made criteria based on the four studies. The criteria for each of the diseases ACS, APO, VTE and AF are discussed further in the next sections. In table 1.1 the diagnoses we have chosen to include in this project are listed with the corresponding ICD8 and ICD10 codes and estimated PPVs.

| Disease | Diagnoses | PPV |
|---|---|---|
| ACS | 410, 42727, I21, I64 | 65.5% |
| APO | 430-434, 436, I60-I64 | 79.3% |
| VTE | I26, I80 | 58.5% |
| AF | 42793, 42794, I48 | 92.6% |

**Table 1.1:** The selected diagnoses for this project based on studies of Joensen et al. [24], Johnsen et al. [25], Severinsen et al. [26] and Rix et al. (2011) along with their PPVs.

### 1.3.1 Validation of Acute Coronary Syndrome Diagnoses

In the study of Joensen et al. [24], PPVs of ACS diagnoses were investigated in the LPR based on the DCH cohort. Participants to be included in the study were based on available hospital discharge history as participants in the DCH cohort with an ACS diagnosis. They received hospital medical records of 1,577 out of 1,654 patients, i.e. 95.3%, who had been hospitalized with an incident ACS diagnosis in the LPR. 96 patients could not be characterized because either the medical record was not available (n=77) or because the medical record was insufficient to classify the patients (n=19). Both primary and secondary diagnoses were included and patients diagnosed with ACS before entering in the DCH cohort were excluded. Medical records corresponding to the discharge diagnosis and date were retrieved from 54 hospitals. The overall PPV for ACS was 65.5% (95%CI: [63.1;67.9]) after exclusion of patients with missing records. Stratisfied by subdiagnoses, the PPV for AMI diagnoses was found to be 81.9% (95%CI [79.5;84.2]), unstable angina pectoris had a PPV of 27.5% (95%CI: [23.4;31.9]) and cardiac arrest had a PPV of 50.0% (95%CI: [34.2;65.8]). Stratisfying on the type of department of discharge, the PPV for patients receiving an ACS diagnosis in a ward

was 80.1% (95%CI: [77.7;82.3]), whereas the PPV for patients receiving an ACS diagnosis in an emergency room or in an outpatient clinic only was 16.1% (95%CI: [12.4;20.4]). After stratifying by type of diagnosis, i.e. primary and secondary diagnosis, a higher PPV for patients registered with a primary diagnosis was found compared to patients registered with a secondary diagnosis , 67.1% (95%CI: [64.6;69.5]) and 47.0% (95%CI: [37.6;56.5], respectively). [24]. Based on the results of this study we only include primary diagnoses in our project of AMI and cardiac arrest discharged from a ward.

### 1.3.2 Validation of Apoplexy Diagnoses

In the study of Johnsen et al. [25] PPVs of apoplexy diagnoses were investigated in the LPR based on the DCH cohort. Participants who had a diagnosis of a cardiovascular disease before entering the DCH cohort were excluded from the study. Hospital medical records were retrieved and validated for 377 out of 389 patients, i.e. 96.9%, who had been hospitalized with an incident apoplexy diagnosis in the LPR. The sample size is small, hence the results to come are rather imprecise. The overall apoplexy PPV was 79.3% (95%CI: [74.9;83.3]), however the PPV differed between apoplexy subgroups. After stratifying by subdiagnosis, the PPV of SAH and ICH were 48.3% (95%CI: [29.4;67.5]) and 65.7% (95%CI: [47.8;80.9]), respectively. The PPV of ischemic stroke was 87.7% (95%CI: [80.1;93.1]) and of unspecified stroke 76.0% (95%CI: [69.5;81.7]). After stratifying by discharge department the PPV also differed. The diagnoses from non-speciality departments had a PPV of 68.8% (95%CI: [61.3;75.5]) and speciality departments had a PPV of 77.9% (95%CI: [72.3;82.7]), whereas emergency rooms only had a PPV of 48.8% (95%CI: [39.9-57.8]). This trend was also found in all subgroups. When stratisfying by age and gender, the PPV did not differ. Hence, based on this study we chose to include primary discharge diagnoses of SAH, ICH, ischemic stroke and unspecified stroke from ward and outpatient. The discharge diagnoses identified by I62 (other and unspecified nontraumatic intracranical hemirrhage) are also included in our study as proposed by Frost et al. (2006) [3], because many strokes are reported as unspecified in the LPR.

### 1.3.3 Validation of Venous Thromboembolism Diagnoses

In the study of Severinsen et al. [26], PPVs of VTE discharge diagnoses were investigated in the LPR based on the DCH cohort. All incident VTE discharge diagnoses were identified and medical records were retrieved from 1,100 of 1,135 participants, i.e. 96.3% with an incident VTE diagnoses. For the last 35 participants, the medical records were not available. Participants registered before entry in the participants cohort with a diagnosis of VTE were excluded. The PPV of VTE was 58.5% (95%CI:[55.5;61.4]). After being stratified by subdiagnosis, the PPV for DVT was 54.6% (95%CI: [50.9;58.2]), whereas the PPV for PE was 66.5% (95%CI: [62.3;72.3]). Stratifying the discharge diagnoses on department, the PPV of discharge diagnoses at wards was 75.0% (95%CI: [71.9;77.9]), whereas discharge diagnoses from emergency rooms only were 31.3% (95%CI: [27.0;35.8]). Stratifying on diagnosis, i.e. primary or secondary diagnosis, the primary diagnosis had a PPV on 77.0% (95%CI: [73.7;80.1]) and secondary diagnosis had a PPV of 66.5% (95%CI: [58.4;73.8]). When stratifying by subdiagnosis and gender the PPV of DVT diagnosis was higher among men than women, with PPVs of 77.2% (95%CI: [72.2;81.6]) and 63.2% (95% CI: [56.7;69,4]), respectively. The PPVs for PE stratified by gender did not differ, nor did the PPV of VTE diagnosis stratified by age. Based on this study we include both DVT and PE discharge diagnosis from a ward and outpatient. Both primary and secondary diagnoses are included.

### 1.3.4 Validation of Atrial Fibrillation Diagnoses

As mentioned previously AF and AFL have a clinical interrelationship between them, so in the study of Rix et al. (2011) [27] both AF and AFL were validated. The PPVs of AF and AFL discharge diagnoses were investigated in the LPR based on the DCH cohort. A random sample of patients with a AF and/or AFL discharge diagnoses were identified and medical records were retrieved if possible from 150 males and 150 females. Out of these 300 cases, 284 were received, i.e. 94.6%. Of the last 16 participants, the clinical department did not respond (n=4), the hospital or department was closed (n=3), or the hospital records were not possible to retrieve (n=9). Participants registered before entry in the DCH cohort with a diagnosis of AF and/or AFL were excluded. The PPV of AF and/or AFL

was 92.6% (95%CI: [88.8;95.2]). The PPV when stratifying on diagnosis, i.e. primary or secondary diagnosis, gave similar results. For diagnosis stratified on emergency rooms the PPV was significantly lower than in-hospital and out-patient PPVs, with a PPV on 64.7% (95%CI: [39.9;83.5]), where as in-hospital and out-patient was 94.0% (95%CI: [90.4;96.3]). The PPVs for AF and/or AFL stratified by gender did not have a significant difference [27]. Based on this study we include AF and AFL discharge diagnosis from a ward or outpatient in this project. Since the emergency room contact PPV was based on very few cases we chose not to include these contacts. Both primary and secondary diagnoses are included.

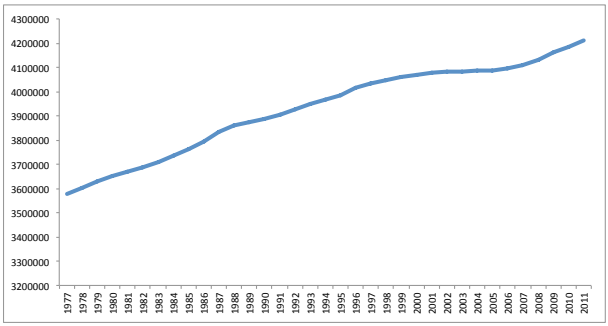
## 1.4 Data

### 1.4.1 Study Period and Population

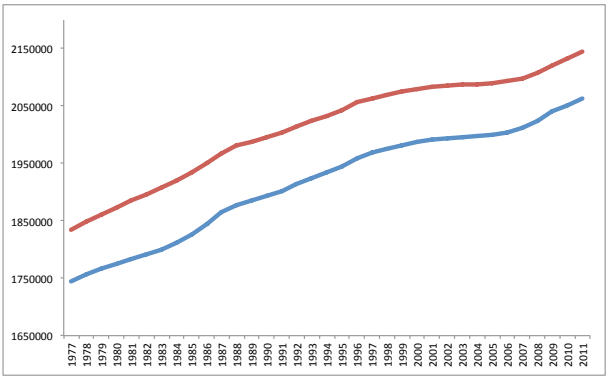
The data for this study was obtained in Denmark from January 1st, 1977 to November 10th, 2011 using the LPR. In this time period the Danish population has grown from approximately 5.08 million people on January 1st, 1977 to 5.56 million people on January 1st, 2011. The population in Denmark consisted of 2.51 million men and 2.57 million women on January 1st, 1977 and on January 1st, 2011 of 2.76 million men and 2.80 million women. Subjects of interest in this study are people of age 20 years and above. The total population of people of age 20 years and above increased from approximately 3.58 million people on January 1st, 1977 to approximately 4.2 million on January 1st, 2011. The population in Denmark in 1977 consisted of approximately 1.74 million men and approximately 1.83 million women on January 1st, 1977 of age 20+ and on January 1st, 2011 of 2.06 million men and 2.15 million women on age 20+, see figure 1.1. [29].

### 1.4.2 Data Preparation

The data preparation was performed in Stata version 11 and as mentioned in section 1.1, the data sources are the LPR, CPR and DMI. The data from LPR is divided in two. The first data contains a record number, classified diagnosis code and information about the diagnosis e.g. is it a primary or secondary. The second data set from LPR consists of additional information e.g. hospital number,




(a) Total Danish population




(b) Total Danish population stratified by gender

**Figure 1.1:** Figure (a) shows the total Danish population from 1977 to 2011 of age 20+. Figure (b) shows the Danish population from 1977 to 2011 stratified by gender of age 20+. The red line represents females and the blue line represents males.

date of admission, date of discharge and municipality codes. The LPR data sets are merged using the record number. By use of the Stata function `FindCardio` written by Anette Luther Christensen subjects having an incident discharge diagnosis of ACS, APO, VTE or AF were identified along with the date of diagnosis. From CPR, information about gender, birth date and possible date of death was merged. An age at diagnosis variable was created using date of diagnosis and birth date. Also the subjects were grouped into two age groups: one containing 20-49 years, and one containing 50+ years. All subjects younger than 20 years were excluded from the data. From DMI we got information about the weather in the ten biggest municipalities in Denmark. If the subject did not live in one of the ten biggest municipalities in Denmark, the subject was

excluded from the data set. Subjects who did not have an incident CVD between 1.1.1995 and 12.31.2006 were excluded, because of the shift from ICD8 to ICD10 in Denmark in 1995 and the municipal merger in 2007. The data set ended up consisting of 104,672 subjects.

The data set was split into four different data set, each containing subjects identified with either ACS, APO, VTE or AF. Since we only include first time occurrences of either ACS, APO, VTE or AF, each subject is only represented in one of the four data sets. In some data sets secondary diagnosis and patient type 2, i.e. outpatient, was deleted according to the validation of diagnoses in section 1.3.

From Statistics Denmark we received information about the Danish population size on January 1st every year since 1977. The population size was linearly interpolated in order to get the population size for each day since January 1st, 1977.

### 1.4.3 Data Description

In this section each of the four data sets for ACS, APO, VTE and AF are described and illustrated.

#### Acute Coronary Syndrome

The ACS data set consists of 20,565 subjects, 12,293 males and 8,272 females, who have had an incident AMI or cardiac arrest found using the diagnosis codes in table 1.1. Subjects in the data set have a median age of 70.19 years (s.d. 13.92), 65.56 years for males (s.d. 13.44) and 76.31 years for females (s.d. 12.91). The age was grouped with a five year interval and the distribution of the incident AMI and cardiac arrest is listed in table 1.2.

Daily incidences of AMI and cardiac arrest from 1.1.1995 to 12.31.2006 is shown in figure 1.2.

The daily count of incident AMI and cardiac arrest in figure 1.2 starts with a small decrease in cases until approximately the year 1996; then it seems stable until approximately the year 2000, where there is a rise in daily incidents of AMI and cardiac arrest. Shortly after the daily cases of incident AMI and cardiac arrest begin to decrease again. The increase in the daily counts of incidences around the year

| | **Male** n (%) | **Female** n (%) |
|---|---|---|
| Age | | |
| 20-24 | 12 (0.10) | 3 (0.04) |
| 25-29 | 41 (0.33) | 19 (0.23) |
| 30-34 | 96 (0.78) | 25 (0.30) |
| 35-39 | 216 (1.76) | 65 (0.79) |
| 40-44 | 471 (3.83) | 129 (1.56) |
| 45-49 | 837 (6.81) | 203 (2.45) |
| 50-54 | 1.233 (10.03) | 305 (3.69) |
| 55-59 | 1.478 (12.02) | 465 (5.62) |
| 60-64 | 1.562 (12.71) | 630 (7.62) |
| 65-69 | 1.603 (13.04) | 787 (9.51) |
| 70-74 | 1.427 (11.61) | 1.165 (14.08) |
| 75-79 | 1.450 (11.80) | 1.341 (16.21) |
| 80-84 | 1.052 (8.56) | 1.400 (16.92) |
| 85-90 | 581 (4.73) | 1.103 (13.33) |
| 90-94 | 204 (1.66) | 525 (6.35) |
| 95+ | 30 (0.24) | 107 (1.29) |
| Total | 12.293 (59.78) | 8.272 (40.22) |

**Table 1.2:** Distribution of incident AMI and cardiac arrest cases with respect to age for 20,565 subjects in Denmark from 1.1.1995 to 12.31.2006.

2000 can be a consequence of changes in the diagnostic definitions for AMI, that happened in 2000. [2] [24]. However Joensen et al. [24] found no change in the PPV before and after 2000, therefore we have not stratified for the change in this study. The daily counts decrease again after 2003.

In figure 1.3 the distribution of incidences are shown in percent with respect to to age groups. It shows that male incidences of AMI and cardiac arrest peak in the age group 65-69, whereas female incidences does not peak until the age group 80-84. This is also evident by the median age of male and female incidences, where the median age was 65.56 years for males and 76.31 years for females. Generally the males tend to have more incidences than females in the age approximately from 20-75 and in this time period the difference is increasing.

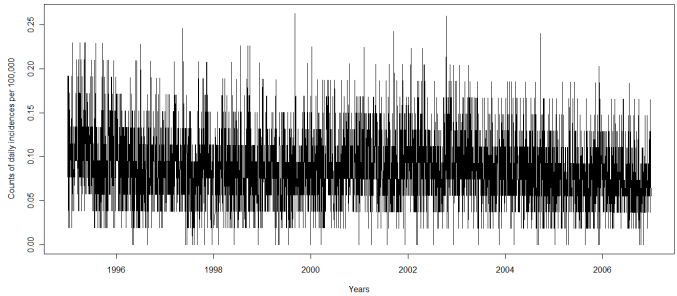



**Figure 1.2:** Daily counts of incidences of AMI and cardiac arrest per 100,000 from 1.1.1995-12.31.2006.

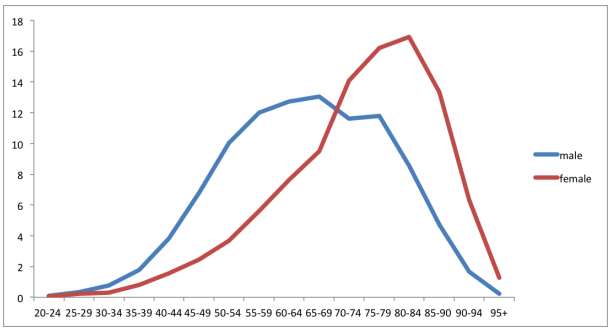



**Figure 1.3:** Incidences of AMI and cardiac arrest from 1.1.1995-12.31.2006 in percent with to respect to age groups.

## Apoplexy

The apoplexy data set consists of 29,822 subjects; 13,614 men and 8,272 females, who have had an incidence of APO found using the diagnosis codes in table 1.1. Subjects in the data has a median age at 73.80 years (s.d. 14.41), 69.53 years for males (s.d. 13.78) and 77.09 years for females (s.d. 14.40). The age was grouped with a five year interval and the distribution of the incident apoplexy is listed in table 1.3.

Daily incidences of apoplexy from 1.1.1995 to 12.31.2006 are shown in figure 1.4.

The daily counts of incidences of apoplexy per 100,000 in figure 1.4 decrease from approximately year 2000. In 1995, the outpatient contacts started to be recorded in the LPR. This would have increased the number of recorded cases from 1995 and in the following years

| | **Male** | **Female** |
|---|---|---|
| | n % | n % |
| Age | | |
| 20-24 | 44 (0.32) | 56 (0.36) |
| 25-29 | 80 (0.59) | 98 (0.60) |
| 30-34 | 125 (0.92) | 129 (0.80) |
| 35-39 | 175 (1.29) | 208 (1.28) |
| 40-44 | 387 (2.84) | 346 (2.13) |
| 45-49 | 629 (4.62) | 490 (3.02) |
| 50-54 | 1,029 (7.56) | 623 (3.84) |
| 55-59 | 1,430 (10.50) | 816 (5.03) |
| 60-64 | 1,473 (10.82) | 1,041 (6.42) |
| 65-69 | 1,588 (11.66) | 1,297 (8.00) |
| 70-74 | 1,841 (13.52) | 1,946 (12.01) |
| 75-79 | 1,949 (14.32) | 2,770 (17.09) |
| 80-84 | 1,598 (11.74) | 2,855 (17.61) |
| 85-90 | 939 (6.90) | 2,290 (14.13) |
| 90-94 | 291 (2.14) | 1,020 (6.29) |
| 95+ | 36 (0.26) | 223 (1.38) |
| Total | 13,614 (45.65) | 16,208 (54.35) |

**Table 1.3:** Distribution of incident APO cases with respect to age for 29,822 subjects in Denmark from 1.1.1995 to 12.31.2006.

compared to years prior to 1995.

In figure 1.5 the distribution of incidences are shown in percent with respect to age groups. It shows that male incidences of APO peak in the age group 70-74, whereas female incidences do not peak until the age group 80-84. This is also evident by the median age of the males and females, where the median age was 69.53 years for males and 77.09 years for females. From approximately age 40-70 the females have lower counts of incidences than males and from 70-95+ the males have a lower counts of incidences than females. There is a sizable change in both curves for age before and after 44 years old.

### Venous Thromboembolism

The VTE data set consists of 14,045 subjects, 5,852 men and 8,193 females, respectively, who had an incidence of VTE found using the diagnosis codes in table 1.1. Subjects in the data has a median age of

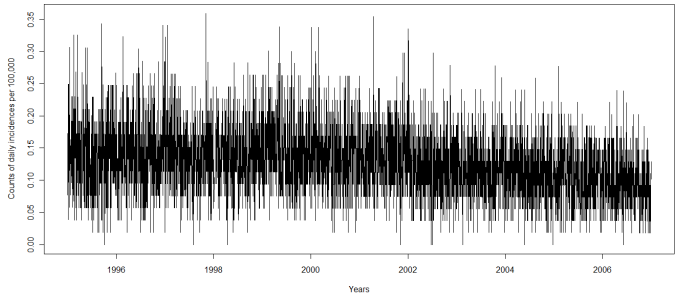



**Figure 1.4:** Daily counts of incidences of apoplexy per 100,000 from 1.1.1995 to 12.31.2006.

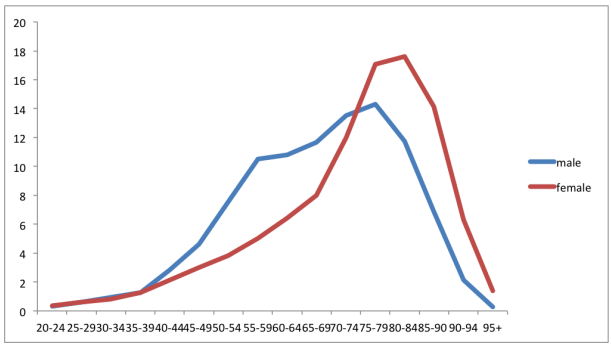



**Figure 1.5:** Incidences of apoplexy from 1.1.1995-12.31.2006 in percent according age groups, blue line representing males and red line representing females.

64.75 years (s.d. 18.79), 59.99 years for males (s.d. 16.92) and 68.76 years for females (s.d. 19.80). The age was grouped with a five year interval and the distribution of the incidence VTE is listed in table 1.4.

Daily incidences of VTE from 1.1.1995 to 12.31.2006 are shown in figure 1.6.

The daily cases of incidences of VTE per 100,000 in figure 1.6 seem to be slightly increasing over the years.

In figure 1.7 the distribution of incidences are shown in percent with respect to age groups. It shows that male incidences of VTE peak in the age group 50-54, whereas female incidences does not peak until approximately the age group 80-84. This is also evident by the median age of the males and females, where the median age was

| | Male | Female |
|---|---|---|
| | n % | n % |
| Age | | |
| 20-24 | 97 (1.66) | 273 (3.33) |
| 25-29 | 177 (3.02) | 357 (4.36) |
| 30-34 | 285 (4.87) | 359 (4.38) |
| 35-39 | 349 (5.96) | 356 (4.35) |
| 40-44 | 388 (6.63) | 358 (4.37) |
| 45-49 | 423 (7.23) | 409 (4.99) |
| 50-54 | 568 (9.71) | 432 (5.27) |
| 55-59 | 640 (10.94) | 483 (5.90) |
| 60-64 | 552 (9.43) | 574 (7.01) |
| 65-69 | 561 (9.59) | 678 (8.28) |
| 70-74 | 555 (9.48) | 805 (9.93) |
| 75-79 | 556 (9.50) | 1,030 (12.57) |
| 80-84 | 386 (6.60) | 953 (11.63) |
| 85-90 | 226 (3.86) | 762 (9.30) |
| 90-94 | 81 (1.38) | 290 (3.54) |
| 95+ | 8 (0.14) | 74 (0.90) |
| Total | 5,852 (41.67) | 8,193 (58.33) |

**Table 1.4:** Distribution of incidences of VTE cases with respect to age for 14,045 subjects in Denmark from 1.1.1995 to 12.31.2006.

59.99 years for males and 68.76 years for females. From age 20-69 the daily counts of incidences of VTE for females is lower than for males. Especially in the age group 55-59 there is a great difference in the two curves.

### Atrial Fibrillation

The AF data set consists of 30,457 subjects; 14,380 men and 16,077 females, who have had an incidence of AF found using the diagnosis codes in table 1.1. Subjects in the data has a median age at 76.7 years (s.d. 13.67), 71.8 years for males (s.d. 14.28) and 80.2 years for females (s.d. 11.71). The age was grouped with a five year interval and the distribution of the incident AF is listed in table 1.5.

Daily incidences of AF from 1.1.1995 to 12.31.2006 is shown in figure 1.8.

The daily cases of incidences of AF per 100,000 in figure 1.8 is in-

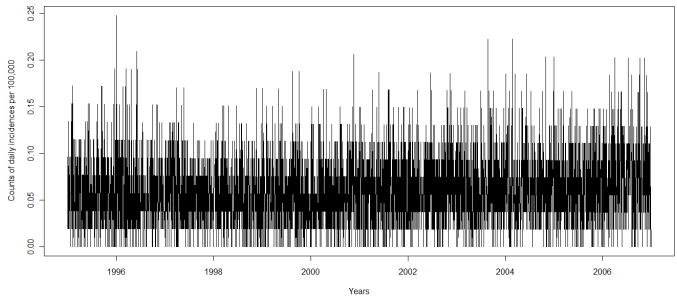



**Figure 1.6:** Daily counts of incidences of VTE per 100,000 from 1.1.1995 to 12.31.2006.

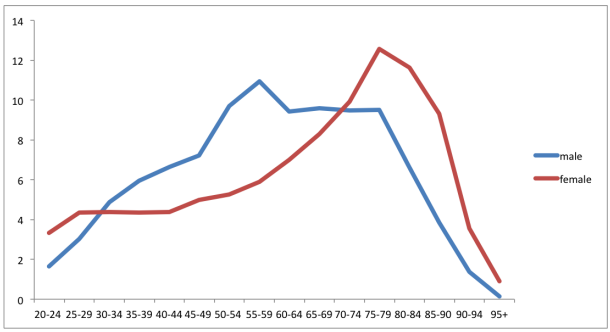



**Figure 1.7:** Incidences of VTE from 1.1.1995-12.31.2006 in percent with respect to age groups. The blue line represents males and the red line represents females.

creasing until approximately year 2001. After this the daily cases seem stable.

In figure 1.9 the distribution of incidences are shown in percent with respect to age groups. It shows that male incidences of AF peak in the age group 75-79, whereas female incidences do not peak until the age group 80-84. This is also evident by the median age of the males and females, where the median age was 71.8 years for males and 80.2 years for females. From approximately 20-75 years the females have lower counts of incidences than males and from 75-95+ years the males have lower counts of incidences than females. There is a sizable change in both curves for age before and after 44 years old.

| | **Male** | **Female** |
|---|---|---|
| | n % | n % |
| Age | | |
| 20-24 | 74 (0.51) | 18 (0.11) |
| 25-29 | 110 (0.76) | 33 (0.21) |
| 30-34 | 139 (0.9) | 51 (0.32) |
| 35-39 | 204 (1.42) | 67 (0.42) |
| 40-44 | 313 (2.18) | 94 (0.58) |
| 45-49 | 523 (3.64) | 153 (0.95) |
| 50-54 | 884 (6.15) | 328 (2.04) |
| 55-59 | 1,244 (8.65) | 527 (3.28) |
| 60-64 | 1,445 (10.05) | 794 (4.94) |
| 65-69 | 1,616 (11.24) | 1,184 (7.36) |
| 70-74 | 1,905 (13.25) | 1,869 (11.63) |
| 75-79 | 2,216 (15.41) | 2,800 (17.42) |
| 80-84 | 1,888 (13.13) | 3,263 (20.30) |
| 85-90 | 1,221 (8.49) | 3,012 (18.73) |
| 90-94 | 501 (3.48) | 1,493 (9.29) |
| 95+ | 97 (0.67) | 391 (2.42) |
| Total | 14,380 (47.2) | 16,077 (52.8) |

**Table 1.5:** Distribution of incidences of AF with respect to age for 30,457 subjects in Denmark from 1.1.1995 to 12.31.2006.

### DMI Data

From DMI we have data about the following meteorological variables; temperature, humidity, wind velocity, atmospheric pressure and downpour from the 10 largest weather stations in Denmark. In table 1.6 the mean, minimum, maximum and standard deviation of each variable is shown. Each meteorological variable is calculated as the mean of the measurements from 10 meteorological variables.

The following figures show a spaghetti plot of the season for every year and the mean of all the meteorological variables with a trend curve. The same figures for maximum and minimum of the meteorological variable can be found in appendix B.

Figure 1.10 shows the mean temperature and mean humidity. Both the temperature and humidity trends seem to be the same through the years with an increase in temperature in the summertime and decrease in the temperature in the wintertime and the opposite in

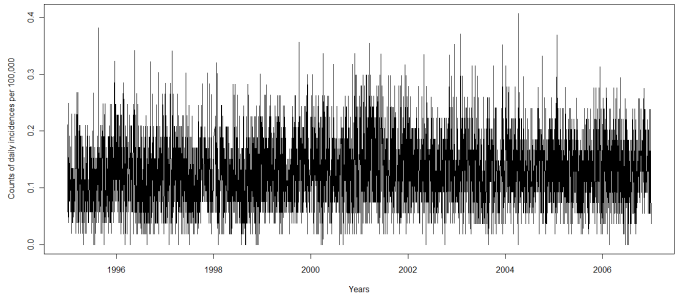



**Figure 1.8:** Daily counts of incidences of incidences per 100,000 of AF from 1.1.1995 to 12.31.2006

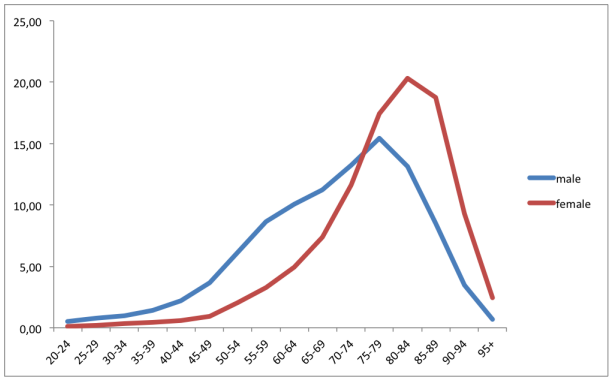



**Figure 1.9:** Incidences of AF from 1.1.1995-12.31.2006 in percent with respect to age groups. The blue line represents males and red line represents females.

humidity.

Figure 1.11 shows the mean wind velocity and mean atmospheric pressure in the ten biggest municipalities in Denmark from 1.1.1995-12.31.2006. The wind velocity have a small increase at summertime and small decrease in the wintertime. It does however looks like the wind velocity is decreasing during the years. The atmospheric pressure trend seems to be the same through the years and it does not seem to change much during the year.

Figure 1.12 shows the mean downpour in the ten biggest municipalities in Denmark from 1.1.1999-12.31.2006. The downpour data was not available from before 1999.

The downpour trend seems to be the same through the years, with

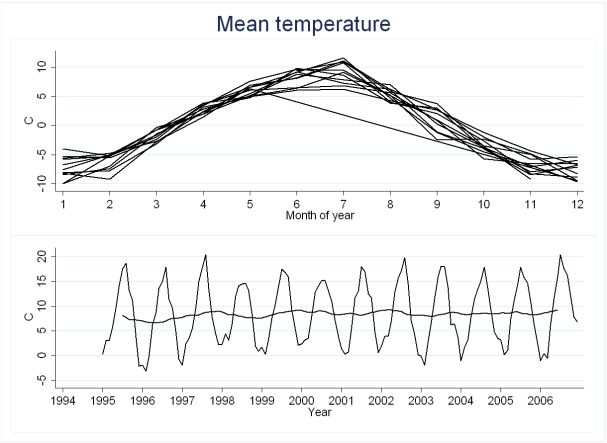
| | Min. | Mean | Max. | sd |
|---|---|---|---|---|
| Temp. | -13.3°C | 8.28°C | 26.1°C | 6.73°C |
| Humidity | 36% | 83.13% | 100 % | 10.34% |
| Wind | 0 m/s | 4.88 m/s | 15.5 m/s | 2.09 m/s |
| Pressure | 968 hPA | 1013.75 hPA | 1045.7 hPA | 10.00 hPA |
| Downpour | -0.1 mm | 1.65 mm | 70 mm | 3.80 mm |

**Table 1.6:** Mean, minimum, maximum and standard deviation of each meteorological variable. Each meteorological variable is calculated as the mean of a measurement from the 10 largest weather stations from 1.1.1995 to 12.31.2006.

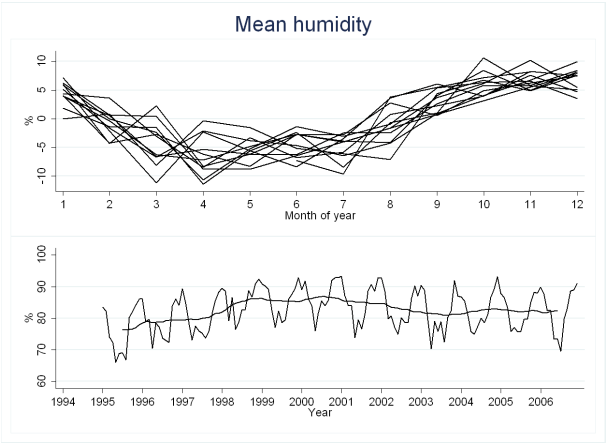
an increase in January 2006.

## 1.5 Study Design

In epidemiology a cohort is defined as a group of individuals who are followed over a given time; the purpose is to measure incidences of a disease in the study cohort. In our study the selected individuals to be observed, i.e. the study population or cohort, are all inhabitants in Denmark, and the period of time in which the study population is observed, i.e. the study period, is from 1.1.1995 to 12.31.2006. To be included in the study, the subjects must fulfill some predefined criteria, also called the inclusion criteria. In our study all subjects has to be 20+ years old. This criterion is included to avoid interference of possibly different pathology in children compared with adults. Also, subjects can only enter in one of the four data sets. An additional characteristic of the cohort is, that the study population may change during the study period, this is referred to as being an open cohort. The primary endpoint of our study is incidents of ACS, APO, VTE and AF and the study is merely observed through registries. [30] [31].

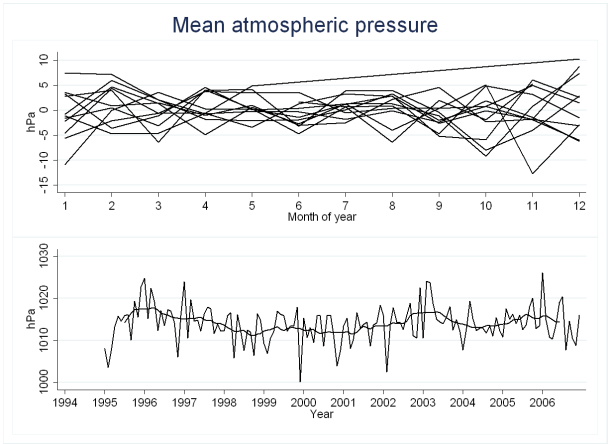



(a) Mean temperature.

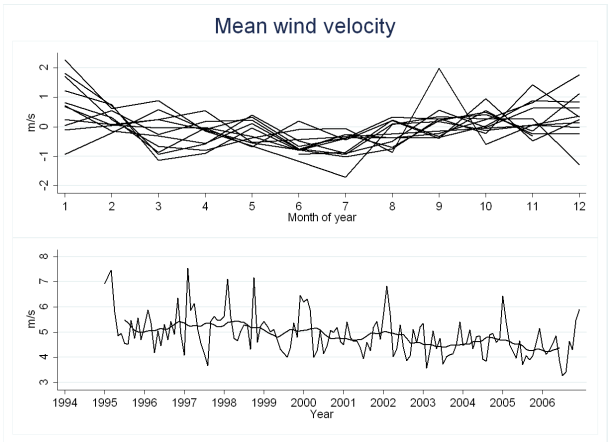



(b) Mean humidity.

**Figure 1.10:** Mean temperature and humidity in the ten largest municipalities in Denmark from 1.1.1995-12.31.2006. The first figure shows a spaghetti plot of the season through the years and the last figure shows the temperature and humidity with trend through the years.




(a) Mean atmospheric pressure.




(b) Mean wind velocity.

**Figure 1.11:** Mean air pressure and wind velocity in the ten largest municipalities in Denmark from 1.1.1995-12.31.2006. The first figure shows a spaghetti plot of the season through the years and the last figure shows the atmospheric pressure and wind velocity with trend through the years.

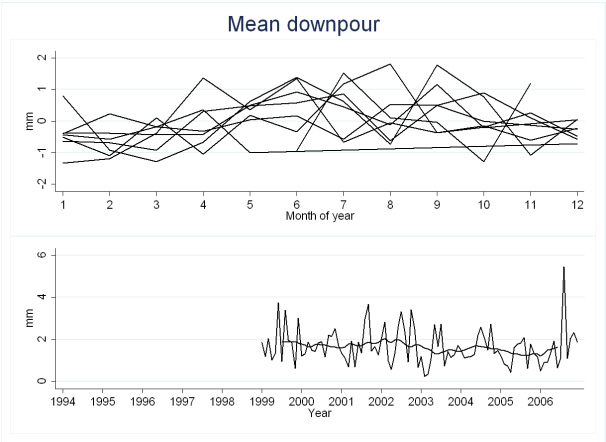



**Figure 1.12:** Mean downpour in the ten largest municipalities in Denmark from 1.1.1999-12.31.2006. The first figure shows a spaghetti plot of the season through the years and the last figure shows the downpour with trend through the years.

# Methods

The purpose of this study is to examine a possible effect of meteorological variables on CVDs. For a first occurrence of incident diagnosis of ACS, APO, VTE or AF it is desired to know:

- Does any of the meteorological variables show an effect on the daily incidences of ACS, APO, VTE and AF?
- If there is an effect, does it change if the value of the variable is high or low?
- Is there a lagged effect?

Besides this it is desired to know if the effects differ according to the disease. First an analysis was made using generalized additive models (GAMs). Another analysis is done by the use of dynamic linear models (DLMs). The mathematical theories behind these models are described later in the third part of the thesis.

## 2.1 Analysis Strategy

In this section the line-up of the two models are given. The models are analyzed using procedures in R.

### 2.1.1 Generalized Additive Models

GAMs are an extension of generalized linear models (GLMs) and specifies effects on the natural parameter scale as additive but possible nonlinear. The GAMs are chosen because it allows regressions to include non-parametric smooth functions to model a potential nonlinear dependence of the log-frequency of daily counts of incidences of either ACS, APO, VTE or AF on the meteorological variables. In our model the response $Y$ is the frequency of daily incidences of ACS, APO, VTE or AF, so $\{Y_t\}$, where $t = 1, \ldots, N$, are possibly serially correlated count data. Subsequently we assume $\{Y_t\}$ is Poisson distributed with intensity parameter $n_t \cdot \mu_t \cdot 100000$, where $\mu_t \cdot 100000$ is the intensity per 100,000.

The GAM is on the form

$$\log(\mathbb{E}[Y_t]) = \text{effects of covariates} + s(t), \tag{2.1}$$

where $t$ is the time, $Y_t \sim \text{Po}(n_t \cdot \mu_t \cdot 100000)$, $n_t$ is daily counts of the number of residents in Denmark at time $t$ and is included in the link function as an offset variable $\log(n_t \cdot 100000)$ and is found by using linear interpolation. $\mathbb{E}[Y_t]$ is the expected value of the frequency of daily incidences of ACS, APO, VTE or AF, $s(\cdot)$ are a smooth function of time. For $s(\cdot)$, we used LOESS, a running-line regression smoother. This allow us to model the potential nonlinear dependence of the counts of daily incidences on time and captures the long-term time trend in the data. A factor indicating the day of the week was included as a covariate. To capture the long-term seasonal variation in the data the model also included a seasonal component as a covariate; $\sin\left(\frac{2\pi t}{T}\right) + \cos\left(\frac{2\pi t}{T}\right)$, where $T$ is the period of seasonality examined, one year in our case, and $t$ is our time variable.

A parallel analysis was performed for each of ACS, APO, VTE and AF. To see the effect of the meteorological variables individually, they are included in an updated model one at the time. The updated model includes a lag model to see how far back in time the variable have an effect, the lag chosen is three weeks i.e. lag 20. The lag model included is a polynomial distributed lag model (PDLM) and the realization behind it is that the meteorological variables can affect the daily counts of incidences not only today, but also on several subsequent days, but in such a way that the effect is modeled by a polynomial.

### Polynomial distributed lag model

The unconstrained Poisson distributed lag model assumes, modulo season and long-term time-trends, that

$$\log(\mathbb{E}[Y_t]) = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_L X_{t-L}, \tag{2.2}$$

where $X_{t-i}$ is the meteorological variable $i$ days before the incidence, $L$ is the assumed maximum lag and where $\boldsymbol{\beta}$ is to be estimated. The lag period could be long and therefore cost a lot of degrees of freedom (d.o.f).

The effects of the meteorological variables on daily incidences of ACS, APO, VTE or AF are usually assumed nonlinear [10], with J-, U-, or

V-shaped relations, therefore we have used a linear and a quadratic term for the meteorological variables at each lag. This means that our lag model has the form

$$\log(\mathbb{E}[Y]) = \alpha \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \mathbf{X} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_L \end{pmatrix} + \mathbf{X}^2 \begin{pmatrix} \gamma_0 \\ \vdots \\ \gamma_L \end{pmatrix},$$

where $X^2$ means that all the entrances in $X$ is squared and e.g. $L = 20$

$$\mathbf{X} = \begin{bmatrix} X_{21} & X_{20} & \dots & X_2 & X_1 \\ X_{22} & X_{21} & \dots & X_3 & X_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_N & X_{N-1} & \dots & X_{N-19} & X_{N-20} \end{bmatrix}.$$

i.e. the first column is the temperature on the same day and the last column is the temperature on lag 20.

There exists correlation between the meteorological variables on days close together, so regression (2.2) will have a high degree of collinearity. This will produce poor estimates of the distribution of the effect over lag. The most common approach to overcome this problem is to set at restriction of how the $\beta_i$'s will evolve over time. This is done by fitting the $\beta_i$'s with lag number to a polynomial function. We used third-degree polynomial constraints for the linear and quadratic temperature terms, because it should be flexible enough to include any plausible pattern of delayed effect over time, so

$$\begin{aligned} \boldsymbol{\beta} &= \tau_1 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \tau_2 \begin{pmatrix} 1 \\ \vdots \\ L \end{pmatrix} + \tau_3 \begin{pmatrix} 1 \\ \vdots \\ L \end{pmatrix}^2 + \tau_4 \begin{pmatrix} 1 \\ \vdots \\ L \end{pmatrix}^3 \\ &= \begin{pmatrix} 1 & 1 & 1^2 & 1^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 21 & 21^2 & 21^3 \end{pmatrix} \begin{pmatrix} \tau_1 \\ \vdots \\ \tau_4 \end{pmatrix} = p_3 \begin{pmatrix} \tau_1 \\ \vdots \\ \tau_4 \end{pmatrix}. \end{aligned}$$

and similar for the quadratic term $\boldsymbol{\gamma}$. This means, modulo the intercept, that

$$\begin{aligned} \log(\mathbb{E}[Y]) &= \mathbf{X}\boldsymbol{\beta} + \mathbf{X}^2\boldsymbol{\gamma} \\ &= \mathbf{X}p_3\boldsymbol{\tau} + \mathbf{X}^2 p_3 \boldsymbol{\varphi}. \end{aligned}$$

By doing this we will have to estimate four parameters for both the linear and the quadratic term leaving us with a 8 d.o.f. surface of the effect of a meteorological variable during the past 21 days on daily incidences from each specific cause. [10] [11].

For the lag model the effect at each lag and its variance was estimated as well as the overall effect and its variance.

An analysis using the difference in the meteorological variables from day to day was also conducted following the same procedure.

The original model and the updated model are tested by use of ANOVA.

### 2.1.2 Dynamic Linear Models

Again the four diseases ACS, APO, VTE and AF are examined separately, i.e. we have four univariate state space models, each describing a disease. The daily counts of cases divided by the number of residents in Denmark at the given time follow Poisson distributions. In order to apply the theories of dynamic linear models, we transform to normally distributed data by raising the relative daily counts to the power of 2/3. [32]

Each observation series of daily counts is treated as a sum of independent components. These components describe the day of the week (incl. an indicator for holidays), a seasonal component and the effect of temperature on daily counts. The seasonal component is described by

$$\cos\left(\frac{2\pi t}{T}\right) + \sin\left(\frac{2\pi t}{T}\right),$$

where $t$ is the time measured in days since 12.31.1994 and $T$ is the period of seasonality examined, i.e. one year. The analyses were performed with components consisting of both the temperature and the temperature squared.

In the first analyses, each of the eight levels of the day of the week component were allowed to have individual variances. We also performed analyses where the variances of the weekday levels were assumed to be the same, with the same arguments as in section 2.1.1.

We also examined the lagged effect of temperature on the daily counts of each disease. This was done similarly to the method used in the GAM analysis using a polynomial distributed lag model, described in section 2.1.1, this time with lags of 0-7 days prior to the event.
The dynamic linear regression model for observations $Y_t$ on covariates is $x_{1,t}, \ldots, x_{p,t}$ is outlined in section 10.4 and described by

$$Y_t = \sum_{j=1}^{p} x_{j,t}\beta_{j,t} + v_t.$$

If we concentrate on the effect of temperature, the explanatory variables are the temperature (here denoted $x_t$) and the temperature squared ($x_t^2$) for each of the eight days of interest. Therefore, in our case, $F_t = [x_t, x_{t-1}, \ldots, x_{t-7}, x_t^2, x_{t-1}^2, \ldots, x_{t-7}^2]$ for each time $t$.

We create binary indicators of weekdays

$$\mathbf{z}_t = (z_{1,t}, \ldots, z_{8,t})$$

and a vectors containing the eight weekday levels (one for each of the seven days of the week and one for holiday)

$$\boldsymbol{\rho}_t = (\rho 1,t, \ldots, \rho_{8,t})^\top .$$

If we let $X_t$ denote the vector of temperatures 0-7 days before time $t$, i.e.

$$X_t = (x_t, \ldots, x_{t-7}),$$

then the dynamic linear regression has the form

$$\begin{aligned} Y_t = \quad & \boldsymbol{\zeta}_1 \cos\left(\frac{2\pi t}{T}\right) + \boldsymbol{\zeta}_2 \sin\left(\frac{2\pi t}{T}\right) \\ & + \mathbf{z}_t\boldsymbol{\rho}_t + X_t p_3 \boldsymbol{\tau}_t + X_t^2 p_3 \boldsymbol{\varphi}_t + v_t \end{aligned}$$

where $\boldsymbol{\beta}_t = p_3\boldsymbol{\eta}$ and $\boldsymbol{\gamma}_t = p_3\boldsymbol{\varphi}$ and where all states $\zeta_{1,t}$, $\zeta_{2,t}$, $\rho_t$, $\eta_t$ and $\varphi_t$ have independent random walks. It is assumed that $\zeta_{1,t}$ and $\zeta_{2,t}$ have the same variance.

# Part II

# Results

# Generalized Additive Models

In the following, results from the analyses of daily incidences of CVDs are presented. Results and figures are outlined thoroughly in the first result section regarding ACS. APO, VTE and AF results are given without comments, but are interpreted in the same way as ACS. In this chapter results from the analysis with GAMs are presented. The mathematical theory behind GAM is outlined in chapter 7 and the analysis strategy in section 2.1.1.

## 3.1 Acute Coronary Syndrome

The daily counts of incidences of ACS were fitted by the use of a GAM, including a time-trend described by a LOESS smoother. The model also included a linear interpolation of the residents in Denmark from 1.1.1995-12.31.2006 as an offset variable aswell as a factor indicating the day of the week. To capture the long-term seasonal variation in the data the model also included a seasonal component as a covariate; $\sin\left(\frac{2\pi t}{365.25}\right) + \cos\left(\frac{2\pi t}{365.25}\right)$, where $t$ is our time variable. To study the effect of the meteorological variables on daily counts of incidences of ACS, the meteorological variables were included in the model one at a time, and to see if the variable significantly improved the model an ANOVA test was applied.

Figure 3.1 shows the trend in ACS incidences over time.

The $y$-axis on figure 3.1 is on the linear scale and since the family is Poisson with canonical link, it is on the log scale and 0 corresponds to the average number of daily counts. The average value of daily counts of incidences of ACS is 4.74. From 1995 to 1998 there is a small decrease in the incidences of ACS of approximately 10% to the average value of daily cases. From 1998 to approximately 2002, the curve is slightly increasing with approximately 7% from the average value of daily counts. From 2002 and until late 2006 there is a decrease of approximately 20% from the average value of daily

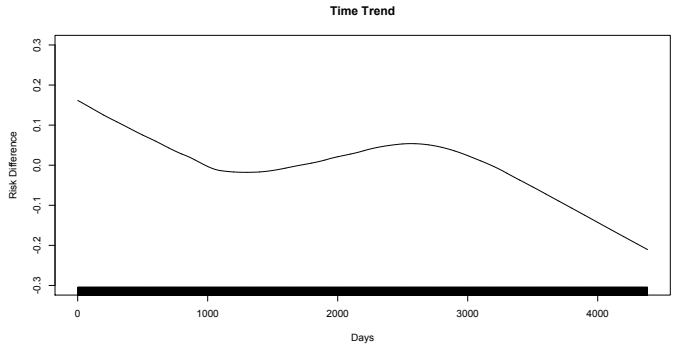



**Figure 3.1:** Trend of ACS incidences over time, with an average value of daily cases of incidences of ACS of 4.74.

counts of incidences of ACS.

Figure 3.2 shows the influence of the day of the week and holidays on daily counts of incidences of ACS. The day of the week

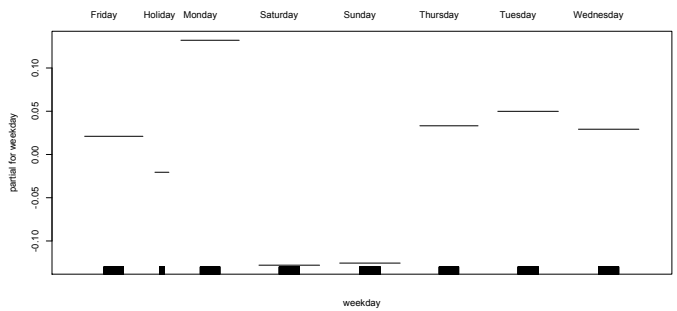


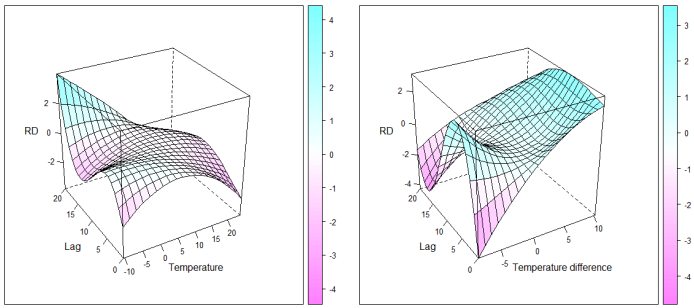

**Figure 3.2:** Influence of the day of the week on daily counts of incidences of ACS.

has an influence on incidences of daily counts of ACS with a p-value of $< 2.2e - 16$ and deviance difference of 135.75, with Monday having the largest influence, with an increase of approximately 10% and the weekends the least with a decrease of approximately 10%. Holidays only show a slight decrease of 2%.

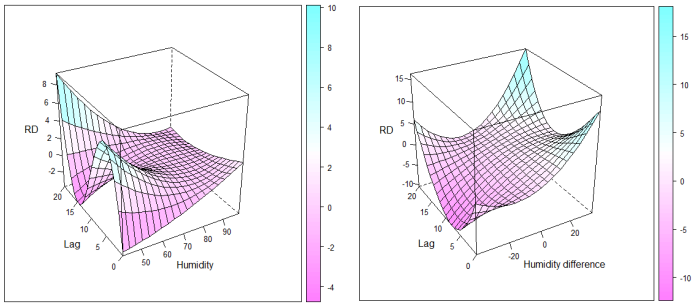
Figure 3.3 and figure 3.4 show the tendencies of the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ values of the PDLM for the effect over time of the five meteorological variables. In the left side of the figure the linear

and squared effects of the meteorological variables on the daily counts of incidences are shown and in the right side the effects of the differences from day to day in the meteorological variables are shown.

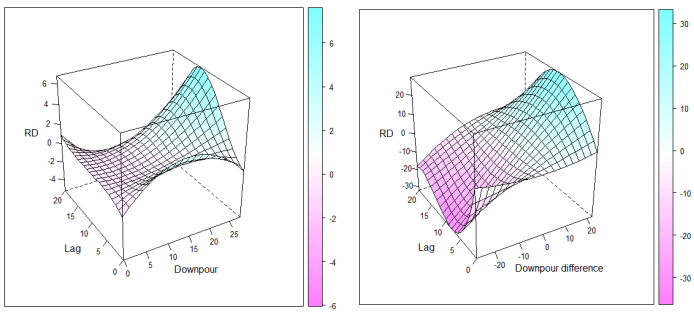
The values on the $y$-axis show the risk difference (RD) in the daily counts of incidences of ACS caused by the meteorological variables in percent points (pp). If the RD is 0, there is no effect of the given meteorological variable. If RD$>$ 0, there is a positive effect, i.e. the given meteorological variable contributes to an increase in daily counts of incidences of ACS. If RD$<$ 0, the meteorological variable has a negative effect, i.e. it contributes to a decrease in daily counts of incidences of ACS. Risk difference in pp is the change in risk measured in percent for each person, when adding a covariate to the model, e.g. if a subject has a 10% risk of developing an incidence of ACS and the RD is $-2$ pp that subject now has a risk of only 8%. The $z$-axis shows lag no. 0 to lag 20, where lag 0 is the same day, lag 1 is the day before and so on. The $x$-axis shows the value of the meteorological variable for the linear and squared effects and the difference in the meteorological variable for the difference effects. By looking at the linear and squared effects, it seems that for high values on temperature there is a small negative effect at lag 0-2 of approximately 2 pp. Small humidity values seems to have a small negative effect at lag 0-5 with a RD of approximately 2 pp. For high humidity values there is a small positive effect at lag 0-5 of 2 pp. Downpour seems to have a small negative effect in the first couple of lags in 2 pp and for for high values on downpour there is an effect around lag 10 of approximately 4 pp. Wind velocity only seems to have a negative effect for all values. Small values of atmospheric pressure seems to have an immediate effect of approximately 1 pp. For the daily difference in temperature it seems that large daily differences have a positive effect of up to 3 pp; however it also seems that a negative change could have an effect of up to 4 pp. The same is the case for humidity, where large daily changes causes a positive effect of up to 20 pp. A large daily difference in downpour seems to have an effect at lag 5-20, where a positive difference causes an positive effect of up to 3 pp and a negative difference causes a negative effect of up to 2 pp. The only effect seen in a daily difference in atmospheric pressure and wind velocity is a small positive effect at lag 0- 2 where a negative difference causes a positive effect of 5 pp and 4 pp, respectively.




(a) Risk difference (pp) caused by temperature (° C) with 0-20 days of lag.

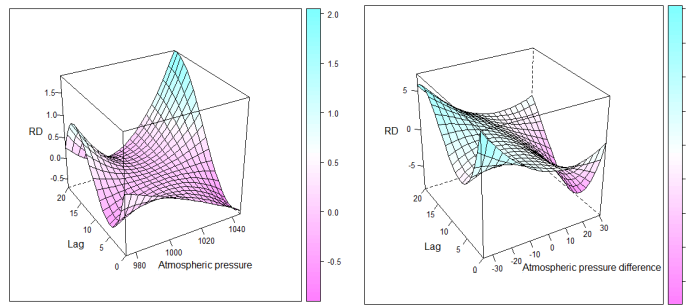



(b) Risk difference (pp) caused by humidity (%) with 0-20 days of lag.

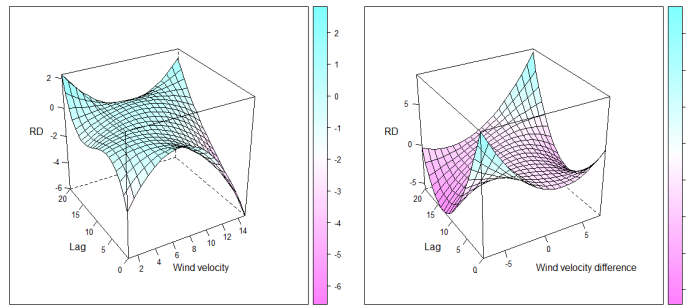



(c) Risk difference (pp) caused by downpour (mm) with 0-20 days of lag.

**Figure 3.3:** Risk difference (pp) caused by the meteorological variables for daily counts of incidences of ACS with 0-20 days of lag.




(a) Risk difference (pp) caused by atmospheric pressure (hPA) with 0-20 days of lag.




(b) Risk difference (pp) caused by wind velocity (m/s) with 0-20 days of lag.

**Figure 3.4:** Risk difference (pp) caused by the meteorological variables for daily counts of incidences of ACS with 0-20 days of lag.

The deviance differences of the models for the various meteorological variables are shown in table 3.1 along with the p-values.

The deviance differences of the models for the daily difference in various meteorological variables are shown in table 3.2 along with the p-values.

Table 3.1 shows that only temperature have a significant p-value of the deviance difference for daily counts of incidences of ACS. Table 3.2 shows that in the daily difference none of the meteorological vari-

| | ACS | |
|---|---|---|
| **Meteorological Variable** | **Deviance Difference** | **p-value** |
| **Temperature** | 25 | 1.8e-03 |
| **Atmospheric Pressure** | 4 | 9.3e-01 |
| **Humidity** | 15 | 1.4e-01 |
| **Wind Velocity** | 6 | 8.0e-01 |
| **Downpour** | 11 | 3.3e-01 |

**Table 3.1:** Deviance difference and p-values for the models for the five different meteorological values for counts of daily incidences of ACS.

| | ACS Daily Difference | |
|---|---|---|
| **Meteorological Variable** | **Deviance Difference** | **p-value** |
| **Temperature** | 3 | 9.3e-01 |
| **Atmospheric Pressure** | 4 | 8.2e-01 |
| **Humidity** | 13 | 1.2e-01 |
| **Wind Velocity** | 5 | 7.6e-01 |
| **Downpour** | 8 | 2.8e-01 |

**Table 3.2:** Deviance difference and p-values for the models for daily difference in the five different meteorological variables for counts of daily incidences of ACS.

ables have a significant p-value. Therefore only further analyses of the temperature variable was made. For the temperature, binary variables defining hot/cold days were made. Hot/cold days were defined as the upper and lower 5% of the meteorological variable. High temperatures are significant with a p-value of $1.5e-02$ and a deviance difference of 12.

## 3.2 Apoplexy

Figure 3.5 shows the trend of daily counts of APO incidences over time.

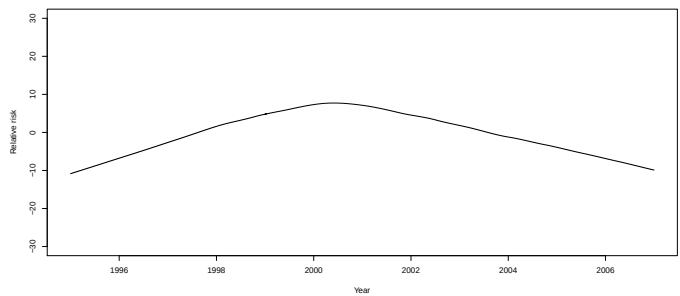



**Figure 3.5:** Trend of daily counts of APO incidences over time, with an average value of daily cases of incidences of APO of 6.82.

Figure 3.6 shows the influence of the day of the week and holidays on daily counts of incidences of APO. The day of the week has an

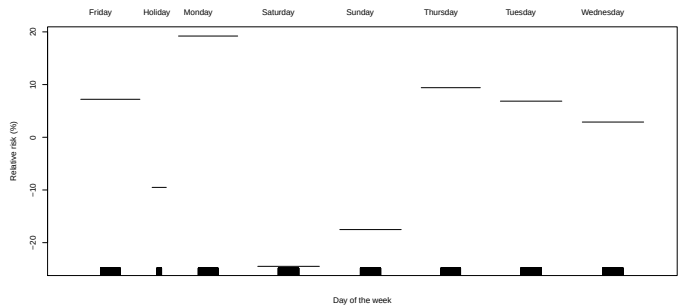
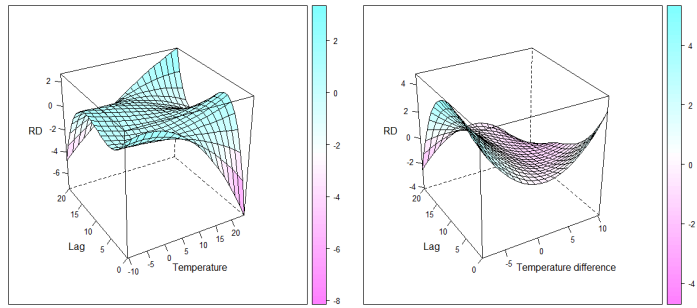



**Figure 3.6:** Influence of the day of the week on daily counts if incidences of APO.

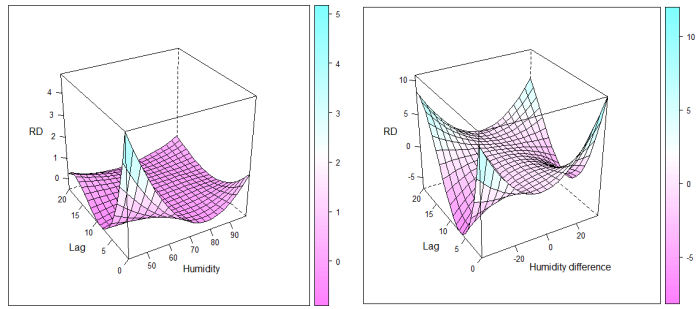
influence on the model with a p-value of $< 2.2e - 16$ and deviance difference of 556.61.

Figure 3.7 and figure 3.8 show the trends of the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ values of the PDLM for the effect over time of the five meteorological variables. In the left side of the figure the linear and squared effects are shown and in the right side the effects of the daily difference in the meteorological variables are shown.

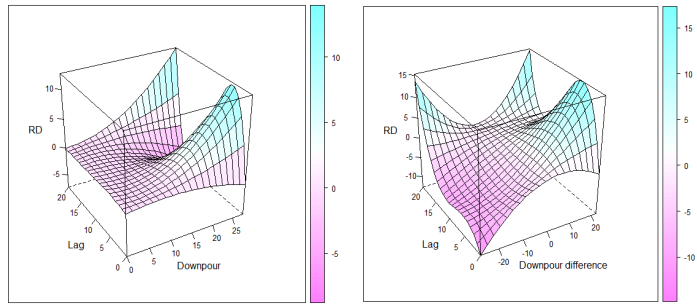
The deviance differences of the models for the various meteorological




(a) Risk difference (pp) caused by temperature (° C) with 0-20 days of lag.

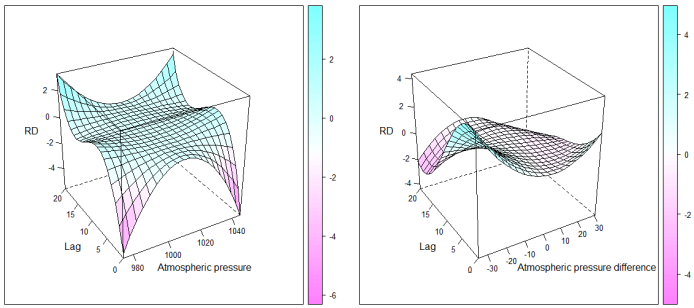



(b) Risk difference (pp) caused by humidity (%) with 0-20 days of lag.

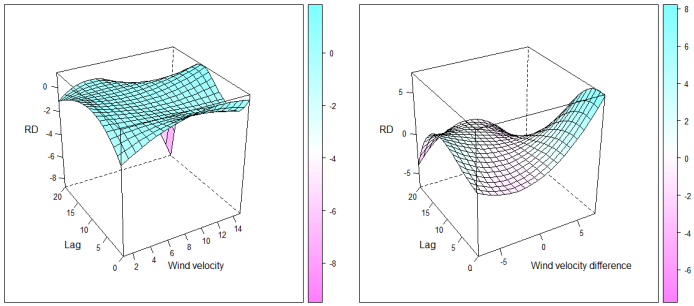



(c) Risk difference (pp) caused by downpour (mm) with 0-20 days of lag.

**Figure 3.7:** Risk difference (pp) caused by the meteorological variables for daily counts of incidences of APO with 0-20 days of lag.




(a) Risk difference (pp) caused by atmospheric pressure (hPA) with 0-20 days of lag.




(b) Risk difference (pp) caused by wind velocity (m/s) with 0-20 days of lag.

**Figure 3.8:** Risk difference (pp) caused by the meteorological variables for daily counts of incidences of APO with 0-20 days of lag.

variables are shown in table 3.3 along with the p-values.

The deviance differences of the models for the daily difference in meteorological variables are shown in table 3.4 along with the p-values.

For apoplexy none of the values in table 3.3 and table 3.4 are significant.

| | APO | |
|---|---|---|
| **Meteorological Variable** | **Deviance Difference** | **p-value** |
| **Temperature** | 14 | 8.4e-02 |
| **Atmospheric Pressure** | 6 | 6.2e-01 |
| **Humidity** | 2 | 9.0e-01 |
| **Wind Velocity** | 5 | 8.0e-01 |
| **Downpour** | 6 | 6.1e-01 |

**Table 3.3:** Deviance differences and p-values for the models for the five different meteorological variables for daily counts of incidences of APO.

| | APO Daily Difference | |
|---|---|---|
| **Meteorological Variable** | **Deviance Difference** | **p-value** |
| **Temperature** | 8 | 4.9e-01 |
| **Atmospheric Pressure** | 3 | 9.2e-01 |
| **Humidity** | 8 | 3.4e-01 |
| **Wind Velocity** | 9 | 3.7e-01 |
| **Downpour** | 8 | 5.6e-01 |

**Table 3.4:** Deviance differences and p-values for the models for daily difference in the five meteorological values for daily counts of incidences of APO.

## 3.3 Venous Thromboembolism

Figure 3.9 shows the trend of daily counts of VTE incidences over time.

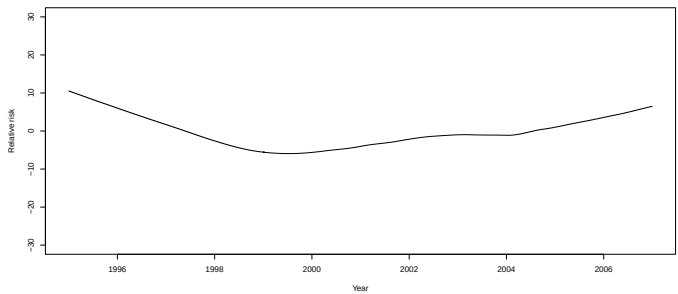



**Figure 3.9:** Trend of the daily counts of VTE incidences over time, with an average value of daily counts of incidences of VTE of 3.46.

Figure 3.10 shows the influence of the day of the week and holidays on

daily counts incidences of VTE. The day of the week has an influence

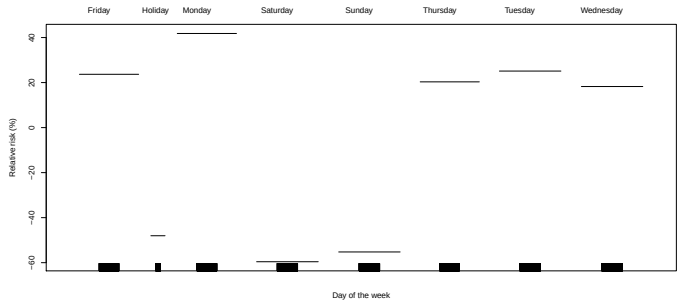



**Figure 3.10:** Influence of the day of the week on daily counts of incidences of VTE.

on the model with a p-value of $< 2.2e-16$ and deviance of 996.97.

Figure 3.11 and figure 3.12 show the trends of the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ values of the PDLM for the effect over time of the five meteorological variables. In the left side of the figure the linear and squared effects are shown and in the right side the effects of the daily difference in the meteorological variables is shown.

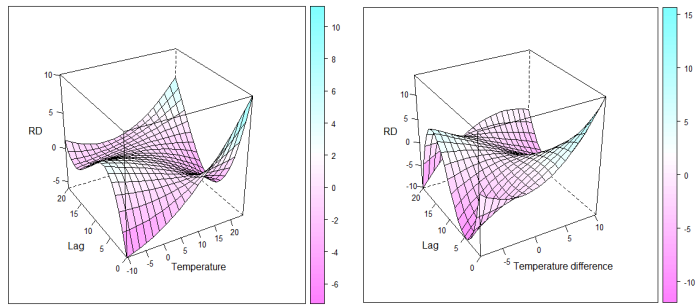
The deviance differences of the models for the various meteorological variables are shown in table 3.5.

| | **VTE** | |
|---|---|---|
| **Meteorological Variable** | **Deviance Difference** | **p-value** |
| **Temperature** | 10 | 2.8e-01 |
| **Atmospheric Pressure** | 13 | 1.0e-01 |
| **Humidity** | 8 | 4.5e-01 |
| **Wind Velocity** | 4 | 8.7e-01 |
| **Downpour** | 6 | 6.6e-01 |

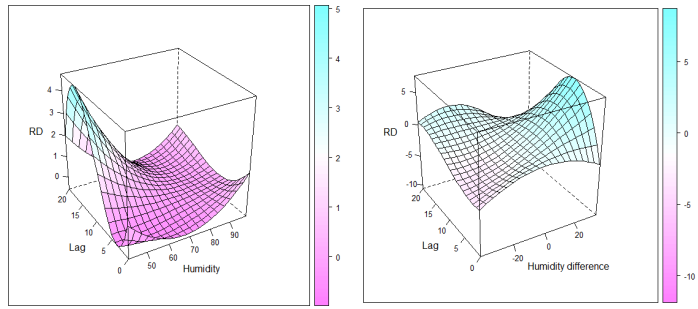
**Table 3.5:** Deviance differences and p-values for the models for the five meteorological variables for daily counts of incidences of VTE.

The deviance differences of the models for the daily difference in meteorological variables are shown in table 3.6 along with the p-values.

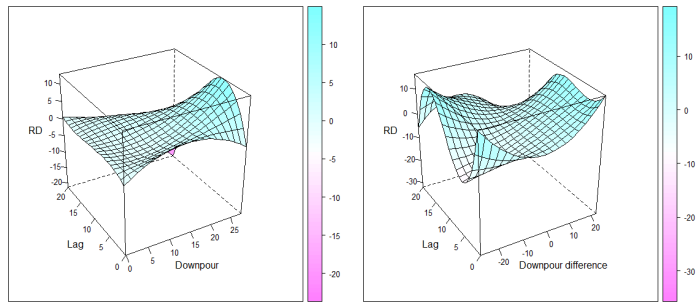
For VTE none of the values in table 3.5 and table 3.6 were significant.




(a) Risk difference (pp) caused by temperature (° C) with 0-20 days of lag.

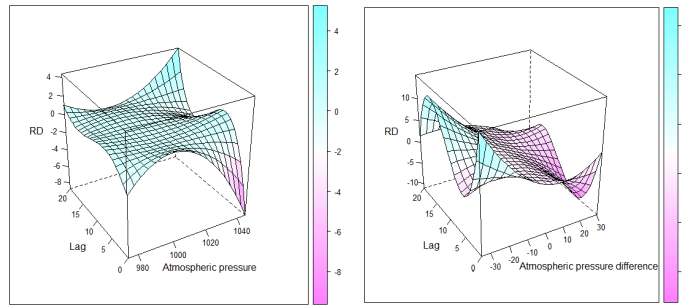



(b) Risk difference (pp) caused by humidity (%) with 0-20 days of lag.

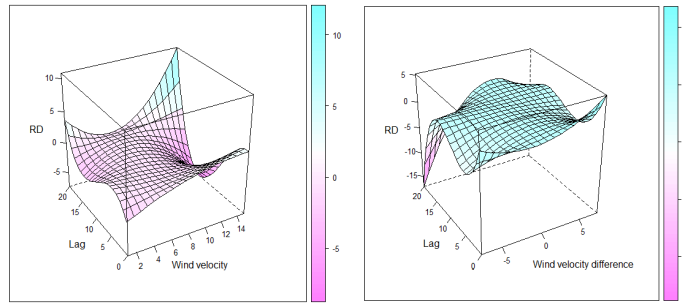



(c) Risk difference (pp) caused by downpour (mm) with 0-20 days of lag.

**Figure 3.11:** Risk difference (pp) caused by the meteorological variables for daily counts of incidences of VTE with 0-20 days of lag.




(a) Risk difference (pp) caused by atmospheric pressure (hPA) with 0-20 days of lag.




(b) Risk difference (pp) caused by wind velocity (m/s) with 0-20 days of lag.

**Figure 3.12:** Risk difference (pp) caused by the meteorological variables for daily counts of incidences of VTE with 0-20 days of lag.

## 3.4 Atrial Fibrillation

Figure 3.13 shows the trend of the daiy counts of AF incidences over time.

Figure 3.14 shows the influence of the day of the week on daily counts incidences of AF. The day of the week has an influence on the model with a p-value of $< 2.2e - 16$ and deviance of 2809.4.

Figure 3.15 and figure 3.16 show the trends of the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ values of the PDLM for the effect over time of the five meteorological variables. In the left side of the figure the linear and squared effects are shown and in the right side the effects of the daily difference in the meteorological variables is shown.

| | VTE Daily Difference | |
|---|---|---|
| **Meteorological Variable** | **Deviance Difference** | **P-value** |
| **Temperature** | 8 | 6.2e-01 |
| **Atmospheric Pressure** | 8 | 5.0e-01 |
| **Humidity** | 8 | 5.0e-01 |
| **Wind Velocity** | 8 | 8.5e-01 |
| **Downpour** | 8 | 5.1e-01 |

**Table 3.6:** Deviance differences and p-values for the models for difference in the five meteorological variables for daily counts of incidences of VTE.

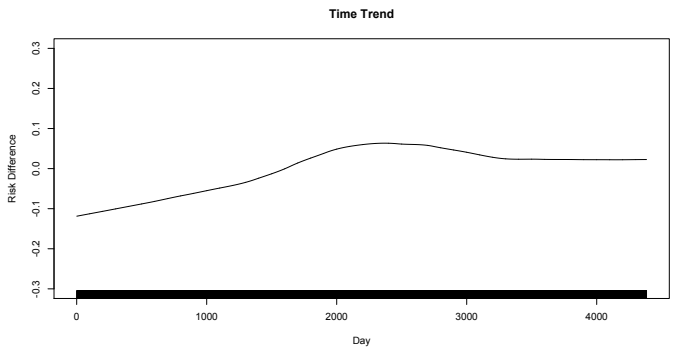



**Figure 3.13:** Trend of daily counts of AF incidences over time, with an average value of daily cases of incidences of AF of 7.00.

The deviance differences of the models for the various meteorological variables are shown in table 3.7.

The deviance differences of the models for the daily difference in the meteorological variables are shown in table 3.8 along with the p-values.

Temperature is only meteorological variables, that has significant p-values in the deviance difference for daily counts of incidences of AF, therefore further analysis of this variable were made by defining hot/cold days. Only high temperatures have a significant effect with a p-value of $1.5e-02$ and a deviance difference of 12.

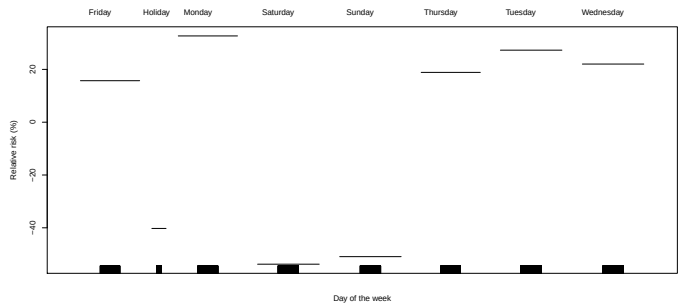


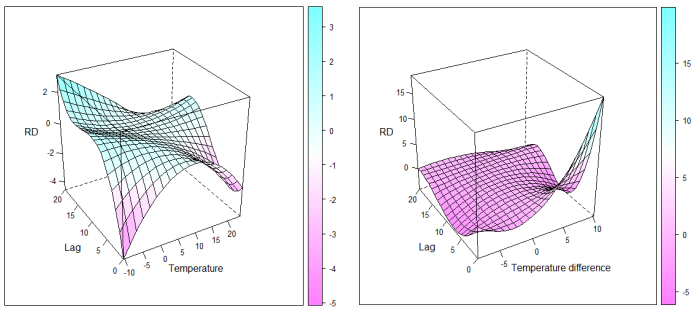

**Figure 3.14:** Influence of the day of the week on daily counts of incidences of AF.

| | **AF** | |
|---|---|---|
| **Meteorological Variable** | **Deviance Difference** | **p-value** |
| **Temperature** | 26 | 1.2e-03 |
| **Atmospheric Pressure** | 10 | 2.9e-01 |
| **Humidity** | 11 | 2.1e-01 |
| **Wind Velocity** | 11 | 2.0e-01 |
| **Downpour** | 5 | 7.9e-01 |

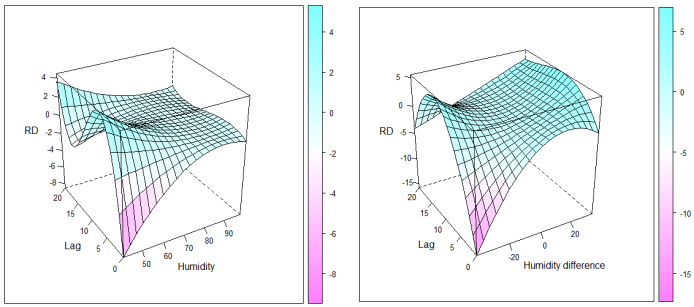
**Table 3.7:** Deviance differences and p-values for the models for the five meteorological variables for daily counts of incidences of AF.

| | **AF Daily Difference** | |
|---|---|---|
| **Meteorological Variable** | **Deviance Difference** | **p-value** |
| **Temperature** | 10 | 2.7e-01 |
| **Atmospheric Pressure** | 10 | 2.6e-01 |
| **Humidity** | 14 | 8.9e-02 |
| **Wind Velocity** | 8 | 6.2e-01 |
| **Downpour** | 5 | 7.5e-01 |

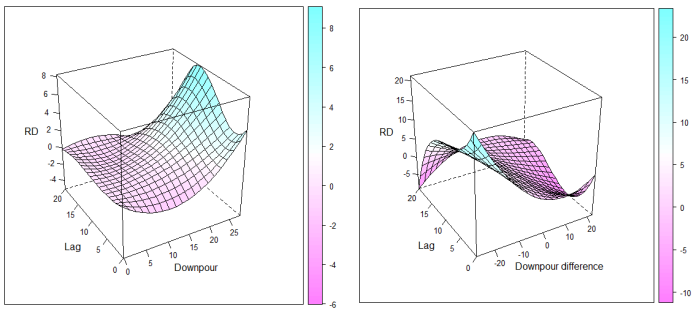
**Table 3.8:** Deviance differences and p-values for the models for the daily difference in the five meteorological variables for daily counts of incidences of AF.




(a) Risk difference (pp) caused by temperature (° C) with 0-20 days of lag.

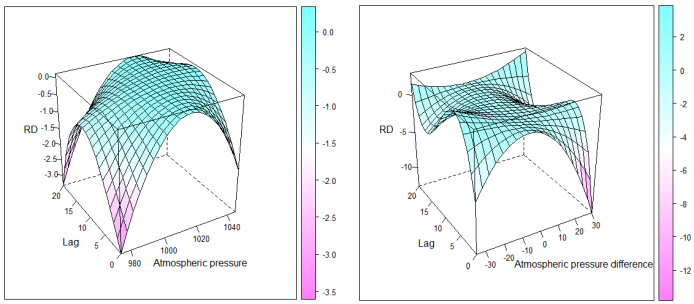



(b) Risk difference (pp) caused by humidity (%) with 0-20 days of lag.

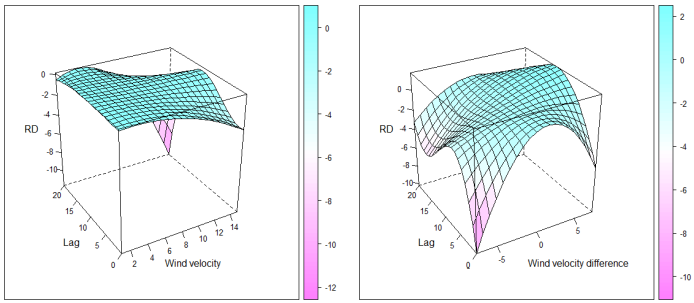



(c) Risk difference (pp) caused by downpour (mm) with 0-20 days of lag.

**Figure 3.15:** Risk difference (pp) caused by the meteorological variables for daily counts of incidences of AF with 0-20 days of lag.




(a) Risk difference (pp) caused by atmospheric pressure (hPA) with 0-20 days of lag.




(b) Risk difference (pp) caused by wind velocity (m/s) with 0-20 days of lag.

**Figure 3.16:** Risk difference (pp) caused by the meteorological variables for daily counts of incidences of AF with 0-20 days of lag.

# Dynamic Linear Models

In this chapter results from the analysis with DLMs are presented. The mathematical theory behind DLM is outlined in chapter 10 and the analysis strategy in section 2.1.2. In the GAM analysis temperature showed the most effect on the daily counts of incidences of CVDs; therefore the DLM analysis is only conducted for this variable.

The daily counts of incidences of ACS, APO, VTE and AF all follow Poisson distributions and are positively skewed data. In order to apply DLM, it is necessary that the observations are normally distributed. Therefore, the data is transformed by raising the daily count vectors to the power of $\frac{2}{3}$, since this transformation makes the distribution as symmetrical as possible. Hereafter, the response is approximately normally distributed. Histograms and QQ-plots can be seen on `http://homes.student.aau.dk/cbisga07/`.

As in the GAM analysis, the DLM included a seasonal component, $\sin\left(\frac{2\pi t}{365.25}\right) + \cos\left(\frac{2\pi t}{365.25}\right)$, where $t$ is our time variable. The seasonal component entered the model as a covariate to capture the long-term seasonal variation in the data. The model also included a factor indicating the day of the week along with both a linear and quadratic term of the temperature variable.

As in the GAM results, the results from ACS are outlined and the results for APO, VTE and AF are given without comments. Results when stratifying on gender are available in appendix C and `http://homes.student.aau.dk/cbisga07/`.

## 4.1 Acute Coronary Syndrome

Figure 4.1 shows the effect of the day of the week on the fitted values. As in the GAM results it is clear that Monday and Tuesday are the days of the week with the most daily cases. As mentioned in GAM, this tendency is possibly caused by some people putting off seeking out medical attention during the weekend until they can see their personal general practitioner on weekdays.

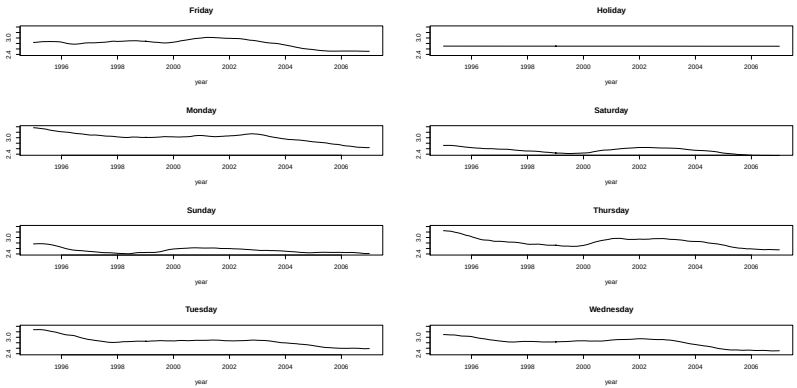



**Figure 4.1:** The effect of the day of the week on daily counts of ACS.

Figure 4.2 shows a plot the effect of both a linear and quadratic effect of temperature of the day in question on the fitted values of the daily counts of ACS. There is a consistent tendency of a negative effect on the counts during summertime of up to approximately 0.15 persons per day compared to wintertime. It tells us that high temperatures have some protective effect on ACS compared to low temperatures.

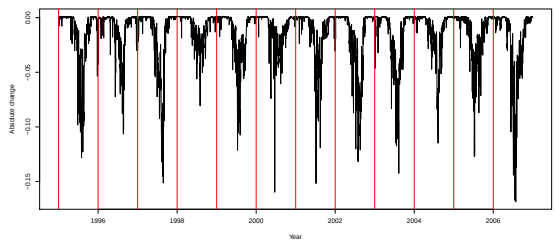



**Figure 4.2:** Effect of linear and quadratic components of temperature on daily counts of ACS. Turn of the year is indicated by vertical lines.

Finally the model was fitted using linear and quadratic components of temperature for 0-7 days of lag prior to the event of ACS, as described in 2.1.2. This yielded the results shown in figure 4.3.

The tendency of fewer daily cases during summer seem quite consistent in this plot. There are up to about 0.3 fewer daily cases of ACS per day in the summer than in the winter, again indicating the protective effect of high temperatures on ACS.
The individual effects of each of the lags are plotted in figure 4.4. The lags all show consistently fewer daily cases in the summer. It

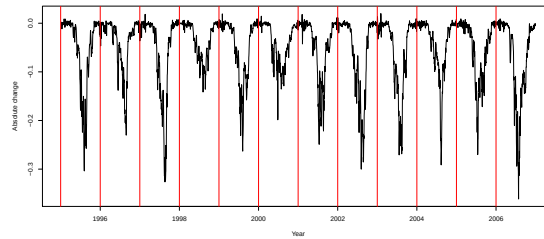



**Figure 4.3:** Effect of linear and quadratic components of temperature with 0-7 days of lag on daily counts of ACS. Turn of the year is indicated by vertical lines.

appears that the temperature on lag 7 has the largest negative effect on daily cases, but the difference between the lags are small.

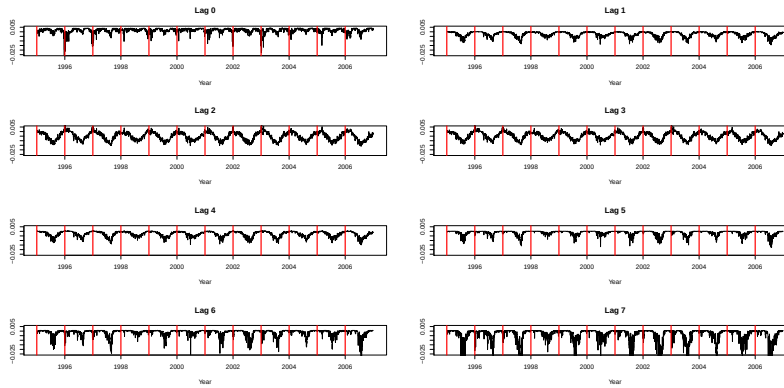



**Figure 4.4:** Individual lag effect of linear and quadratic components of temperature on daily counts of ACS. Turn of the year is indicated by vertical lines.

## 4.2 Apoplexy

Figure 4.5 illustrates the influence of the day of the week on the fitted values.

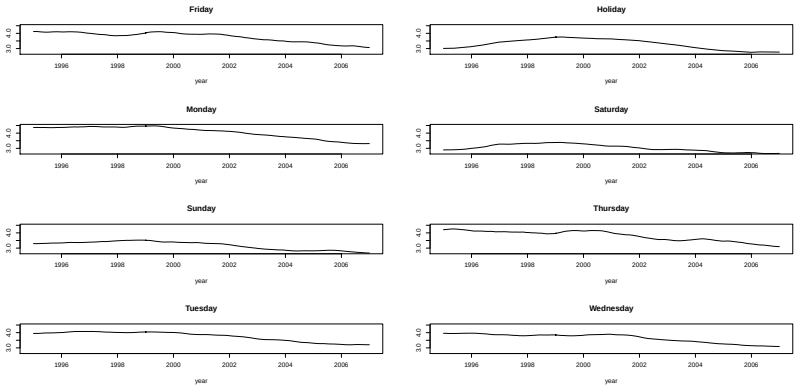



**Figure 4.5:** The effect of the day of the week on daily counts of incidences of APO.

Figure 4.6 illustrates the effect of both a linear and quadratic effect of temperature of the day in question on the fitted values of the daily counts of incidences of APO.

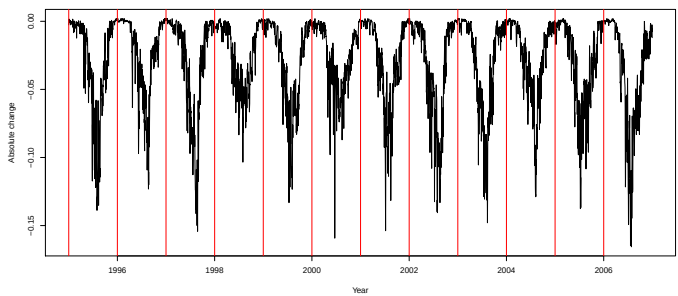



**Figure 4.6:** Effect of linear and quadratic components of temperature on daily counts of incidences of APO. Turn of the year is indicated by vertical lines.

The results from the analysis using 0-7 days of lag are shown in figure 4.7.

The individual effects of each of the lags are plotted in figure 4.8.

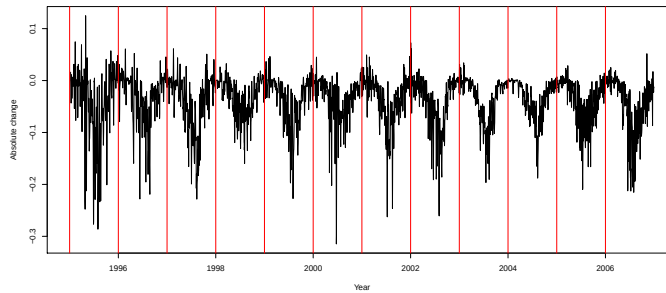



**Figure 4.7:** Effect of linear and quadratic components of temperature with 0-7 days of lag on daily counts of incidences APO. Turn of the year is indicated by vertical lines.

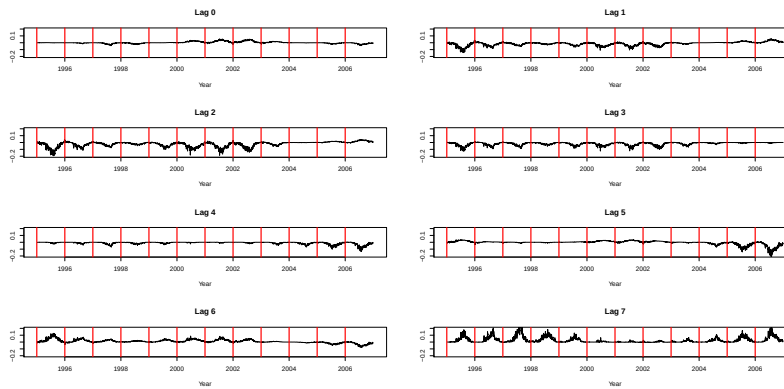



**Figure 4.8:** Individual lag effect of linear and quadratic components of temperature on daily counts of incidences APO. Turn of the year is indicated by vertical lines.

## 4.3 Venous Thromboembolism

Figure 4.9 illustrates the influence of the day of the week on the fitted values.

Figure 4.10 illustrates the effect of both a linear and quadratic effect of temperature of the day in question on the fitted values of the daily counts of incidences of VTE.

The results from the analysis using 0-7 days of lag are shown in figure 4.11.

The individual effects of each of the lags are plotted in figure 4.12.

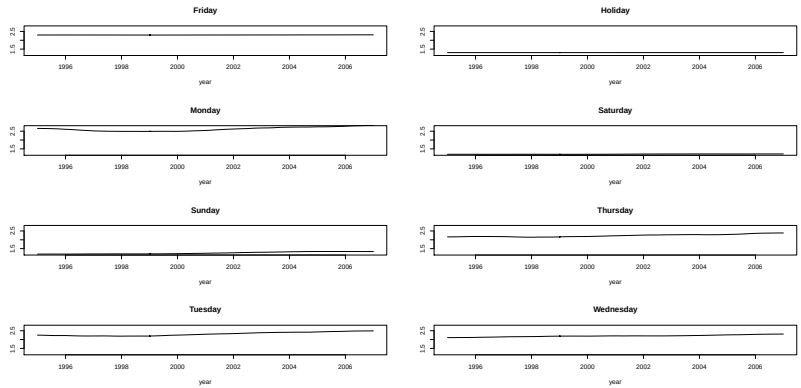



**Figure 4.9:** The effect of the day of the week on daily counts of VTE.

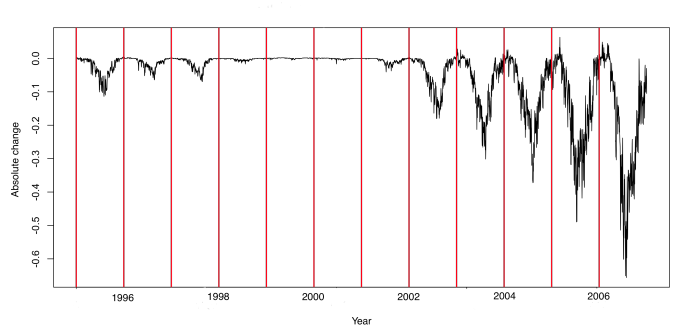



**Figure 4.10:** Effect of linear and quadratic components of temperature on daily counts of incidences of VTE. Turn of the year is indicated by vertical lines.

## 4.4 Atrial Fibrillation

Figure 4.13 illustrates the influence of the day of the week on the fitted values.

Figure 4.14 illustrates the effect of both a linear and quadratic effect of temperature of the day in question on the fitted values of the daily counts of incidences of AF.

The results from the analysis using 0-7 days of lag are shown in figure 4.15.

The individual effects of each of the lags are plotted in figure 4.16.

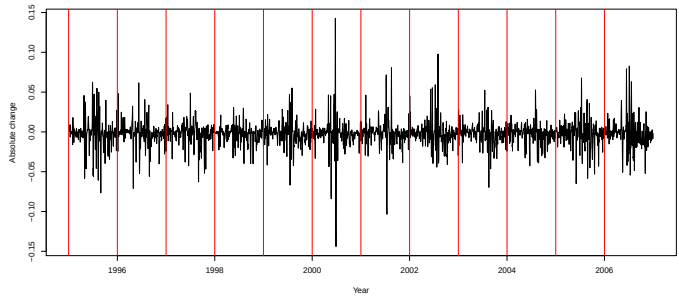



**Figure 4.11:** Effect of linear and quadratic components of temperature with 0-7 days of lag on daily counts of incidences of VTE. Turn of the year is indicated by vertical lines.

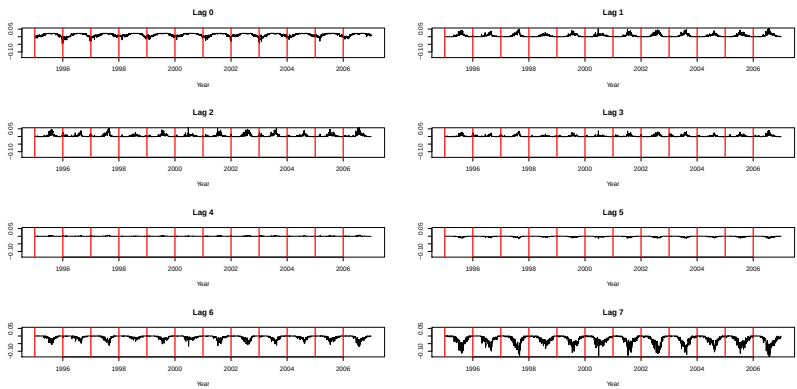



**Figure 4.12:** Individual lag effect of linear and quadratic components of temperature on daily counts of incidences of VTE. Turn of the year is indicated by vertical lines.

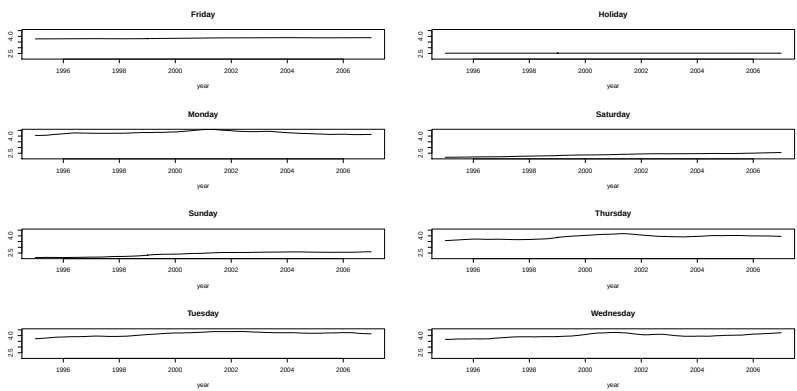



**Figure 4.13:** The effect of the day of the week on daily counts of incidences of AF.

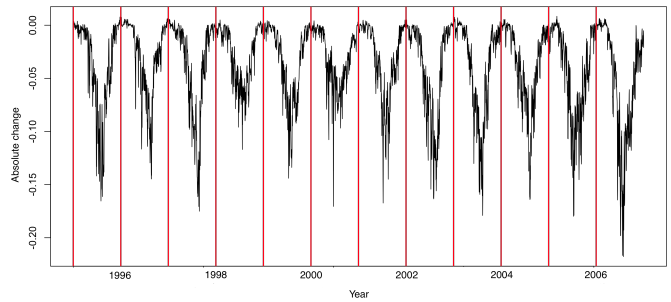



**Figure 4.14:** Effect of linear and quadratic components of temperature on daily counts of incidences of AF. Turn of the year is indicated by vertical lines.

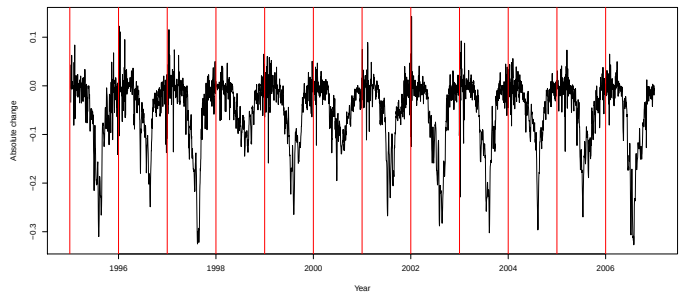



**Figure 4.15:** Effect of linear and quadratic components of temperature with 0-7 days of lag on daily counts of incidences of AF. Turn of the year is indicated by vertical lines.

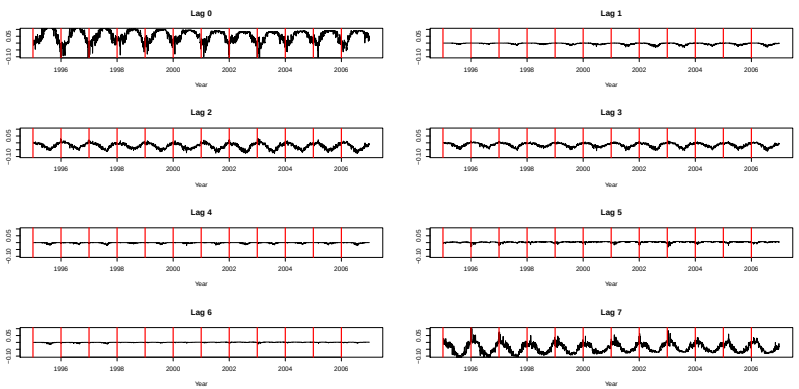



**Figure 4.16:** Individual lag effect of linear and quadratic components of temperature on daily counts of incidences of AF. Turn of the year is indicated by vertical lines.

# Synthesis

## 5.1 Conclusion

In this Master of Science Thesis we have investigated a number of meteorological variables and their influence on the daily incidences of CVDs in Denmark along with their lagged effect. First an analysis using GAMs was conducted. GAMs were chosen since they allow regressions to include non-parametric smooth functions to model a potential nonlinear dependence of the log-frequency of daily incidences of CVDs of interest on the meteorological variables. This model consists of an offset variable of daily counts of the number of residents in Denmark at time $t$, a LOESS smooth function of time to capture the long-term time-trend in data and a factor indicating the day of the week included as a covariate. To capture the long-term seasonal variation in the data the model also included a seasonal component as a covariate. To capture the lagged effect of the meteorological variables a PDLM was included in the model, with the motivation that the weather today can have an influence on incidences of CVDs not only occurring today, but also on several subsequent days. For the DLM analysis the daily counts of incidences was divided by the relative change in the number of residents in Denmark with respect to the mean population at the given time and transformed to normally distributed data by raising the relative daily counts to the power of 2/3. Like the GAMs, the DLMs also included a factor indicating the day of the week, a seasonal component and a PDLM. Both models included both a linear and a quadratic term of the meteorological variable. For both models parallel analyses were conducted for each of ACS, APO, VTE and AF.

### GAM

The GAM analysis gave significant results for temperature, that showed some effect on all diagnoses, but it changes from diagnosis to diagnosis when the effect is highest during the year. Analysis made on the daily difference in the weather did not give any significant results. For all GAM analyses the day of the week had a significant

effect on incidences of CVDs. All CVDs showed a large increase on Mondays and a large decrease in the weekends and a small decrease on holidays.

## DLM

The analysis using DLMs did not give any significant effects of the temperature variable. However it gives a more realistic and consistent image of how the temperature influence the daily counts of incidences of CVDs today and with a lag of up to 7 days. On all diagnoses the analyses show a decrease in the counts of daily incidences caused by high temperatures in the summer. As in the GAM analyses the day of the week had an significant effect on incidents of CVDs. For DLM the data was separated to analyze the age groups 20-49 and 50+ and each gender individually. However most of the subjects in the data were age 50+ and because of this we could not normalize data for the age group 20-49, since there were too few subjects per day in this group. Making the analysis for the age group 50+ made no change in the results, since most of subjects in data already was in this group. Analyzing males and females separately showed the same tendencies as when not separated, i.e. the temperature has the same influence on daily counts of incidences on both males and females.

## 5.2 Discussion

The first issue related to the data in this thesis was to decide whether we would look at all the diagnoses separately or together. Ongoing studies are trying to show that ACS, APO and VTE are basically the same disease, which is interesting, since the approach now is to look at each diagnosis individually. A problem in this is however that the cases of ACS and APO completely outweigh the VTE cases (20,565 and 29,822, respectively against 14,045 subjects) so when looking at CVDs in relation to meteorology it would make more sense to look at them separately. Pathologically AF is different from the other diseases, and therefore it was always meant for this to be studied alone.
Another discussion in relation to the data was which diagnosis codes to use. Since all hospitalizations in Denmark are registered in the LPR from hospitals and wards nationally, and since the registration of diagnosis codes is a subjective consideration, they are not always correct. Validation studies were made on each of the diseases and from these studies we decided which diagnosis codes, diagnosis types and which patient types to include in the study. The sample size for the APO validation is small and therefore the results of this could be rather imprecise. A larger validation study of APO is ongoing, but results of the study was not available at the deadline of this thesis. A brief discussion of this is given in chapter 1.3.

The original data set consists of incidences of CVDs, from subjects of age 20 and above, from all of Denmark from 1980 up until 2011 and was a large data set. The reason the subjects are age 20+ is because we are interested in subjects with arteriosclerosis. If subjects get a CVD before they turn 20 it is almost always because of a congenital disease. When collecting data we also chose the criterion, that the subjects only should be included with the first occurrence of one of all four diagnoses. We could have chosen, that subjects could enter in all four data sets if they had all diagnosis, but we do not know if one disease could cause one of the other diseases. In this case the occurrence of the disease could be because of the old disease and not because of the weather. The ICD diagnosis codes changed from ICD8 to ICD10 in Denmark in 1994 and here some definitions were changed, so to make sure all definitions were the same we chose to narrow the data down to only consist of incidences of CVDs from 1995 and above. We

decided to narrow the data further down by only looking at the 10 biggest municipalities, since the subject from this population probably will give us the same results, as including all incidents. Because of the municipal merger in 2007 we only used data from incidents up to this year giving us 12 years with incident CVDs in the 10 biggest municipalities in Denmark. From DMI we got the meteorological variables from these municipalities and originally we wanted to match each subject to the meteorological variable from the municipality they stayed in for the last three weeks; this idea did however not work in practice, since a lot of the subjects municipality codes was not consistent with the municipality codes from the CPR. The reason for this is that the subjects not necessarily live in the same municipality as the hospital they are admitted to. Ideally we would had assigned each subject the meteorological variables that corresponded to the municipality they have stayed in for the last three weeks, but since this is not possible we decided to take the mean of the meteorological variables for the 10 municipalities. Since Denmark is a relatively small country and the weather therefore does not change notably depending on location, we decided the mean of the variables were suitable.

Our analysis took its point of reference in a study of Braga et al.[10], who conducted a time-series analysis estimating both the acute and lagged effect of weather on respiratory and cardiovascular deaths in 12 US cities. Braga et al. used additive Poisson regressions for each city, with a smooth function of time to capture the long-term time-trend. To capture the lagged effect they used a PDLM. When looking at the time-trend for all diagnoses, ACS and APO incidences decrease from the year 2000, whereas VTE incidences increase. This could indicate, that VTE has replaced ACS and APO in some cases.

When doing the analysis we expected to find that the weather, and especially the temperature, had high influence on the incidences of CVDs when low and high temperatures occurred, but the results using this method was equivocal on this point for the four diagnoses. Actually, for temperature, APO was the only diagnosis, which gave the outcome we expected. The results are shown in chapter 3 and we would say that conclusions should be drawn with care, meaning that in the model we included both a linear and quadratic term and in the PDLM we fitted to a polynomial of degree three, this way we

force the effect of the meteorological variables on the incidence CVDs to look a certain way and it is not clear if the effect shown on the figures are actually real or we created them ourselves, e.g. the effects showing on lag 20 are rather high and are probably only presence because we force it to fit a polynomial of third degree. Ideally we would see the effect starting high and decreasing the further away from lag 0 we get, but this was not possible to do with the chosen model. An obvious idea was to not only look at the meteorological variables, but also look at the difference from day to day to see if it was the change in the weather that caused more or fewer incidences of CVDs. These results was as equivocal as the previous results and showed that there has to be quite a large change in the weather for it to have an effect. None of the variables showed significant effect, so we chose not to include this analysis when using DLM.

DLM allows the parameters to change over time and this is perhaps why the results from the analysis using these models give a more consistent image of the impact of temperature on daily counts of CVDs. Unlike GAM, high temperature seems to have a negative effect in the daily incidences of CVDs for all diagnoses. In the DLM analysis the eight levels of the day of the week have separate variances. The analysis was also made by letting the variances of the day of the week levels be the same, which seems to be a fair assumption looking at the time-trend for each day of the week during the years. When doing this the cycles turned odd; therefore we chose to give the levels separate variances. It is worrying that the cycle can vary this much by changing such a small thing in the model and it is perhaps an indication that there is not a lot of information in data. Therefore conclusion of the results should be drawn with care.

When looking at age groups, the subjects aged 50+ more often develop incidences of CVDs in comparison with subjects aged 20-49, this is due to the fact that arteriosclerosis biologically is something that comes with age. When analyzing on each gender separately the temperature did not show different influence according to gender and we do not see a reason why this should not be correct.

Originally the lag time was chosen to be 20 days, since the study og Braga et al. used this amount of lag. We do not however

find it probable that the temperature has an influence on the daily counts of CVDs for such a long period; therefore we chose the lag time to be a week when conducting the DLM analyses. Optimally one could test all lag times to see when there is the largest effect on the daily counts of incidences of CVDs.

For all analyses the variable showing the most interesting pattern was the factor indicating the day of the week/holiday. All days had an significant effect on incidences of CVDs and for all CVDs it showed a large increase on Monday and a large decrease in the weekends and small decrease on holidays. The explanation for this could be that people have a tendency not to go to the hospital and/or call a general practitioner in the weekends and holidays, they wait until Monday. Therefore you do not have increased risk of getting a incidence of CVD on Monday, the results simply reflect the general behavior of people regarding the hospital services in Denmark. Another explanation could also be that people relax in their holidays and weekends and the beginning of a new week can be stressful. This may lead to an incidence of CVD.

Earlier studies have shown that the daily incidence rate for CVDs in Denmark was highest during the winter [5][6] and it was therefore obvious to think that the weather had a say in this. According to our results the weather do not have a lot of influence on the incidences of CVDs, but in clarifying the etiology of CVDs, it it worth investigating further and future work could encompass the influence of the temperature on the reappearance of CVDs in subjects. Study in this field is still essential, because it contributes to clarifying the etiology of CVDs, and it could therefore improve treatment and preventive strategies. Future work in this field could include analyses of the association between incidences of CVDs and air-pollution. It is well known, that meteorological conditions and air pollution are closely related, [14], and low temperatures prevent air pollutants from dispersing. This could maybe be an explanation of why the incidence rate og CVDs are highest during the winter in Denmark. Other things that could be interesting to investigate, that we did not include in this study is the hours of sun and change in lifestyle. When receiving a lot of sun we also receive a lot of vitamin D and maybe the lack of vitamin D in the winter period in Denmark could cause an increase in the daily incidence rate for CVDs. The season can also cause people to change their lifestyle, e.g. change in

the diet and amount of exercise, which could also very well cause a increase in the daily incidence rate for CVDs during the winter in Denmark.

As mentioned before, subjects are only allowed in one of the four data sets since we do not know if one of the diseases could cause one of the other diseases. Another issue in the same category is to adjust for other possible comorbid diseases, such as diabetes and cancer since these diseases could possibly increase the risk of getting CVDs and therefore it would not be the weather causing this. This is not adjusted for in our study.

## 5.3 Perspectives

To firmly establish the results presented in this thesis, further analysis needs to be conducted. Conclusion of the results should be drawn with care, but assuming they are reliable the hospital services in Denmark should definitely take the temperature and therefore season of the year into consideration when evaluation risk factors and people at risk.
Optimally this means that everyone with high risk of developing a CVD should travel to warmer climates during the winter season in Denmark or alternatively stay indoor with higher temperatures. More realistically people at risk should look into their lifestyle, exercise habits and other risk factors during cold weather and perhaps make a diet/exercise plan that could counteract the damaging effects of the cold weather.

# Part III

# Theory

# Generalized Linear Models

In this chapter the basic theory of generalized linear models is covered including the theory behind exponential families. Also a method for estimating parameters is described.

A statistical model includes parameters that describe the patterns of the data. A simple model is a linear model connecting the two quantities $y$ and $x$ through a parameter pair $(\alpha, \beta)$. The linear model has the form

$$y = \alpha + \beta x,$$

so if the values of $\alpha$ and $\beta$ are known, the values of $y$ can be reconstructed for a given $x$. However, in practice the relationship between $y$ and $x$ is only approximately linear and therefore we choose values of $\alpha$ and $\beta$, say $a$ and $b$, that are suitable to describe the approximately linear relationship. The quantities $a + bx_1, \ldots, a + bx_n$, denoted by $\hat{y}_1, \ldots, \hat{y}_n$, are the fitted values and are generated by the model and the data. In fitting a linear relationship we need to choose the pair of parameter values ($a$,$b$) that makes the $\hat{y}$'s as close to the observed data as possible. [33, p. 4].

One might think a good model is a model, that fits the observed data well, but in making a model that fits data perfectly there is no reduction in complexity. Therefore simplicity is desired in any model, so parameters that are not needed should not be in the model. A substantially simple model gives better predictions than a more complex model. Also if a model fits very closely to a particular data set it will probably not fit very well to another data set related to the same phenomenon. [33, pp. 7-8].

## 6.1 The Exponential Family and Generalized Linear Models

Linear models on the form

$$\mathbb{E}[Y_i] = \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{6.1}$$

are the basis of analyses of continuous data, with Gaussian errors. In the model, $y_i$, $i = 1, \ldots, n$ are independent random variables, where each $y_i \sim \mathrm{N}(\mu_i, \sigma^2)$, with mean $\mu_i$ and variance $\sigma^2$ and $\mathbf{x}_i$ is the $i$'th row of a $n \times p$ design matrix $X$ containing the covariates. Associated with each covariate is an usually unknown parameter $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$, that needs to be estimated from the data. The reason to model the dependence of $y_i$ on $\mathbf{x}_i$ is to learn more about the process that produces $y_i$ to asses the relative contributions in explaining $y_i$ and to predict $y_i$ for some $\mathbf{x}_i$. [34, p.45] [33, p.9]. In our case $y_i$ describes the daily counts of incidences of CVDs and $\mathbf{x}_i$ consists of seasonality, day of the week/holiday and meteorological variables for day $i$.

By using linear models there are some clear limitations e.g. the response variables have a normal distribution and the relationship between the response and covariates have to be on the simple linear form in equation (6.1). A class of models that allows more possible situations is called generalized linear models (GLMs). GLMs are a class of models that can be used if the response variables $y_i$ have distributions other than the normal distribution. Furthermore the relationship between the response and explanatory variables does not need to be of the simple linear form in equation (6.1). Finally, the GLMs have the advantage of allowing a common approach for a set of relevant specific models. In GLM these appear as special instances of the same approach. [35, p. 224] [34, p.45].

### The Exponential Family

GLMs are defined in terms of the exponential family. Let $\mathbf{Y}$ be a random variable, which probability distribution depends on a single parameter $\boldsymbol{\theta}$. The distribution is part of the exponential family, if the density function can be written on the form

$$f(\mathbf{y}|\boldsymbol{\theta}) = s(\mathbf{y})t(\boldsymbol{\theta})e^{a(\mathbf{y})b(\boldsymbol{\theta})}, \tag{6.2}$$

where $a$, $b$, $s$ and $t$ are known functions. Equation (6.2) can be rewritten as

$$f(\mathbf{y}|\boldsymbol{\theta}) = \exp\left(a(\mathbf{y})b(\boldsymbol{\theta}) + c(\boldsymbol{\theta}) + d(\mathbf{y})\right), \tag{6.3}$$

where $s(\mathbf{y}) = \exp(d(\mathbf{y}))$ and $t(\boldsymbol{\theta}) = \exp(c(\boldsymbol{\theta}))$.

The distribution is said to be on canonical form if $a(\mathbf{y}) = \mathbf{y}$. This is the standard form. The function $b(\boldsymbol{\theta})$ is called the natural parameter of the distribution. If there are other parameters than the one of interest, $\boldsymbol{\theta}$, they are regarded as nuisance parameters forming parts of the functions $a$, $b$, $c$ and $d$. Nuisance parameters are treated as if they are known.

Many well-known distributions belong to the exponential family, e.g. the normal, Poisson and binomial distributions. [34, p. 46].

### Some Properties of the Exponential Family

Here some of the most important properties of the exponential family are stated.

*Expected Value*

From the definition of a probability density function the area under the curve is

$$\int f(\mathbf{y}|\boldsymbol{\theta})d\mathbf{y} = 1. \tag{6.4}$$

If this equation is differentiated on both sides with respect to $\boldsymbol{\theta}$,

$$\frac{d}{d\boldsymbol{\theta}} \int f(\mathbf{y}|\boldsymbol{\theta})d\mathbf{y} = \frac{d}{d\boldsymbol{\theta}} 1 = 0. \tag{6.5}$$

If the differential and integral are interchanged

$$\int \frac{df(\mathbf{y}|\boldsymbol{\theta})}{d\boldsymbol{\theta}} d\mathbf{y} = 0. \tag{6.6}$$

If equation (6.5) is differentiated twice with respect to $\boldsymbol{\theta}$ and the integral and differential are interchanged, the equation is

$$\int \frac{d^2 f(\mathbf{y}|\boldsymbol{\theta})}{d\boldsymbol{\theta}^2} d\mathbf{y} = 0. \tag{6.7}$$

From equation (6.3) the distribution of an exponential family is

$$f(\mathbf{y}|\boldsymbol{\theta}) = \exp\left(a(\mathbf{y})b(\boldsymbol{\theta}) + c(\boldsymbol{\theta}) + d(\mathbf{y})\right),$$

so

$$\frac{df(\mathbf{y}|\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \left(a(\mathbf{y})b'(\boldsymbol{\theta}) + c'(\boldsymbol{\theta})\right) f(\mathbf{y}|\boldsymbol{\theta}).$$

From equation (6.6)

$$\int (a(\mathbf{y})b'(\boldsymbol{\theta}) + c'(\boldsymbol{\theta}))f(\mathbf{y}|\boldsymbol{\theta}) = 0,$$

since $\int a(\mathbf{y})f(\mathbf{y}|\boldsymbol{\theta})d\mathbf{y} = \mathbb{E}[a(\mathbf{Y})]$ from the definition of the expected value and since $\int c'(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})d\mathbf{y} = c'(\boldsymbol{\theta})$ and because of the area under the curve is one according to equation (6.4) we get

$$\begin{aligned} & b'(\boldsymbol{\theta})\mathbb{E}[a(\mathbf{Y})] + c'(\boldsymbol{\theta}) = 0 \\ & \Updownarrow \\ & \mathbb{E}[a(\mathbf{Y})] = -\frac{c'(\boldsymbol{\theta})}{b'(\boldsymbol{\theta})}. \end{aligned} \tag{6.8}$$

*Variance*
In the same way we obtain the variance. By use of simple differential rules

$$\begin{aligned} \frac{d^2 f(\mathbf{y}|\boldsymbol{\theta})}{d\boldsymbol{\theta}^2} &= \left(a(\mathbf{y})b''(\boldsymbol{\theta}) + c''(\boldsymbol{\theta})\right) f(\mathbf{y}|\boldsymbol{\theta}) \\ &+ \left(a(\mathbf{y})b'(\boldsymbol{\theta}) + c'(\boldsymbol{\theta})\right)^2 f(\mathbf{y}|\boldsymbol{\theta}). \end{aligned} \tag{6.9}$$

The second term on the right hand side of equation (6.9) can be rewritten as

$$\begin{aligned} \left(a(\mathbf{y})b'(\boldsymbol{\theta}) + c'(\boldsymbol{\theta})^2\right) f(\mathbf{y}|\boldsymbol{\theta}) &= b'(\boldsymbol{\theta})^2 \cdot \left(a(\mathbf{y}) + \frac{c'(\boldsymbol{\theta})}{b'(\boldsymbol{\theta})}\right)^2 f(\mathbf{y}|\boldsymbol{\theta}) \\ &= b'(\boldsymbol{\theta})^2 \left(a(\mathbf{y}) - \mathbb{E}[a(\mathbf{Y})]\right) f(\mathbf{y}|\boldsymbol{\theta}), \end{aligned}$$

so the right hand side of equation (6.9) is

$$\left(a(\mathbf{y})b''(\boldsymbol{\theta}) + c''(\boldsymbol{\theta})\right) f(\mathbf{y}|\boldsymbol{\theta}) + b'(\boldsymbol{\theta})^2 \left(a(\mathbf{y}) - \mathbb{E}[a(\mathbf{Y})]\right) f(\mathbf{y}|\boldsymbol{\theta}).$$

By the use of equation (6.7) and since per definition

$$\int \left(a(\mathbf{y}) - \mathbb{E}[a(\mathbf{Y})]\right)^2 f(\mathbf{y}|\boldsymbol{\theta})d\mathbf{y} = \mathrm{Var}[a(\mathbf{Y})]$$

and because of the calculations already shown in finding the expected value,

$$\int \frac{d^2 f(\mathbf{y}|\boldsymbol{\theta})}{d\boldsymbol{\theta}^2} d\mathbf{y} = b''(\boldsymbol{\theta})\mathbb{E}[a(\mathbf{Y})] + c''(\boldsymbol{\theta}) + b'(\boldsymbol{\theta})^2 \mathrm{Var}[a(\mathbf{Y})] = 0.$$

This can be rearranged to

$$\begin{aligned} \mathrm{Var}[a(\mathbf{Y})] &= -\frac{b''(\boldsymbol{\theta})\mathbb{E}[a(\mathbf{Y})] + c''(\boldsymbol{\theta})}{b'(\boldsymbol{\theta})^2} \\ &= -\frac{b''(\boldsymbol{\theta})\frac{-c'(\boldsymbol{\theta})}{b'(\boldsymbol{\theta})} + c''(\boldsymbol{\theta})}{b'(\boldsymbol{\theta})^2} \\ &= \frac{b''(\boldsymbol{\theta})c'(\boldsymbol{\theta}) - c''(\boldsymbol{\theta})b'(\boldsymbol{\theta})}{b'(\boldsymbol{\theta})^3}. \end{aligned} \tag{6.10}$$

*Score Statistic and Information*
From equation (6.3) the log-likelihood function for the distribution of the exponential family is

$$l(\boldsymbol{\theta}|\mathbf{y}) = a(\mathbf{y})b(\boldsymbol{\theta}) + c(\boldsymbol{\theta}) + d(\mathbf{y}), \tag{6.11}$$

and the score statistic and the information are derived from the derivatives of the log-likelihood function w.r.t. $\boldsymbol{\theta}$. The score statistic is

$$U(\boldsymbol{\theta}|\mathbf{y}) = \frac{dl(\boldsymbol{\theta}|\mathbf{y})}{d\boldsymbol{\theta}} = a(\mathbf{y})b'(\boldsymbol{\theta}) + c'(\boldsymbol{\theta}).$$

As $U$ depends on $\mathbf{y}$, it can be regarded as a random variable, i.e.

$$U = a(\mathbf{Y})b'(\boldsymbol{\theta}) + c'(\boldsymbol{\theta}).$$

Using equation (6.8), the expected value of $U$ is

$$\begin{aligned} \mathbb{E}[U] &= \mathbb{E}[a(\mathbf{Y})]\, b'(\boldsymbol{\theta}) + c'(\boldsymbol{\theta}) \\ &= -\frac{c'(\boldsymbol{\theta})}{b'(\boldsymbol{\theta})} b'(\boldsymbol{\theta}) + c'(\boldsymbol{\theta}) \\ &= 0. \end{aligned} \tag{6.12}$$

The information $\mathcal{J}$ is the variance of $U$ and, using equation (6.10),

it is given by

$$\begin{aligned}\mathcal{J} &= \mathrm{Var}\left[U\right] \\ &= b'(\boldsymbol{\theta})^2 \mathrm{Var}\left[a(\mathbf{Y})\right] \\ &= b'(\boldsymbol{\theta})^2 \frac{b''(\boldsymbol{\theta})c'(\boldsymbol{\theta}) - c''(\boldsymbol{\theta})b'(\boldsymbol{\theta})}{b'(\boldsymbol{\theta})^3} \\ &= \frac{b''(\boldsymbol{\theta})c'(\boldsymbol{\theta})}{b'(\boldsymbol{\theta})} - c''(\boldsymbol{\theta}). \end{aligned} \tag{6.13}$$

Furthermore

$$\mathcal{J} = \mathbb{E}\left[U^2\right] = -\mathbb{E}\left[U'\right],$$

where the first equality comes from the fact that for any random variable

$$\mathrm{Var}(U) = \mathbb{E}[U^2] - (\mathbb{E}[U])^2$$

and from equation (6.12). The second equality comes from differentiation of $U$ w.r.t. $\boldsymbol{\theta}$

$$U' = \frac{dU}{d\boldsymbol{\theta}} = a(\mathbf{Y})b''(\boldsymbol{\theta}) + c''(\boldsymbol{\theta}),$$

and taking the expected value of this using equation (6.8) and equation (6.13):

$$\begin{aligned}\mathbb{E}[U'] &= \mathbb{E}[a(\mathbf{Y})]b''(\boldsymbol{\theta}) + c''(\boldsymbol{\theta}) \\ &= -\frac{c'(\boldsymbol{\theta})}{b'(\boldsymbol{\theta})} b''(\boldsymbol{\theta}) + c''(\boldsymbol{\theta}) \\ &= -\mathrm{Var}[U] = -\mathcal{J}. \end{aligned}$$

### Generalized Linear Models

In a GLM, $\mathbf{Y} = [Y_1, \ldots, Y_n]^\top$ is a set of independent random variables, each with a distribution from the exponential family and it has mean vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^\top$. If $\mathbf{Y}$ satisfies the following properties it can be described by a GLM:

1. The distribution of each $Y_i$ is on canonical form and depends on a single parameter $\theta_i$, i.e.

$$f(y_i|\theta_i) = \exp\left(y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)\right).$$

2. The distributions of all the $Y_i$'s are of the same form, so the indices on $b$, $c$ and $d$ are not needed. Thus the joint pdf of $Y_1, \ldots, Y_N$ is

$$\begin{aligned} f(y_1, \ldots, y_n \quad | \quad \theta_1, \ldots \theta_n) &= \prod_{i=1}^{n} \exp\left(y_i b(\theta_i) + c(\theta_i) + d(y_i)\right) \\ &= \exp\left(\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right). \end{aligned}$$

Usually, for model specification, it is a smaller set of unknown parameters $\beta_1, \ldots, \beta_p$, where $p \leq n$ that is interesting rather than the $\theta_i$'s.

Suppose that $\mathbb{E}[Y_i] = \mu_i$, where $\mu_i$ is some function of $\theta_i$ and that $\mathbf{x}_i$ is a $p \times 1$ vector of covariates, then the linear predictor $\boldsymbol{\eta}$, a linear combination of the covariates, is given by

$$\boldsymbol{\eta} = \sum_{i=1}^{p} \mathbf{x}_i^\top \boldsymbol{\beta}.$$

For a GLM there also exists a link function

$$g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{6.14}$$

for $i = 1, \ldots, n$, where $g$ is a monotone, differentiable function, that describes the relationship between the expected value and the linear predictor. $\mathbf{x}_i$ is a $p \times 1$ vector of explanatory variables:

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}, \text{so } \mathbf{x}_i^\top = [x_{i1} \ldots x_{ip}],$$

where $\mathbf{x}_i^\top$ is the $i$'th row of the design matrix $X$. The parameter vector $\boldsymbol{\beta}$ is a $p \times 1$ vector:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

Thus the GLM has three components:

1. Response variables $Y_1, \ldots, Y_N$, which are assumed to share the same distribution from the exponential family.

2. A set of parameters $\boldsymbol{\beta}$ and covariates

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix}$$

3. A monotone, differentiable link function $g$ such that

$$g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}, \text{where } \mu_i = \mathbb{E}[Y_i]$$

for $i = 1, \ldots, n$.

[34, pp. 51-52] [33, pp. 26-28].

## 6.2 Estimation of Parameters

When a model is chosen, it is required to estimate the unknown parameters and to obtain the precision of these. This section describes how to obtain these parameter estimates for GLMs using methods based on maximum likelihood estimation (MLE).

### 6.2.1 Maximum Likelihood Estimation

Let $Y_i, \ldots, Y_n$ be independent random variables that satisfy the properties of GLMs described in section 6.1. It is desired to estimate the parameters $\boldsymbol{\beta}$, which are related to the $Y_i$'s through $\mathbb{E}[Y_i] = \mu_i$ and through the link function $g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$. For GLMs the estimation is done by defining a measure of the goodness of fit between the observed data and the fitted values. Described in this section is estimates obtained by maximizing the likelihood of the parameters for the observed data.

Assuming canonical form and using equation (6.11), the log-likelihood function for each $Y_i$ is defined as

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i),$$

where $i = 1, \ldots, n$ and the functions $b$, $c$ and $d$ are as in section 6.1. The joint log-likelihood function for all the $Y_i$'s is therefore

$$l = \sum_{i=1}^{N} l_i = \sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i).$$

Also, since GLMs are defined in terms of the exponential family, $\mathbb{E}[Y_i]$ and $\mathrm{Var}[Y_i]$ are given by

$$\mathbb{E}[Y_i] = \mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)}, \tag{6.15}$$

$$\mathrm{Var}[Y_i] = \frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{b'(\theta_i)^3}, \tag{6.16}$$

as shown in section 6.1. The link function is given by

$$g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}. \tag{6.17}$$

To maximize the log-likelihood function it is necessary to find the derivative with respect to the parameter $\beta_j$. This is the score function and is found by use of the multivariable chain rule:

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \left( \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \right), \tag{6.18}$$

for all $j = 1,2, \ldots, p$.

Each of the terms on the right hand side of equation (6.18) is computed separately. Using equation (6.15), the first term is

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i).$$

The second term can be rewritten as

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}}.$$

Differentiation of equation (6.15) yields

$$\begin{aligned} \frac{\partial \mu_i}{\partial \theta_i} &= \frac{-c''(\theta_i)b'(\theta_i) + c'(\theta_i)b''(\theta_i)}{b'(\theta_i)^2} \\ &= b'(\theta_i)\mathrm{Var}\left[Y_i\right], \end{aligned}$$

using equation (6.16). Therefore,

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b'(\theta_i)\text{Var}\left[Y_i\right]}.$$

The last term is rewritten using equation (6.17) and the chain rule:

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}x_{ij}.$$

Thus, the scores $U_j$ are given by

$$\begin{aligned} U_j &= \sum_{i=1}^{N} b'(\theta_i)(y_i - \mu_i)\frac{1}{b'(\theta_i)\text{Var}\left[Y_i\right]}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} \\ &= \sum_{i=1}^{N}\left[\frac{y_i - \mu_i}{\text{Var}\left[Y_i\right]}x_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right], \end{aligned} \tag{6.19}$$

for all $j = 1,2,\ldots,p$.

The expected information matrix $\mathcal{J}$ is found in terms of the variance-covariance of the $U_j$'s:

$$\mathcal{J}_{jk} = \mathbb{E}[U_j U_k] = \mathbb{E}\left[\frac{\partial l}{\partial \beta_j}\frac{\partial l}{\partial \beta_k}\right].$$

By use of the score function given in equation (6.19), the information matrix is

$$\begin{aligned} \mathcal{J}_{jk} &= \mathbb{E}\left[\left(\sum_{i=1}^{n}\left[\frac{Y_i - \mu_i}{\text{Var}\left[Y_i\right]}x_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]\right)\right. \\ &\qquad \left.\left(\sum_{h=1}^{n}\left[\frac{Y_h - \mu_h}{\text{Var}\left[Y_h\right]}x_{hk}\left(\frac{\partial \mu_h}{\partial \eta_l}\right)\right]\right)\right]. \end{aligned}$$

Since the $Y_i$'s are independent, $\mathbb{E}[(Y_i - \mu_i)(Y_h - \mu_h)] = 0$ for $i \neq h$, and therefore

$$\mathcal{J}_{jk} = \sum_{i=1}^{N}\frac{\mathbb{E}\left[(Y_i - \mu_i)^2\right]}{\text{Var}\left[Y_i\right]^2}x_{ij}x_{ik}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{6.20}$$

Equation (6.20) can be simplified by applying that $\mathbb{E}[(Y_i - \mu_i)^2] = \text{Var}[Y_i]$:

$$\mathcal{J}_{jk} = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{\text{Var}\left[Y_i\right]}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \tag{6.21}$$

for all $j$ and $k$.

### Fisher Scoring Method

The MLE is the solution in which the scores are all equal to zero. In order to solve this equation, a Newton-Raphson algorithm can be used. Normally this algorithm uses the Hessian matrix, but using Fisher scoring it will be replaced by the expected information $\mathcal{J}$. This is done, since it is easier to calculate. Let $\hat{\boldsymbol{\beta}}^{(m)}$ denote the estimation of the vector $\boldsymbol{\beta}$ at the $m$'th iteration. Then the $(m+1)$'th estimate is given by

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} - J^{-1}\mathbf{U}(\hat{\boldsymbol{\beta}}^{(m)}), \tag{6.22}$$

where $\hat{\boldsymbol{\beta}}^{(m+1)}$ is the vector of estimates of the vector parameter $\boldsymbol{\beta}$ at the $(m+1)$'th iteration. The term $\left[\mathcal{J}^{(m)}\right]^{-1}$ is the inverse of the expected information matrix with elements $\mathcal{J}_{ik}$ given in equation (6.21) evaluated at $\hat{\boldsymbol{\beta}}^{(m)}$, and $\mathbf{U}^{(m)}$ is the vector of elements given in equation (6.19), both evaluated at $\hat{\boldsymbol{\beta}}^{(m)}$.

If both sides of equation (6.22) are multiplied by $\mathcal{J}^{(m)}$, we get

$$\mathcal{J}^{(m)}\hat{\boldsymbol{\beta}}^{(m+1)} = \mathcal{J}^{(m)}\hat{\boldsymbol{\beta}}^{(m)} + \mathbf{U}(\hat{\boldsymbol{\beta}}^{(m)}), \tag{6.23}$$

and from equation (6.21) the information $\mathcal{J}$ can be written as

$$\mathcal{J} = X^{\top}WX,$$

where $W$ is a $n \times n$ diagonal matrix with elements

$$w_{ii} = \frac{1}{\mathrm{Var}[Y_i]}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{6.24}$$

The expression on the right side of equation (6.22) evaluated at $\hat{\boldsymbol{\beta}}^{(m)}$. is the vector with elements

$$\begin{aligned}
&\sum_{k=1}^{p}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{\mathrm{Var}[Y_i]}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\hat{\beta}_k^{(m)} + \sum_{i=1}^{N}\frac{(y_i-\mu_i)x_{ij}}{\mathrm{Var}[Y_i]}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\\
= &\sum_{k=1}^{p}\sum_{i=1}^{n}x_{ij}w_{ii}^{(m)}x_{ik}\hat{\beta}_k^{(m)} + \sum_{i=1}^{N}x_{ij}w_{ii}^{(m)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)(y_i-\mu_i)\\
= &\sum_{i=1}^{n}x_{ij}w_{ii}^{(m)}\left(\sum_{k=1}^{p}x_{ik}\hat{\beta}_k^{(m)} + \left(\frac{\partial \mu_i}{\partial \eta_i}\right)(y_i-\mu_i))\right)\\
= &\sum_{i=1}^{n}x_{ij}w_{ii}^{(m)}z_i^{(m)},
\end{aligned}$$

where $z_i^{(m)} = \sum_{k=1}^{p} x_{ik}\hat{\beta}_k^{(m)} + \left(\frac{\partial \mu_i}{\partial \eta_i}\right)(y_i - \mu_i))$ and with $\mu_i$ and $\frac{\partial \eta_i}{\partial \mu_i}$ evaluated at $\hat{\boldsymbol{\beta}}^{(m)}$. This follows from equations (6.21) and (6.19). Therefore the right hand side of equation (6.23) can be written as

$$X^\top W\mathbf{z},$$

where $\mathbf{z}$ has elements

$$z_i = \sum_{k=1}^{p} x_{ik}\hat{\beta}_k^{(m)} + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right), \tag{6.25}$$

with $\mu_i$ and $\frac{\partial \eta_i}{\partial \mu_i}$ evaluated at $\hat{\boldsymbol{\beta}}^{(m)}$ and is called the adjusted dependent variable. Hence equation (6.23) can be written as

$$X^\top W X\hat{\boldsymbol{\beta}}^{(m+1)} = X^\top W\mathbf{z}. \tag{6.26}$$

This is on the same form as normal equations for a linear model obtained by weighted least squares, except this has to be solved iteratively. Thus by using the Fisher scoring method to solve the scores the MLE of $\boldsymbol{\beta}$ is equivalent to the MLE of $\boldsymbol{\beta}$ produced by the iterative weighted least squares procedure.

Most statistical packages that includes procedures for fitting GLM have an algorithm based on equation (6.26). Begin by using an initial approximation $\hat{\boldsymbol{\beta}}^{(0)}$ to evaluate $\mathbf{z}$ and $W$. Then equation (6.26) is solved to yield $\hat{\boldsymbol{\beta}}^{(1)}$, which in turn is used to obtain better approximation for $\mathbf{z}$ and $W$, and so on until the change between succeeding iterations $\hat{\boldsymbol{\beta}}^{(m)}$ and $\hat{\boldsymbol{\beta}}^{(m-1)}$ are sufficiently small. When the difference is sufficiently small, $\hat{\boldsymbol{\beta}}^{(m)}$ is chosen as the maximum likelihood estimate. [34, ch.4] [33, pp. 23-25] [36, pp. 137-139].

## 6.3 Models for Count Data

In this thesis we are dealing with count data. Count data is for example a number of certain events within a fixed period of time or frequencies in cells of contingency tables. In our case we have daily counts of incidences of CVDs as response. When modeling count data it is often reasonable to assume an underlying Poisson distribution $\text{Po}(\mu)$. The Poisson distribution is defined as follows

**Definition 6.1** (Poisson Distribution)
Let $Y$ be the number of occurrences, then the probability distribution can be written as

$$f(y) = \frac{\mu^y e^{-\mu}}{y!},$$

where $y = 0,1,2,\ldots$ and $\mu$ is the average number of occurrences. For the Poisson distribution, $\mathbb{E}[Y] = \mu$ and $\text{Var}[Y] = \mu$.

The parameter $\mu$ is often described as a rate in terms of units of exposure. The effect of covariates on the response $Y$ is modeled through $\mu$.

In this chapter two situations are described. In the first situation the events are subject to varying amounts of exposure, which needs to be taken into account in the modeling of the events. The other covariates may be continuous or categorical. In this situation Poisson regression is used. In the other situation the events have a constant amount of exposure and the covariates are usually categorical. If there is only a few covariates the data is usually summarized in a cross-classified table, where the response is the frequency in every cell of the table. The variables which define the table is treated as covariates. The study design can cause constraints on the cell frequencies e.g. the population is equal to 400. These constraints need to be taken into account in the modeling. The term log-linear model is used for GLMs appropriate for this situation. [34, pp.165-166].

### 6.3.1 Poisson Regression

Let $Y_1, \ldots, Y_n$ be independent random variables with $Y_i$ denoting the number of events observed from the exposure $n_i$ for the $i$th covariate pattern. The expected value of $Y_i$ can be written as

$$\mathbb{E}[Y_i] = \mu_i = n_i\theta_i.$$

For example, suppose $Y_i$ is the number of insurance claims for a particular make and model of a car. This number $Y_i$ will depend on the number of cars of this type that are insured at the insurance company, $n_i$, and other variables that effect $\theta_i$, such as age of the car and where it is used.

The dependence of $\theta_i$ on the explanatory variables is usually modeled by

$$\theta_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}), \tag{6.27}$$

so the GLM is

$$\mathbb{E}[Y_i] = \mu_i = n_i \exp(\mathbf{x}_i^\top \boldsymbol{\beta}),\ Y_i \sim \mathrm{Po}(\mu_i)$$

and the natural link function is the logarithmic function

$$\log(\mu_i) = \log(n_i) + \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Here the term $\log(n_i)$ is called the offset and is a known constant, which is readily incorporated into the estimation procedure. $\mathbf{x}_i$ is the covariate pattern and $\boldsymbol{\beta}$ is the parameter vector as usual.

### 6.3.2 Log-Linear Models

Before specifying log-linear models in contingency tables it is important to consider how the study design has limited the data, since the limitations will influence the choice of the probability models to describe the data.

#### Probability models for contingency tables

Let $\mathbf{y}$ be a vector describing the frequencies $Y_i$ in $N$ cells of the cross-tabulated table.

#### Poisson Model

If there is no constraints on the $Y_i$'s they can be modeled as independent random variables with parameters $\mathbb{E}[Y_i] = \mu_i$ and joint pdf

$$f(\mathbf{y},\mu) = \prod_{i=1}^{N} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!},$$

where $\mu$ is a vector of the $\mu_i$'s.

### Multinomial Model

If the only constraint is, that the sums of the $Y_i$'s is $n$, then a multinomial model is used:

$$f(\mathbf{y},\mu|n) = n! \prod_{i=1}^{n} \frac{\theta_i^{y_i}}{y_i!},$$

where $\sum_{i=1}^{n} \theta_i = 1$ and $\sum_{i=1}^{N} y_i = n$. Here $\mathbb{E}[Y_i] = n\theta_i$. For a 2 dimensional table, let $j$ denote the rows and $k$ denote the columns, then the most common hypothesis is if the rows and columns are independent, so

$$\theta_{jk} = \theta_{j\cdot}\theta_{\cdot k},$$

where $\theta_{j\cdot}$ and $\theta_{\cdot k}$ is the marginal probabilities with $\sum_j \theta_{j\cdot} = 1$ and $\sum_k \theta_{\cdot k} = 1$. This hypothesis can be tested by comparing the fit of two linear models for the logarithm of $\mu_{jk} = \mathbb{E}[Y_{jk}]$;

$$\begin{aligned} \log(\mu_{jk}) &= \log(n) + \log(\theta_{jk}) \text{ and} \\ \log(\mu_{jk}) &= \log(n) + \log(\theta_{j\cdot}) + \log(\theta_{\cdot k}). \end{aligned}$$

### Product Multinomial Model

If there are more fixed marginals than just the total $n$, then an appropriate product of multinomial distributions can be used as model.

For example a three dimensional table with $J$ rows, $K$ columns and $L$ layers, if the row totals are fixed in every layer the joint pdf of the $Y_{jkl}$'s is

$$f(\mathbf{y}|y_{j\cdot l}, j = 1, \cdots, J, l = 1, \cdots, L) = \prod_{j=1}^{J}\prod_{l=1}^{L} y_{j\cdot l}! \prod_{k=1}^{K} \frac{\theta_{jkl}^{y_{jkl}}}{y_{jkl}!},$$

where $\sum_j \sum_k \theta_{jkl} = 1$ for $l = 1, \ldots, L$ and $\mathbb{E}[Y_{jkl}] = y_{\cdot\cdot l}\theta_{jkl}$.

All the above mentioned is based on the Poisson distribution and the $\mathbb{E}[Y_i]$'s can be written as a product of parameters and other terms. Therefore the natural link function for the Poisson distribution yields a linear component

$$\log(\mathbb{E}[Y_i]) = \text{constant} + \mathbf{x}_i^\top \boldsymbol{\beta}.$$

The term log-linear model is used to describe all these GLMs. They are typically hierarchical, which means that if a higher order term is included in the model, all the related lower order terms have to be included in the model as well. [34, pp. 171-178].

# Generalized Additive Models

A generalized additive model (GAM) is an extension of the GLM, covered in chapter 6.

## 7.1 Additive Models

Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote the response variable and let $\mathbf{x}_i^\top = (x_{i1}, \ldots, x_{ip})$ be the $i$'th row of an $n \times p$ design matrix $X$, where each $\mathbf{x}_i^\top$ is a $1 \times p$ vector of explanatory variables associated with $y_i$. It is desired to model the dependence of $y_i$ on $\mathbf{x}_i$. As mentioned in section 6.1, the reason to do this is to learn more about the process that produces $y_i$ and to asses the relative contributions in explaining $y_i$ and maybe to predict $y_i$ for $\mathbf{x}_i$.

A standard tool for this is a multiple linear regression given by

$$\mathbb{E}\left[Y|\mathbf{x}_i\right] = \alpha + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + \epsilon_i,$$

where $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}[\epsilon_i] = \sigma^2$ and $i = 1, \ldots, n$. The problem with this regression model is that it assumes the dependence of $y_i$ on $\mathbf{x}_i^\top$ is linear. A way to generalize this model, so the dependence of $y_i$ on $\mathbf{x}_i^\top$ is no longer linear is by use of a smoothing function, often just called a smoother. This can be thought of as a description of the dependence of $y_i$ on $\mathbf{x}_i^\top$ and as nonparametric estimates of the regression model

$$\mathbb{E}\left[Y_i|\mathbf{x}_i\right] = s(x_{i1}, \ldots, x_{ip}) + \epsilon_i,$$

where $i = 1, \ldots, n$ . Smoothers are described further in the next section.

An additive model (AM) extends the standard linear regression by modeling the mean of the response $y_i$ as an additive sum of smoothed effects from the covariates and is defined by

$$\mathbb{E}\left[Y_i|\mathbf{x}_i\right] = \alpha_i + \sum_{j=1}^{p} s_j(x_{ij}) + \epsilon_i,$$

where $i = 1, \ldots, n$ and the errors $\epsilon_i$ are independent of the $x_{ij}$'s, $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$. The smoothers, $s_j(\cdot)$, are arbitrary univariate functions for each covariate and $\mathbb{E}[s_j(x_{ij})] = 0$. [36, pp. 1-5, 82-89].

### 7.1.1 Smoothing

A smoother is a tool that summarizes the trend of the response $y_i$ as a function of measurements $\mathbf{x}_i^\top$, but has less variation than $y_i$ itself. An important property of a smoother is that it is nonparametric, i.e. that the smoother does not assume a rigid form for dependence of $y_i$ on $\mathbf{x}_i^\top$. For a single covariate we have $\mathbb{E}[Y_i|x_i] = s(x_i) + \epsilon_i$, where $i = 1, \ldots, n$ and $s(x_i)$ is an arbitrary smooth function to be estimated by any scatterplot smoother, e.g. a running mean, running line, kernel estimate or a spline. A scatterplot smoother is defined as following

**Definition 7.1** (Scatterplot Smoothing)
Suppose we have data of the form $(x_1,y_1), \ldots, (x_n, y_n)$ and let $\mathbf{y} = (y_1, \ldots, y_n)^\top$ be the response and $\mathbf{x} = (x_1, \ldots, x_n)^\top$ denote the covariates. A scatterplot smoother of $\mathbf{y}$ against $\mathbf{x}$ is then defined as a function

$$s(x) = \mathcal{S}(x|\mathbf{x}, \mathbf{y}),$$

which at each $x \in \mathbb{R} \in \{x_1, \ldots, x_n\}$ estimates the response $y_i$ on $x$.

Often there is more than one covariate, and for $p$ covariates $\mathbf{x}_j^\top = (x_{1j}, x_{2j}, \ldots, x_{nj})$, where $j = 1, \ldots, p$ it can be solved as in the single covariate case, by adding polynomial terms in a $p$-dimensional scatterplot smoother, but it will be rather difficult to choose which terms will be appropriate. For simplicity, let the $p$ covariates be denoted $\mathbf{x}_j^\top = (x_{1j}, x_{2j}, \ldots, x_{nj}) = X_j$. Let $X = (X_1, \ldots, X_p)$. The additive model is described by $\mathbb{E}[Y_i|X] = \sum_{j=1}^{p} s_j(X_j)$ and can be estimated iterative by a so called Backfitting Algorithm. Consider two covariates $X_1$ and $X_2$ then $\mathbb{E}[Y_i|X_1, X_2] = \sum_{j=1}^{2} s_j(X_j)$. The estimate of $s_1$ is found as in the single-case manner. Given the estimate of $s_1$ an intuitive way to estimate $s_2$ is by smoothing the residual $\mathbf{y} - s_1$ on $X_2$. With this estimate of $s_2$ an improved estimate of $s_1$ is given by smoothing the residual $\mathbf{y} - s_2$ on $X_1$ an so on until the estimates of $s_1$ and $s_2$ are such that the smooth of

$\mathbf{y} - s_1$ on $X_2$ is the estimate of $s_2$ and the other way around. This iterative smoothing process is an example of a backfitting algorithm.

There are two main uses for smoothers. The first is description; the scatterplot smoother can be used visually to see the trend in a scatterplot of $y_i$ against $x_{ij}, j = 1, \ldots, p$. The second use of the smoother is to estimate the dependence of the mean of $y_i$ on $\mathbf{x}_i^\top$, and thus serve as a foundation for the estimation of AMs. An example of a simple smoother is in the case of a categorical predictor, e.g. gender. In this case, to smooth $\mathbf{y}$, the values of $\mathbf{y}$ can be averaged in each category. This satisfies the requirements for a scatterplot smooth; it captures the trend of $y_i$ on $\mathbf{x}_i^\top$ and is smoother than the $\mathbf{y}$ values themselves. This is the basic concept for smoothing in the most general setting. Most smoothers imitate this category averaging through local averaging, that is, averaging the $\mathbf{y}$-values of observations having the covariate values close to a target value. The averaging is then done in neighborhoods around the target value. This sums up to two main decisions that has to be made in scatterplot smoothing:

1. How to average the $\mathbf{y}$ values in each neighborhood.
2. How large the neighborhoods should be.

How to average the $\mathbf{y}$ values in a neighborhood is a question of which type of smoother to use, since smoothers differs mainly in their method of averaging.Intuitively, large neighborhoods will produce an estimate with low variance but potentially high bias, and vice versa for small neighborhoods. Thus there is a fundamental trade-off between bias and variance, determined by the smoothing parameter. [36, pp. 1-13] [37].

**Smoothers**

There are a lot of different smoothers, e.g. parametric regression, bin smoother, running-mean and splines. For the GAM analysis in this thesis we have used a LOESS smoother.

Let the target value $x_0$ denote one of the $x_j$s, and assume there are no duplicates of it. If there is, the average of the $\mathbf{y}$ values for the $x_0$'s is simply used as the estimate of $s(x_0)$. If there are no

duplicates of $x_0$, the $\mathbf{y}$ values corresponding to the $x$ values close to $x_0$ is averaged, this is what a running-mean smoother does. A way to choose which $x$ values are close to $x_0$ is to choose $x_0$ itself and the $k$ points that are closest to $x_0$ on the left and right side. This is called a symmetric nearest neighborhood and is denoted $N^s(x_0)$. A definition for a symmetric nearest neighborhood for a arbitrary $x_i$ is given by

$$N^s(x_i) = \{j | \max(i-k, 1) \leq j \geq \min(i+k, n)\},$$

and the running-mean for the target value $x_0$ is then defined as

$$s(x_0) = \frac{\sum_{j \in N^s(x_0)}(y_i)}{|N^s(x_0)|}.$$

If it is not possible to choose $k$ points at each side of $x_i$, as many as possible are chosen. How to define the symmetric nearest neighbors at target points $x_0$ other than the $x_i$ in the sample is not obvious. The fit between the two values in $x$ in the sample adjacent to $x_0$ can be interpolated linearly, or alternatively the symmetry can be ignored and the $r$ points closest to $x_0$, no matter which side, are chosen. This is called nearest neighborhood. In practice this simple smoother does not work very well and tends to be very wiggly and have a tendency to flatten out near the endpoints. Because of this, it can be severely biased. A generalization of the running-mean that eases the bias problem is to compute a least-squares line instead of a mean in each neighborhood. The running-line smoother is defined by

$$s(x_i) = \hat{\alpha}(x_i) + \hat{\beta}(x_i)x_i.$$

were $\hat{\alpha}(x_i)$ and $\hat{\beta}(x_i)$ are least-square estimates for the data points $(x_j, y_j)$, where $j \in N^s(x_i)$. The estimated smooth at $x_i$ is then the value of the fitted line at $x_i$. This is done for each $x_i$. The running-line captures the trend in the data, but is still jagged.

The appearance of the running-line smoother is controlled by the parameter $k$. Large values of $k$ tend to produce smoother curves than small values of $k$. In extreme cases where each neighborhood contains all the data, the running-line is the least-squares. If each neighborhood consists of only the point itself and one neighbor, the smoother interpolates the data.

A way to improve the appearance of the running-line smooth is by using weighted least-squares fit in each neighborhood. The running-line can produce jagged output because points outside the neighborhood are given zero weight and points inside the neighborhood are given equal (nonzero) weight. Thus as the neighborhoods move from left to right, there are changes in the weight given to the leftmost and rightmost points. This problem can be solved by giving the highest weight to $x_i$ and let the weight smoothly decrease when moving further away from $x_i$. This is called a locally-weighted running-line smoother (LOESS), where the weights are assigned to each data point in the neighborhood by use of the tri-cube weight function

$$W\left(\frac{|x_0 - x_j|}{\delta(x_j)}\right),$$

where $\delta(x_0) = \max_{j \in N^s(x_0)} |x_0 - x_j|$ is the distance of the nearest neighbor furthest away and the weights are then assigned by

$$W(u) = \begin{cases} (1-u^3)^3 & \text{for} \quad 0 \leq u < 1 \\ 0 & \text{otherwise.} \end{cases}$$

[36, pp. 14-18] [38].

## 7.2 Generalized Additive Models

In section 6.1 GLMs were defined. GLMs, that themselves are a generalization of linear regression models can also have an additive extension. As in linear models the predictor effects in a GLM are assumed to be linear, GAMs extend these models the same way additive models extend a linear regression model, by replacing some of the linear terms $\sum_{j=1}^{p} X_j \boldsymbol{\beta}_j$ with smoothers.

### 7.2.1 Estimation of a GAM via local scoring

In section 6.2.1 a way to compute the MLE for a GLM was presented. The method was Fisher scoring and was shown to be equivalent to the iterative weighted least squares procedure, where the estimates were found by repeatedly regressing the adjusted dependent variable $z_i$, equation (6.25), on $\mathbf{x}_i$ with weights $w_i$ from equation (6.24). This technique can also be used to estimate the smoothing functions $s(\cdot)$'s in GAMs, by repeatedly smoothing the adjusted dependent variable

on $X$. Therefore the estimation of the $s(\cdot)$'s is done by replacing the linear predictor $\eta_i = \sum_{j=1}^{p} X_j \boldsymbol{\beta}_j$ in the adjusted dependent variable, equation (6.25), with the additive predictor $\eta_i = \sum_{j=1}^{p} s_j(X_j)$, where some of the functions $s_j$ may be linear. Apart from this change in $z_i$ the procedure is the same as shown in section 6.2.1 and is called local scoring since local averaging is used to generalize the procedure. [37] [36, pp. 136-141].

# Inference and Model Validation

## 8.1 Inference

A main tool for statistical inference is hypothesis testing, where two related models are compared to see how they fit data. For a GLM, the two models in the hypothesis test should have the same probability distribution and the same link function, but the parameters of the linear component of one model is an extension of the parameters of the linear components of the other model. To make comparisons of the two models, summary statistics are used. The summary statistics are used to describe how well the model fits the data and are therefore called goodness of fit statistics. Goodness of fit statistics can among others be score, Wald or likelihood ratio statistics.

### 8.1.1 Goodness of Fit and Deviance

The goodness of fit of a statistical model describes how well the model fits the data. Goodness of fit can be assessed using a number of measures that generally compare the observed values with the fitted values generated by the model.

Two models can be tested by comparing their goodness of fit. The two models need to be either nested or hierarchical, in other words the models have to have the same probability distribution and the same link function, but the linear component of the simpler model $M_0$ is a special case of the linear component of the more general model $M_1$.

Let $\hat{\boldsymbol{\beta}}_{max}$ denote the MLE of the parameter vector for the saturated model, i.e. the model with as estimated parameters number of observations. Let $\hat{\boldsymbol{\beta}}$ denote the MLE for the simpler model $M_0$. The likelihood function for the saturated model evaluated at $\hat{\boldsymbol{\beta}}_{max}$ is denoted by $L(\hat{\boldsymbol{\beta}}_{max}|\mathbf{y})$ and $L(\hat{\boldsymbol{\beta}}|\mathbf{y})$ denotes the maximum value of the

likelihood function of the model $M_0$ evaluated at $\hat{\boldsymbol{\beta}}$. Then the likelihood ratio is given by

$$\lambda = \frac{L(\hat{\boldsymbol{\beta}}_{max}|\mathbf{y})}{L(\hat{\boldsymbol{\beta}}|\mathbf{y})}.$$

Often the logarithm of the likelihood ratio is used. It gives the difference between the log-likelihood functions,

$$\log(\lambda) = l(\hat{\boldsymbol{\beta}}_{max}|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}|\mathbf{y}). \tag{8.1}$$

Large values of equation (8.1) imply that the model of interest, $M_0$, does not fit the data well compared to the saturated model.

The likelihood ratio statistic, which is also called the deviance, has a chi-squared distribution and for that reason it is more used than $\log(\lambda)$. The deviance is given by

$$D = 2\log(\lambda)\phi = 2(l(\hat{\boldsymbol{\beta}}_{max}|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}|\mathbf{y})).$$

If $_O$ is an adequate model, then $D$ has a $\chi^2_{(m-p)}$ distribution, where $m$ is the number of parameters in the saturated model and $p$ is the number of parameters in the simpler model. [39, section 2.1.6] [34, chapter 5]

### 8.1.2 Hypothesis testing

Consider a hypothesis $H_1$, corresponding to an extensive model $M_1$ of $p$ parameters. The null hypothesis $H_0$ states that a simpler model $M_0$ of $q$ parameters, of which $M_1$ is an extension, is as suitable to describe data as $M_1$. The hypotheses are

$$\begin{aligned} H_1 \quad &: \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \\ \beta_{q+1} \\ \vdots \\ \beta_p \end{bmatrix}, \\ H_0 \quad &: \quad \beta_{q+1} = \ldots = \beta_p = 0, \end{aligned}$$

where $q < p < N$ and N is number of observations, i.e. the null hypothesis states that the extra parameters in model $M_1$ are all zero. The two hypotheses are tested by using the difference in the deviances:

$$\triangle D = D_1 - D_0 = 2\left(l(\hat{\boldsymbol{\beta}}_1|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}_0|\mathbf{y})\right),$$

where $\hat{\boldsymbol{\beta}}_1$ is the MLE of model $M_1$ and $\hat{\boldsymbol{\beta}}_0$ is the MLE of model $M_0$. If both of the models describe data well then $D_0 \sim \chi^2_{(n-q)}$, $D_1 \sim \chi^2_{(n-p)}$, and $\triangle D \sim \chi^2_{(p-q)}$. If $\triangle D$ is consistent with the $\chi^2_{(p-q)}$ distribution, the model $M_0$ is chosen, since it is the simplest one. If $D_0$ is larger than what is expected from $\chi^2_{(n-q)}$, then model $M_0$ does not describe the data well. If $M_1$ describes data well, but $M_0$ does not, then $\triangle D$ is bigger than what would be expected from $\chi^2_{(p-q)}$. For nonparametric and additive models, the deviance still makes sense in assessing models and their differences, but distribution theory is undeveloped. [34, chapter 5] [36, pp. 155-158].

If $M_0$ and $M_1$ do not satisfy that $M_0 \subseteq M_1$ the above described goodness of fit test does not apply. Instead Akaike's Information Criterion (AIC) can be used. AIC is also based on the log-likelihood function and has an adjustment for the number of parameters estimated and for the amount of data. For a model AIC is defined as

$$AIC = -2l(\hat{\boldsymbol{\beta}}|\mathbf{y}) + 2p,$$

where $p$ is the number of estimated parameters. AIC is calculated for both models and the one with the lowest AIC is chosen. [34, p.137].

### 8.1.3 Goodness of Fit for Poisson Models

In this study we use models with Poisson distributions; therefore the goodness of fit for Poisson models is given.
The fitted values of a Poisson model are given by

$$\hat{y}_i = \hat{\mu}_i = n_i \exp(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}),$$

where $i = 1, \ldots, N$. For a Poisson distribution $\mathbb{E}[Y_i] = \mathrm{Var}[Y_i]$, so the standard error of $y_i$ is estimated by $\sqrt{\hat{y}_i}$ and the Pearson residuals are given by

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}, \tag{8.2}$$

where $y_i$ is the observed values of $Y_i$.

In the Poisson distribution the Pearson residuals in equation (8.2) and the chi-squared goodness of fit statistic are related by

$$X^2 = \sum_{i=1}^{n} r^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}.$$

*Deviance for a Poisson Model*
If the responses $Y_1, \ldots, Y_n$ are independent and $Y_i \sim \text{Po}(\lambda_i)$, the log-likelihood function is

$$l(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} y_i \log(\lambda_i) - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \log(y_i!).$$

In the saturated model the $\lambda_i$'s are all different, so $\boldsymbol{\beta} = [\lambda_1, \ldots, \lambda_N]^\top$. The maximum value of the log-likelihood function are

$$l(\hat{\boldsymbol{\beta}}_{max}|\mathbf{y}) = \sum_{i=1}^{n} y_i \log(y_i) - \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log(y_i!),$$

since the maximum likelihood estimates are $\hat{\lambda}_i = y_i$. Suppose that the model of interest have $p < N$ parameters. The MLE $\hat{\boldsymbol{\beta}}$ can be used to calculate estimates $\hat{\lambda}_i$ and, hence, fitted values $\hat{y}_i = \hat{\lambda}_i$, since $\mathbb{E}[Y_i] = \lambda_i$. In this case the maximum value of the log-likelihood is

$$l(\hat{\boldsymbol{\beta}}|\mathbf{y}) = \sum_{i=1}^{n} y_i \log(\hat{y}_i) - \sum_{i=1}^{n} \hat{y}_i - \sum_{i=1}^{n} \log(y_i!).$$

Hence, the deviance is

$$\begin{aligned} D &= 2(l(\hat{\boldsymbol{\beta}}_{max}|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}|\mathbf{y})) \\ &= 2\sum_{i=1}^{n} \left[ y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i) \right]. \end{aligned} \tag{8.3}$$

The goodness of fit statistic $X^2$ and the deviance $D$ are closely related; by the use of the Taylor expansion

$$y \log\left(\frac{y}{\hat{y}}\right) = (y - \hat{y}) + \frac{1}{2}\frac{(y - \hat{y})^2}{\hat{y}} + \ldots,$$

so approximately from equation (8.3)

$$\begin{aligned} D &= 2\sum_i \left[ (y_i - \hat{y}_i) + \frac{1}{2}\frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} - (y_i - \hat{y}_i) \right] \\ &= \sum_{i=1}^{N} \left[ \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i} \right] \\ &= X^2. \end{aligned}$$

The statistics $D$ and $X^2$ can be used directly as measures of goodness of fit, since they both can be calculated from the data and the fitted model, because they do not contain any nuisance parameters. They can be compared with the central chi-squared distribution with $N-p$ d.o.f., where $p$ is the number of estimated parameters. [34, pp. 166-171].

## 8.2 Model Validation

Model validation is an essential part of statistics. It is an important tool used to asses the accuracy of the chosen model and the achieved results, and any thorough statistical analysis is completed by checking the model. When a model fits the data well, the predicted data generated by the model are similar to the observed data.

The residuals from the fitted model are the difference between the observed data, $\mathbf{y} = (y_1, \ldots, y_n)$, and the predicted data, $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)$, generated by the chosen model, i.e. for each $i = 1, \ldots, n$ the $i$'th residual is

$$r_i = y_i - \hat{y}_i.$$

If the model is adequate, the residuals should appear completely random and should, since presumably random, not have any structural relationship. On the other hand, if there seem to be a non-random structure the model fits the data poorly. The residuals can be evaluated graphically. The different aspects of the model can be evaluated using different types of plots.

### Standardized Residuals Plot and QQ-plot

Residuals are standardized by subtracting the mean of the residuals and dividing by their standard deviation, i.e. if the model is appropri-

ate, the standardized residuals approximately follow a standard normal distribution. When plotted against the fitted values in a scatter plot, the standardized residuals should lie close to the line $y = 0$, typically within $y = \pm 2$, and an equal number of points should be above and below the line and should appear non-systematic. When plotted in a QQ-plot, the standardized residuals should follow a straight line.

### Leverage Plot

Leverage points are observations made at extreme or outlying values of the independent variables. The measure of leverage is given by the diagonal of the hat matrix, i.e. the matrix that maps the vector of observations into the vector of fitted values. The leverages, denoted $h_{ii}$, can be plotted on the $y$-axis against the squared residuals on the $x$-axis. Observations with large residuals and large leverages indicate that the model fits the data poorly. [33, pp. 404-405][36, p. 75].

### Cook's Distance

Cook's distance is commonly used as an estimate of the influence of an observation on the regression model; it measures the effect it would have to delete a given observation from the dataset. Especially observations with large residuals or high leverage are of interest and could possibly distort the outcome of the regression. Cook's distance is calculated as

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p \cdot \text{MSE}},$$

where $\hat{y}_j$ is the prediction from the full regression model, $\hat{y}_{j(i)}$ is the prediction from the model where observation $i$ is omitted, $p$ is the number of fitted parameters, and MSE is the mean square error of the regression model. The larger $D_i$ is, the more influence the $i$'th observation has on the model. [36, p. 256][33, pp. 406-407].

Plots from the validation of our model can be found at `http://homes.student.aau.dk/cbisga07/`.

# State Space Models

When serial correlation is present in data, GLMs, chapter 6, may not be adequate for analysis. One approach to solve this problem is to use dynamic generalized linear models (DGLM), also referred to as state space models (SSMs). In this chapter, general SSMs are defined, and the filtering, smoothing and forecasting processes for SSMs are derived.

Consider a time series $(Y_t)_{t\geq 1}$. Let $y_{1:t}$ denote the set of observations $\{y_1, y_2, \ldots, y_{t-1}, y_t\}$. The time series is said to be a Markov chain if, for any $t > 1$,

$$p\left(y_t|y_{1:t-1}\right) = p\left(y_t|y_{t-1}\right),$$

that is, if the future state of the time series depends only on the present state and not on past states. This can also be expressed as $Y_t$ and $Y_{1:t-2}$ being conditionally independent given $y_{t-1}$.
For a Markov chain the finite-dimensional joint distributions can be written as

$$p\left(y_{1:t}\right) = p\left(y_1\right)\prod_{i=2}^{t} p\left(y_i|y_{i-1}\right).$$

A SSM is defined as follows.

**Definition 9.1** (State Space Model)
A state space model consists of a latent $\mathbb{R}^p$-valued time series $(\boldsymbol{\theta}_t)$, $t = 0,1,\ldots,$ and an observed $\mathbb{R}^m$-valued time series $(\mathbf{Y}_t)$, $t = 1,2,\ldots,$ that satisfy the following conditions:

1. $(\boldsymbol{\theta}_t)$ is a Markov chain.
2. Conditionally on $(\boldsymbol{\theta}_t)$, the $\mathbf{Y}_t$'s are independent and $\mathbf{Y}_t$ depends on $\boldsymbol{\theta}_t$ only.

As a result of definition 9.1, a SSM is completely specified by the initial distribution $p\left(\boldsymbol{\theta}_0\right)$ and the conditional densities $p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}\right)$ and $p\left(\mathbf{y}_t|\boldsymbol{\theta}_t\right)$ for $t \geq 1$. Therefore, for any $t > 0$, the following equation

holds:

$$p\left(\boldsymbol{\theta}_{0:t}, \mathbf{y}_{1:t}\right) = p\left(\boldsymbol{\theta}_0\right) \prod_{j=1}^{t} p\left(\boldsymbol{\theta}_j | \boldsymbol{\theta}_{j-1}\right) p\left(\mathbf{y}_j | \boldsymbol{\theta}_j\right). \tag{9.1}$$

Any other distribution of interest can be derived from equation (9.1). The information flow of the SSM is illustrated in figure 9.1. The figure illustrates the Markov property of the $\boldsymbol{\theta}_t$'s and that for any $t$, $\mathbf{Y}_t$ depends only on $\boldsymbol{\theta}_t$. [40, chapter 2], [41, chapter 4].

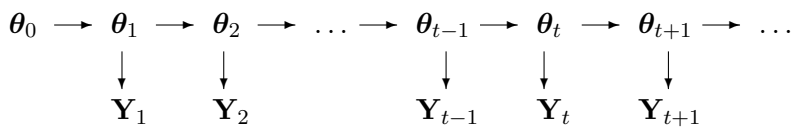



**Figure 9.1:** The information flow of a SSM

## 9.1 Filtering

For a SSM the challenge is to make inference on the unobserved states or to predict future observations based on the observation sequence. This is done by computing the densities of interest given the available information.
Estimation of the state vector $\boldsymbol{\theta}$ is done by computing the density $p\left(\boldsymbol{\theta}_t | \mathbf{y}_{1:s}\right)$, where $t$ is the time of interest. If $t = s$, i.e. if all observations up to the time of interest are available, the process is called filtering and can be performed using the following theorem.

**Theorem 9.2** (Filtering Recursions)
For a general SSM, the following statements hold.

**a)** The one-step-ahead predictive density for the states can be computed from the filtering density $p(\boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1})$ as

$$p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) = \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})\, p(\boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1}) d\boldsymbol{\theta}_{t-1}.$$

**b)** The one-step-ahead predictive density for the observations can be computed from the predictive density for the states as

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{y}_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) d\boldsymbol{\theta}_t.$$

**c)** The filtering density can be computed from the predictive densities for the states and observations as

$$p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) = \frac{p(\mathbf{y}_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})}.$$

**Proof**. **a)** Given $\boldsymbol{\theta}_{t-1}$, $\boldsymbol{\theta}_t$ is independent of $\mathbf{Y}_{1:t-1}$, according to the definition of a SSM, definition 9.1. Therefore,

$$\begin{aligned} p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) &= \int p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1}) d\boldsymbol{\theta}_{t-1} \\ &= \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \mathbf{y}_{1:t-1})\, p(\boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1}) d\boldsymbol{\theta}_{t-1} \\ &= \int p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})\, p(\boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1}) d\boldsymbol{\theta}_{t-1}. \end{aligned}$$

**b)** Given $\boldsymbol{\theta}_t$, $\mathbf{Y}_t$ is independent of $\mathbf{Y}_{1:t-1}$. Therefore,

$$\begin{aligned} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) &= \int p(\mathbf{y}_t, \boldsymbol{\theta}_t|\, \mathbf{y}_{1:t-1}) d\boldsymbol{\theta}_t \\ &= \int p(\mathbf{y}_t|\boldsymbol{\theta}_t, \mathbf{y}_{1:t-1})\, p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) d\boldsymbol{\theta}_t \\ &= \int p(\mathbf{y}_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) d\boldsymbol{\theta}_t. \end{aligned}$$

**c)** Given $\boldsymbol{\theta}_t$, $\mathbf{Y}_t$ and $\mathbf{Y}_{1:t-1}$ are conditionally independent. There-

fore, using Bayes' formula:

$$\begin{aligned} p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) &= \frac{p(\mathbf{y}_{1:t-1}, \mathbf{y}_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t)}{p(\mathbf{y}_{1:t-1}, \mathbf{y}_t)} \\ &= \frac{p(\mathbf{y}_t|\boldsymbol{\theta}_t, \mathbf{y}_{1:t-1})\, p(\mathbf{y}_{1:t-1}|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t)}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})\, p(\mathbf{y}_{1:t-1})} \\ &= \frac{p(\mathbf{y}_t|\boldsymbol{\theta}_t, \mathbf{y}_{1:t-1})\, p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})\, p(\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})\, p(\mathbf{y}_{1:t-1})} \\ &= \frac{p(\mathbf{y}_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})}. \end{aligned}$$

□

[40, chapter 2], [41, chapter 4].

## 9.2 Smoothing

When $t < s$, calculation of the density $p(\boldsymbol{\theta}_t|\mathbf{y}_{1:s})$ and estimation of the state vector is referred to as smoothing. It is quite similar to filtering, but is used when e.g. all observations are available up to the time of interest and beyond. We let $n$ denote the number of observations used in the smoothing process.

**Theorem 9.3** (Smoothing Recursions)
For a general SSM, the following statements hold.

**a)** Conditional on $\mathbf{y}_{1:n}$, the state sequence $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ has backward transition probabilities given by

$$p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1}, \mathbf{y}_{1:n}) = \frac{p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t})}{p(\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:t})},$$

where $t < n$.

**b)** Conditional on $\mathbf{y}_{1:n}$, the smoothing distributions of $\boldsymbol{\theta}_t$ can be computed according to the following backward recursion in $t$, starting from $p(\boldsymbol{\theta}_n|\mathbf{y}_{1:n})$:

$$p(\boldsymbol{\theta}_t|\mathbf{y}_{1:n}) = p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) \int \frac{p(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)}{p(\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:t})} p(\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:n}) d\boldsymbol{\theta}_{t+1}.$$

**Proof**. **a)** By definition of a SSM, definition 9.1, $\boldsymbol{\theta}_t$ and $\mathbf{Y}_{t+1:T}$ are conditionally independent given $\boldsymbol{\theta}_{t+1}$. Furthermore, $\boldsymbol{\theta}_{t+1}$ and $\mathbf{Y}_{1:t}$ are conditionally independent given $\boldsymbol{\theta}_t$. Therefore, using Bayes' formula,

$$\begin{aligned} p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathbf{y}_{1:T}\right) &= p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathbf{y}_{1:t}\right) \\ &= \frac{p\left(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}\right)p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t,\mathbf{y}_{1:t}\right)}{p\left(\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:t}\right)} \\ &= \frac{p\left(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}\right)p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\right)}{p\left(\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:t}\right)}. \end{aligned}$$

**b)** Using the result of **a)**, we obtain

$$\begin{aligned} p\left(\boldsymbol{\theta}_t|\mathbf{y}_{1:T}\right) &= \int p\left(\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:T}\right)d\boldsymbol{\theta}_{t+1} \\ &= \int p\left(\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:T}\right)p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathbf{y}_{1:T}\right)d\boldsymbol{\theta}_{t+1} \\ &= \int p\left(\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:T}\right)\frac{p\left(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}\right)p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\right)}{p\left(\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:t}\right)}d\boldsymbol{\theta}_{t+1} \\ &= p\left(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}\right)\int \frac{p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\right)}{p\left(\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:t}\right)}p\left(\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:T}\right)d\boldsymbol{\theta}_{t+1}. \end{aligned}$$

□

[40, chapter 2], [41, chapter 4].

## 9.3 Forecasting

When $t > s$, state estimation and the calculation of the density $p\left(\boldsymbol{\theta}_t|\mathbf{y}_{1:s}\right)$ is referred to as state prediction or forecasting.
After observing $\mathbf{y}_{1:t}$, prediction of the latent process and future observations may be of interest. This requires that the distributions of $\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}$ and $\mathbf{y}_{t+k}|\mathbf{y}_{1:t}$ are determined for the $k$ of interest.

**Theorem 9.4** (Forecasting Recursions)
For a general SSM, the following statements hold for any $k > 0$.

**a)** The $k$-steps ahead forecast distribution of the state is

$$p\left(\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}\right) = \int p\left(\boldsymbol{\theta}_{t+k}|\boldsymbol{\theta}_{t+k-1}\right) p\left(\boldsymbol{\theta}_{t+k-1}|\mathbf{y}_{1:t}\right) d\boldsymbol{\theta}_{t+k-1}.$$

**b)** The $k$-steps ahead forecast distribution of the observation is

$$p\left(\mathbf{y}_{t+k}|\mathbf{y}_{1:t}\right) = \int p\left(\mathbf{y}_{t+k}|\boldsymbol{\theta}_{t+k}\right) p\left(\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}\right) d\boldsymbol{\theta}_{t+k}.$$

**Proof**. **a)** Using the conditional independence property of $\boldsymbol{\theta}_{t+k}$ and $\mathbf{y}_{1:t}$ given $\boldsymbol{\theta}_{t+k-1}$, pertaining to a SSM, definition 9.1, we obtain

$$\begin{aligned} p\left(\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}\right) &= \int p\left(\boldsymbol{\theta}_{t+k}, \boldsymbol{\theta}_{t+k-1}|\mathbf{y}_{1:t}\right) d\boldsymbol{\theta}_{t+k-1} \\ &= \int p\left(\boldsymbol{\theta}_{t+k}|\boldsymbol{\theta}_{t+k-1}, \mathbf{y}_{1:t}\right) p\left(\boldsymbol{\theta}_{t+k-1}|\mathbf{y}_{1:t}\right) d\boldsymbol{\theta}_{t+k-1} \\ &= \int p\left(\boldsymbol{\theta}_{t+k}|\boldsymbol{\theta}_{t+k-1}\right) p\left(\boldsymbol{\theta}_{t+k-1}|\mathbf{y}_{1:t}\right) d\boldsymbol{\theta}_{t+k-1}. \end{aligned}$$

**b)** Similarly,

$$\begin{aligned} p\left(\mathbf{y}_{t+k}|\mathbf{y}_{1:t}\right) &= \int p\left(\mathbf{y}_{t+k}, \boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}\right) d\boldsymbol{\theta}_{t+k} \\ &= \int p\left(\mathbf{y}_{t+k}|\boldsymbol{\theta}_{t+k}, \mathbf{y}_{1:t}\right) p\left(\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}\right) d\boldsymbol{\theta}_{t+k} \\ &= \int p\left(\mathbf{y}_{t+k}|\boldsymbol{\theta}_{t+k}\right) p\left(\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}\right) d\boldsymbol{\theta}_{t+k}, \end{aligned}$$

using the conditional independence property of $\mathbf{y}_{t+k}$ and $\mathbf{y}_{1:t}$ given $\boldsymbol{\theta}_{t+k}$.

□

[40, chapter 2], [41, chapter 4].

# Dynamic Linear Models

In this chapter, the important class of Gaussian linear SSMs, also called dynamic linear models (DLMs), are defined. Furthermore, the Kalman filtering, smoothing and forecasting processes are derived. These are special cases of the processes derived in chapter 9, utilizing the linearity and normality of the DLM.

First a DLM is defined

**Definition 10.1** (Dynamic Linear Model)
Let $\mathbf{Y}_t \in \mathbb{R}^m$ denote an observation vector and $\boldsymbol{\theta}_t \in \mathbb{R}^p$ denote a state vector.
A DLM is characterized by an initial Normal prior distribution for the parameter vector,

$$\boldsymbol{\theta}_0 \sim \mathrm{N}_p\left(\mathbf{m}_0, C_0\right), \tag{10.1}$$

where $\mathbf{m}_0$ and $C_0$ is the mean and variance, respectively, and a dynamic set of four matrices $\{F_t, G_t, V_t, W_t\}$, that for each time $t \geq 1$ are known matrices of appropriate dimensions.
The set $\{F_t, G_t, V_t, W_t\}$ defines the model relating the observation vector $\mathbf{Y}_t$ to the state vector $\boldsymbol{\theta}_t$ at time $t$, and the $\boldsymbol{\theta}_t$ sequence through time by satisfying the equations

$$\begin{aligned} \mathbf{Y}_t &= F_t\boldsymbol{\theta}_t + \mathbf{v}_t, & \mathbf{v}_t &\sim \mathrm{N}_m\left(\mathbf{0}, V_t\right), & (10.2)\\ \boldsymbol{\theta}_t &= G_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t, & \mathbf{w}_t &\sim \mathrm{N}_p\left(\mathbf{0}, W_t\right). & (10.3) \end{aligned}$$

Furthermore, it is assumed that $\boldsymbol{\theta}_0$ is independent of both $(\mathbf{v}_t)$ and $(\mathbf{w}_t)$, which are independent noise sequences.

Equation (10.2) is called the observation equation for the model, and equation (10.3) is the state equation.
From equation (10.3) it is easy to see that, given the known matrices $G_t$ and $W_t$, $\boldsymbol{\theta}_t$ depends only on the previous state $\boldsymbol{\theta}_{t-1}$ and not on earlier information, i.e. $(\boldsymbol{\theta}_t)$ is a Markov chain. From equation (10.2) it is clear that, conditionally on $(\boldsymbol{\theta}_t)$, the $\mathbf{Y}_t$'s are independent and $\mathbf{Y}_t$ depends on $\boldsymbol{\theta}_t$ only. In other words, a DLM

satisfies the two properties of a SSM, stated in definition 9.1, with $\mathbf{Y}_t|\boldsymbol{\theta}_t \sim \mathrm{N}_m(F_t\boldsymbol{\theta}_t, V_t)$ and $\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1} \sim \mathrm{N}_p(G_t\boldsymbol{\theta}_{t-1}, W_t)$. The DLM is completely specified by these conditional densities combined with the initial prior distribution of equation (10.1), as mentioned in chapter 9.

If the matrices $F_t$ and $G_t$ are constant for all $t$, the model is referred to as a time series DLM (TSDLM). A TSDLM with constant variance matrices $V_t$ and $W_t$ for all $t$ is called a constant DLM. Thus a constant DLM is characterized by a single set of matrices $\{F, G, V, W\}$ for all times $t$. This special case of DLMs includes essentially all classical linear time series models.

## 10.1 Kalman Filtering

The filtering for a general SSM is done by use of theorem 9.2. For DLMs the computations needed for filtering are considerably simplified, as stated in the next theorem. The conditional distributions are derived by induction in time, where the initial step at time $t = 0$ is specified by equation (10.1).

**Theorem 10.2** (Kalman Filter)
Consider a DLM. Let

$$\boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1} \sim \mathrm{N}_p\left(\mathbf{m}_{t-1}, C_{t-1}\right). \tag{10.4}$$

Then the following statements hold.

**a)** The one-step-ahead predictive distribution of $\boldsymbol{\theta}_t$ given $\mathbf{y}_{1:t-1}$ is Gaussian with mean and variance

$$\begin{aligned} \mathbf{a}_t &= G_t \mathbf{m}_{t-1}, \\ R_t &= G_t C_{t-1} G_t^\top + W_t. \end{aligned}$$

**b)** The one-step-ahead predictive distribution of $\mathbf{Y}_t$ given $\mathbf{y}_{1:t-1}$ is Gaussian with mean and variance

$$\begin{aligned} \mathbf{f}_t &= F_t \mathbf{a}_t, \\ Q_t &= F_t R_t F_t^\top + V_t. \end{aligned}$$

**c)** The filtering distribution of $\boldsymbol{\theta}_t$ given $\mathbf{y}_{1:t}$ is Gaussian with mean and variance

$$\begin{aligned} \mathbf{m}_t &= \mathbf{a}_t + R_t F_t^\top Q_t^{-1} \mathbf{e}_t, \\ C_t &= R_t - R_t F_t^\top Q_t^{-1} F_t R_t, \end{aligned}$$

where $\mathbf{e}_t = \mathbf{y}_t - \mathbf{f}_t$ is the forecast error.

**Proof**. The theorem is proved by induction using the theory of the multivariate normal distribution, see [40, appendix A, p. 233].

**a)** The state equation (10.3) and the assumption of equation (10.4) yield

$$\begin{aligned} \boldsymbol{\theta}_t|\mathbf{y}_{1:t-1} &= G_t \cdot (\boldsymbol{\theta}_{t-1}|\mathbf{y}_{1:t-1}) + \mathbf{w}_t \\ &\sim \mathrm{N}_p\left(G_t \mathbf{m}_{t-1}, G_t C_{t-1} G_t^\top + W_t\right), \end{aligned}$$

using the rules for multiplying a normal distribution by a matrix and adding two independent normal distributions. Using the notation defined in the theorem statement, we obtain $\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1} \sim \mathrm{N}_p\left(\mathbf{a}_t, R_t\right)$.

**b)** Similarly, using the observation equation (10.2) and **a)**,

$$\begin{aligned}\mathbf{Y}_t|\mathbf{y}_{1:t-1} &= F_t \cdot (\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) + \mathbf{v}_t \\ &\sim \mathrm{N}_m\left(F_t\mathbf{a}_t, F_tR_tF_t^\top + V_t\right).\end{aligned}$$

Using the notation defined in the theorem statement, $\mathbf{Y}_t|\mathbf{y}_{1:t-1} \sim \mathcal{N}_m(\mathbf{f}_t, Q_t)$.

**c)** From Theorem 9.2 **c)** it is known, that

$$\begin{aligned}p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}) &= \frac{p(\mathbf{y}_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})} \\ &\propto p(\mathbf{y}_t|\boldsymbol{\theta}_t)\, p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}), \qquad (10.5)\end{aligned}$$

where $p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1})$ is already known from **a)**:

$$p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}) \propto \exp\left(-\frac{(\boldsymbol{\theta}_t - \mathbf{a}_t)^\top R_t^{-1}(\boldsymbol{\theta}_t - \mathbf{a}_t)}{2}\right).$$

The first density is found using the observation equation (10.2):

$$\begin{aligned}\mathbf{Y}_t &= F_t\boldsymbol{\theta}_t + \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathrm{N}_m(\mathbf{0}, V_t) \\ &\Downarrow \\ p(\mathbf{y}_t|\boldsymbol{\theta}_t) &\propto \exp\left(-\frac{(\mathbf{y}_t - F_t\boldsymbol{\theta}_t)^\top V_t^{-1}(\mathbf{y}_t - F_t\boldsymbol{\theta}_t)}{2}\right).\end{aligned}$$

Taking the logarithm and multiplying equation (10.5) by $-2$ yield

$$\begin{aligned}-2\log(p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t})) &= (\boldsymbol{\theta}_t - \mathbf{a}_t)^\top R_t^{-1}(\boldsymbol{\theta}_t - \mathbf{a}_t) \qquad (10.6) \\ &+ (\mathbf{y}_t - F_t\boldsymbol{\theta}_t)^\top V_t^{-1}(\mathbf{y}_t - F_t\boldsymbol{\theta}_t) + c_1,\end{aligned}$$

where $c_1$ is a constant not depending on $\boldsymbol{\theta}_t$. Rearranging equation (10.6) yields

$$\begin{aligned}-2\log(p(\boldsymbol{\theta}_t|\mathbf{y}_{1:t})) &= \boldsymbol{\theta}_t^\top\left(R_t^{-1} + F_t^\top V_t^{-1}F_t\right)\boldsymbol{\theta}_t \qquad (10.7) \\ &- 2\boldsymbol{\theta}_t^\top\left(R_t^{-1}\mathbf{a}_t + F_t^\top V_t^{-1}\mathbf{y}_t\right) + c_2,\end{aligned}$$

where $c_2$ is a new constant not depending on $\boldsymbol{\theta}_t$.
Using $C_t$ defined in the theorem statement,

$$\left(R_t^{-1} + F_t^\top V_t^{-1}F_t\right)C_t = I,$$

where $I$ is the $p \times p$ identity matrix. Thus

$$R_t^{-1} + F_t^\top V_t^{-1} F_t = C_t^{-1}.$$

Using $\mathbf{m}_t$ defined in the theorem statement,

$$C_t^{-1}\mathbf{m}_t = R_t^{-1}\mathbf{a}_t + F_t^\top V_t^{-1}\mathbf{y}_t.$$

Therefore, equation (10.7) becomes

$$\begin{aligned} -2\ln\left(p\left(\boldsymbol{\theta}_t|\mathbf{y}_t\right)\right) &= \boldsymbol{\theta}_t^\top C_t^{-1}\boldsymbol{\theta}_t - 2\boldsymbol{\theta}_t^\top C_t^{-1}\mathbf{m}_t + c_3 \\ &= \left(\boldsymbol{\theta}_t - \mathbf{m}_t\right)^\top C_t^{-1}\left(\boldsymbol{\theta}_t - \mathbf{m}_t\right) + c_4, \end{aligned}$$

where $c_3$ and $c_4$ are constants. Hence,

$$p\left(\boldsymbol{\theta}_t|\mathbf{y}_t\right) \propto \exp\left(-\frac{\left(\boldsymbol{\theta}_t - \mathbf{m}_t\right)^\top C_t^{-1}\left(\boldsymbol{\theta}_t - \mathbf{m}_t\right)}{2}\right),$$

i.e. $\boldsymbol{\theta}_t|\mathbf{y}_t \sim \mathrm{N}_p\left(\mathbf{m}_t, C_t\right)$.

□

[40, section 2.7.2]

## 10.2 Kalman Smoothing

For a DLM the smoothing recursions can be expressed more specifically, as stated in the next theorem, which is a special case of theorem 9.3.

**Theorem 10.3** (Kalman Smoother)
Consider a DLM. Let $\boldsymbol{\theta}_{t+1}|\mathbf{y}_{1:n} \sim \mathrm{N}_p\left(\mathbf{s}_{t+1}, S_{t+1}\right)$ for $t < n$. Then $\boldsymbol{\theta}_t|\mathbf{y}_{1:n} \sim \mathrm{N}_p\left(\mathbf{s}_t, S_t\right)$, where

$$\mathbf{s}_t = \mathbf{m}_t + C_t G_{t+1}^\top R_{t+1}^{-1}\left(\mathbf{s}_{t+1} - \mathbf{a}_{t+1}\right), \tag{10.8}$$

$$S_t = C_t - C_t G_{t+1}^\top R_{t+1}^{-1}\left(R_{t+1} - S_{t+1}\right) R_{t+1}^{-1} G_{t+1} C_t. \tag{10.9}$$

**Proof**. By the definition of a DLM, the distributions of $\boldsymbol{\theta}_{1:n}$ and $\mathbf{y}_{1:n}$ are Gaussian. It follows from the properties of the multivariate normal distribution, see [40, appendix A, p. 233], that $\boldsymbol{\theta}_t$ given $\mathbf{y}_{1:n}$

is also Gaussian. Therefore it suffices to compute the mean and variance of this distribution, which are

$$\begin{aligned}\mathbf{s}_t &= \mathbb{E}\left[\boldsymbol{\theta}_t|\mathbf{y}_{1:n}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathbf{y}_{1:n}\right]|\mathbf{y}_{1:n}\right],\end{aligned}$$

and

$$\begin{aligned}S_t &= \mathrm{Var}\left[\boldsymbol{\theta}_t|\mathbf{y}_{1:n}\right] \\ &= \mathrm{Var}\left[\mathbb{E}\left[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathbf{y}_{1:n}\right]|\mathbf{y}_{1:n}\right] + \mathbb{E}\left[\mathrm{Var}\left[\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathbf{y}_{1:n}\right]|\mathbf{y}_{1:n}\right].\end{aligned}$$

As in the proof of theorem 9.3, $\boldsymbol{\theta}_t$ and $\mathbf{y}_{t+1:n}$ are conditionally independent given $\boldsymbol{\theta}_{t+1}$. Thus, $p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathbf{y}_{1:n}\right) = p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathbf{y}_{1:t}\right)$. This posterior density can be calculated using Bayes' formula. Applying the conditional independence of $\boldsymbol{\theta}_{t+1}$ and $\mathbf{y}_{1:t}$ given $\boldsymbol{\theta}_t$, the likelihood is $p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t,\mathbf{y}_{1:t}\right) = p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\right)$, which can be derived from the state equation (10.3) as

$$\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t \sim \mathrm{N}_p\left(G_{t+1}\boldsymbol{\theta}_t, W_{t+1}\right).$$

As for the prior, $\boldsymbol{\theta}_t|\mathbf{y}_{1:t} \sim \mathrm{N}_p\left(\mathbf{m}_t, C_t\right)$ according to theorem 10.2 **c)**. Hence, the posterior is

$$\begin{aligned}p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathbf{y}_{1:t}\right) &\propto p\left(\boldsymbol{\theta}_t|\mathbf{y}_{1:t}\right) p\left(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}_t - \mathbf{m}_t\right)^\top C_t^{-1}\left(\boldsymbol{\theta}_t - \mathbf{m}_t\right)\right) \\ &\quad \cdot \exp\left(-\frac{1}{2}\left(\boldsymbol{\theta}_{t+1} - G_{t+1}\boldsymbol{\theta}_t\right)^\top W_{t+1}\right. \\ &\quad \left.\cdot\left(\boldsymbol{\theta}_{t+1} - G_{t+1}\boldsymbol{\theta}_t\right)\right) \\ &\Downarrow \\ -2\log\left(p\left(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1},\mathbf{y}_{1:t}\right)\right) &= \left(\boldsymbol{\theta}_t - \mathbf{m}_t\right)^\top C_t^{-1}\left(\boldsymbol{\theta}_t - \mathbf{m}_t\right) \\ &\quad + \left(\boldsymbol{\theta}_{t+1} - G_{t+1}\boldsymbol{\theta}_t\right)^\top W_{t+1} \\ &\quad \cdot\left(\boldsymbol{\theta}_{t+1} - G_{t+1}\boldsymbol{\theta}_t\right) + c_1 \\ &= \boldsymbol{\theta}_t^\top\left(C_t^{-1} + G_{t+1}W_{t+1}^{-1}G_{t+1}\right)\boldsymbol{\theta}_t \\ &\quad -2\boldsymbol{\theta}_t^\top\left(C_t^{-1}\mathbf{m}_t + G_{t+1}^\top W_{t+1}^{-1}G_{t+1}\boldsymbol{\theta}_{t+1}\right) \\ &\quad +c_2,\end{aligned}$$

where $c_1$ and $c_2$ are constants not depending on $\boldsymbol{\theta}_t$. Furthermore,

$$\left(C_t^{-1} + G_{t+1}W_{t+1}^{-1}G_{t+1}\right)\left(C_t - C_tG_{t+1}^\top R_{t+1}^{-1}G_{t+1}C_t\right) = I,$$

so

$$C_t^{-1} + G_{t+1} W_{t+1}^{-1} G_{t+1} = \left( C_t - C_t G_{t+1}^\top R_{t+1}^{-1} G_{t+1} C_t \right)^{-1}.$$

Furthermore,

$$\begin{aligned} &\left( C_t - C_t G_{t+1}^\top R_{t+1}^{-1} G_{t+1} C_t \right)^{-1} \left( \mathbf{m}_t + C_t G_{t+1}^\top R_{t+1}^{-1} \left( \boldsymbol{\theta}_{t+1} - \mathbf{a}_{t+1} \right) \right) \\ &= C_t^{-1} \mathbf{m}_t + G_{t+1}^\top W_{t+1}^{-1} G_{t+1} \boldsymbol{\theta}_{t+1}. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[ \boldsymbol{\theta}_t | \boldsymbol{\theta}_{t+1}, \mathbf{y}_{1:t} \right] &= \mathbf{m}_t + C_t G_{t+1}^\top R_{t+1}^{-1} \left( \boldsymbol{\theta}_{t+1} - \mathbf{a}_{t+1} \right), \\ \mathrm{Var}\left[ \boldsymbol{\theta}_t | \boldsymbol{\theta}_{t+1}, \mathbf{y}_{1:t} \right] &= C_t - C_t G_{t+1}^\top R_{t+1}^{-1} G_{t+1} C_t. \end{aligned}$$

From this, it follows that

$$\begin{aligned} s_t &= \mathbb{E}\left[ \mathbb{E}\left[ \boldsymbol{\theta}_t | \boldsymbol{\theta}_{t+1}, \mathbf{y}_{1:n} \right] | \mathbf{y}_{1:n} \right] \\ &= \mathbf{m}_t + C_t G_{t+1}^\top R_{t+1}^{-1} \left( \mathbf{s}_{t+1} - \mathbf{a}_{t+1} \right), \end{aligned}$$

and

$$\begin{aligned} S_t &= \mathrm{Var}\left[ \mathbb{E}\left[ \boldsymbol{\theta}_t | \boldsymbol{\theta}_{t+1}, \mathbf{y}_{1:n} \right] | \mathbf{y}_{1:n} \right] + \mathbb{E}\left[ \mathrm{Var}\left[ \boldsymbol{\theta}_t | \boldsymbol{\theta}_{t+1}, \mathbf{y}_{1:n} \right] | \mathbf{y}_{1:n} \right] \\ &= C_t - C_t G_{t+1}^\top R_{t+1}^{-1} G_{t+1} C_t + C_t G_{t+1}^\top R_{t+1}^{-1} S_{t+1} R_{t+1}^{-1} G_{t+1} C_t \\ &= C_t - C_t G_{t+1}^\top R_{t+1}^{-1} \left( R_{t+1} - S_{t+1} \right) R^{-1} G_{t+1} C_t, \end{aligned}$$

where $\mathbb{E}\left[ \boldsymbol{\theta}_{t+1} | \mathbf{y}_{1:n} \right] = \mathbf{s}_{t+1}$ and $\mathrm{Var}\left[ \boldsymbol{\theta}_{t+1} | \mathbf{y}_{1:n} \right] = S_{t+1}$, by assumption. □

Kalman smoothing is performed after Kalman filtering and runs backwards in time using the recursions (10.8) and (10.9). The starting point for the algorithm is $\boldsymbol{\theta}_n | \mathbf{y}_{1:n} \sim \mathrm{N}_p\left( \mathbf{m}_n, C_n \right)$, obtained using the Kalman filter.

## 10.3 Kalman Forecasting

The next theorem concerning forecasting for a DLM is a special case of theorem 9.4.

**Theorem 10.4** (Kalman Forecaster)
For a DLM, let $\mathbf{a}_{t,0} = \mathbf{m}_t$ and $R_{t,0} = C_t$. For any $k \geq 1$, the following statements hold.

**a)** The distribution of $\boldsymbol{\theta}_{t+k}$ given $\mathbf{y}_{1:t}$ is Gaussian with mean and covariance

$$\begin{aligned} \mathbf{a}_{t,k} &= G_{t+k}\mathbf{a}_{t,k-1}, \\ R_{t,k} &= G_{t+k}R_{t,k-1}G_{t+k}^{\top} + W_{t+k}. \end{aligned}$$

**b)** The distribution of $\mathbf{Y}_{t+k}$ given $\mathbf{y}_{1:t}$ is Gaussian with mean and covariance

$$\begin{aligned} \mathbf{f}_{t,k} &= F_{t+k}\mathbf{a}_{t,k}, \\ Q_{t,k} &= F_{t+k}R_{t,k}F_{t+k}^{\top} + V_t. \end{aligned}$$

**Proof**. Similarly to the the proof for theorem 10.3, the forecast distributions are Gaussian by definition of a DLM, and it suffices to derive the parameters of the distributions. This is done by induction. Both statements hold for $k = 1$ according to theorem 10.2 **a)** and **b)**.

**a)** For $k > 1$, assume that the statement holds for $k - 1$. Then,

$$\begin{aligned} \mathbf{a}_{t,k} &= \mathbb{E}\left[\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}, \boldsymbol{\theta}_{t+k-1}\right]|\mathbf{y}_{1:t}\right] \\ &= \mathbb{E}\left[G_{t+k}\boldsymbol{\theta}_{t+k-1}|\mathbf{y}_{1:t}\right] \\ &= G_{t+k}\mathbf{a}_{t,k-1}, \end{aligned}$$

and

$$\begin{aligned} R_{t,k} &= \mathrm{Var}\left[\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}\right] \\ &= \mathrm{Var}\left[\mathbb{E}\left[\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}, \boldsymbol{\theta}_{t+k-1}\right]|\mathbf{y}_{1:t}\right] \\ &\quad + \mathbb{E}\left[\mathrm{Var}\left[\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}, \boldsymbol{\theta}_{t+k-1}\right]|\mathbf{y}_{1:t}\right] \\ &= G_{t+k}R_{t,k-1}G_{t+k}^{\top} + W_{t+k}. \end{aligned}$$

**b)** For $k > 1$, assume that the statement holds for $k-1$. Then, using

**a)**,

$$\begin{aligned}\mathbf{f}_{t,k} &= \mathbb{E}\left[\mathbf{Y}_{t+k}|\mathbf{y}_{1:t}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{Y}_{t+k}|\mathbf{y}_{1:t},\boldsymbol{\theta}_{t+k}\right]|\mathbf{y}_{1:t}\right] \\ &= \mathbb{E}\left[F_{t+k}\boldsymbol{\theta}_{t+k}|\mathbf{y}_{1:t}\right] \\ &= F_{t+k}\mathbf{a}_{t,k},\end{aligned}$$

and

$$\begin{aligned}Q_{t,k} &= \mathrm{Var}\left[\mathbf{Y}_{t+k}|\mathbf{y}_{1:t}\right] \\ &= \mathrm{Var}\left[\mathbb{E}\left[\mathbf{Y}_{t+k}|\mathbf{y}_{1:t},\boldsymbol{\theta}_{t+k}\right]|\mathbf{y}_{1:t}\right] \\ &\quad +\mathbb{E}\left[\mathrm{Var}\left[\mathbf{Y}_{t+k}|\mathbf{y}_{1:t},\boldsymbol{\theta}_{t+k}\right]|\mathbf{y}_{1:t}\right] \\ &= F_{t+k}R_{t,k}F_{t+k}^{\top}+V_{t+k}.\end{aligned}$$

□

[40, chapter 2], [41, chapter 4].

## 10.4 Model Specification

The Kalman filter and smoother provide methods for estimation and prediction when the DLM is completely specified, i.e. when all of the matrices $F_t$, $G_t$, $V_t$ and $W_t$ are known. However, that is rarely the case in practice. In fact, completely specifying a DLM can be very difficult.
A general approach is to decompose the time series into simple components that each capture a specific feature of the time series, e.g. seasonal component, trend or dependence on covariates (regression). Each of these components can be thought of as individual time series, each described by a DLM that, when added together, form the DLM that describes the complete time series.

Consider a time series $(\mathbf{Y}_t)$ and assume that it can be written as the sum of $h$ independent components, i.e.

$$\mathbf{Y}_t = \mathbf{Y}_{1,t} + \cdots + \mathbf{Y}_{h,t},$$

where each time series $(\mathbf{Y}_{i,t})$, $i = 1,\ldots,h$ represents a component of the model and is described by a DLM:

$$\begin{aligned}\mathbf{Y}_{i,t} &= F_{i,t}\boldsymbol{\theta}_{i,t}+\mathbf{v}_{i,t}, & \mathbf{v}_{i,t} &\sim \mathrm{N}_m\left(\mathbf{0},V_{i,t}\right), \\ \boldsymbol{\theta}_{i,t} &= G_{i,t}\boldsymbol{\theta}_{i,t-1}+\mathbf{w}_{i,t}, & \mathbf{w}_{i,t} &\sim \mathrm{N}p_i\left(\mathbf{0},W_{i,t}\right),\end{aligned}$$

where the $p_i$-dimensional state vectors $\boldsymbol{\theta}_{i,t}$ are distinct. We assume that the time series $(\mathbf{Y}_{i,t}, \boldsymbol{\theta}_{i,t})$ and $(\mathbf{Y}_{j,t}, \boldsymbol{\theta}_{j,t})$ are mutually independent for all $i \neq j$. By assumption of independence it follows that $\mathbf{Y}_t = \sum_{i=1}^{h} \mathbf{Y}_{i,t}$ is described by the DLM

$$\begin{aligned} \mathbf{Y}_t &= F_t \boldsymbol{\theta}_t + \mathbf{v}_t, & \mathbf{v}_t &\sim \mathrm{N}_m(\mathbf{0}, V_t), \\ \boldsymbol{\theta}_t &= G_t \boldsymbol{\theta}_{t-1} + \mathbf{w}_t, & \mathbf{w}_t &\sim \mathrm{N}_p(\mathbf{0}, W_t), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\theta}_t &= \begin{bmatrix} \boldsymbol{\theta}_{1,t} \\ \vdots \\ \boldsymbol{\theta}_{h,t} \end{bmatrix}, \\ F_t &= [F_{1,t} | \ldots | F_{h,t}], \\ V_t &= \sum_{i=1}^{h} V_{i,t}, \end{aligned}$$

and where $G_t$ and $W_t$ are the block diagonal matrices

$$G_t = \begin{bmatrix} G_{1,t} & & \\ & \ddots & \\ & & G_{h,t} \end{bmatrix}, \qquad W_t = \begin{bmatrix} W_{1,t} & & \\ & \ddots & \\ & & W_{h,t} \end{bmatrix}.$$

### 10.4.1 Regression Models

In our study we wish to find the dependence of the response on the covariates, i.e. regression. A linear regression model for univariate observations $Y_t$ on $p$ covariates $x_1, \ldots, x_p$ is described by

$$\begin{aligned} Y_t &= x_{1,t}\beta_1 + \cdots + x_{p,t}\beta_p + v_t \\ &= \sum_{j=1}^{p} x_{j,t}\beta_j + v_t, \end{aligned}$$

where $v_t \overset{iid}{\sim} \mathrm{N}\left(0, \sigma_t^2\right)$ per definition. If the observations are taken over time, the assumption of i.i.d. errors is often not very realistic. A solution to this problem is to assume that the relationship between $y$ and the $x_j$'s evolves over time, i.e. to assume a dynamic linear regression model

$$\begin{aligned} Y_t &= x_{1,t}\beta_{1,t} + \cdots + x_{p,t}\beta_{p,t} + v_t \\ &= \sum_{j=1}^{p} x_{j,t}\beta_{j,t} + v_t, \end{aligned}$$

and model the temporal evolution of $(\beta_{1,t}, \ldots, \beta_{j,t})$. This would yield a DLM with the parameters $F_t = [x_{1,t}, \ldots, x_{p,t}]$, $\boldsymbol{\theta}_t = [\beta_{1,t}, \ldots, \beta_{p,t}]^\top$ and $V_t = \sigma_t^2$.
The model is completed by a state equation. An often used default approach is to choose the evolution matrix $G_t$ to be the identity matrix and $W_t$ as a diagonal matrix, corresponding to modeling the regression coefficients as independent random walks. [40, chapter 3].

## 10.5 Parameter Estimation

In previous sections, the matrices $F_t$, $G_t$, $V_t$ and $W_t$ have been assumed known. This was done to more easily study their properties, but it is rarely the case in practice. In this section we assume the matrices depend on an unknown parameter vector $\boldsymbol{\psi}$ consisting of so-called hyper parameters $\psi_t$. These are often constant, but may also evolve over time.

The hyper parameters, $\psi_t$, can be estimated using different methods. The estimation can be of interest in itself or used for smoothing or forecasting using the methods discussed in previous sections of this chapter.

This section presents the method for estimating the hyper parameters used in our study, the direct MLE. [40, chapter 4].

### 10.5.1 Direct Maximum Likelihood Estimation

Suppose $n$ random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ to have distributions depending on an unknown parameter vector $\boldsymbol{\psi}$. The joint density of the observations for a given value of the parameter is denoted by $p(\mathbf{y}_1, \ldots, \mathbf{y}_n|\boldsymbol{\psi})$. The likelihood of the hyper parameters conditional on the observations is $L(\boldsymbol{\psi}|\mathbf{y}_1, \ldots, \mathbf{y}_n) = p(\mathbf{y}_1, \ldots, \mathbf{y}_n|\boldsymbol{\psi})$. The likelihood and loglikelihood are therefore

$$\begin{aligned} L(\boldsymbol{\psi}|\mathbf{y}_1, \ldots, \mathbf{y}_n) &= \prod_{t=1}^{n} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\psi}), \\ l(\boldsymbol{\psi}|\mathbf{y}_1, \ldots, \mathbf{y}_n) &= \sum_{t=1}^{n} \log\left(p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\psi})\right). \end{aligned} \tag{10.10}$$

From theorem 10.2 **b)** it is known, that the densities on the right hand side of equation (10.10) are $m$-dimensional Gaussian densities with means $\mathbf{f}_t$ and variances $Q_t$. Hence, the loglikelihood is

$$\begin{aligned} l\left(\boldsymbol{\psi}|\mathbf{y}_1,\ldots,\mathbf{y}_n\right) &= -\frac{mn}{2}\log\left(2\pi\right) - \frac{1}{2}\sum_{t=1}^{n}\log|Q_t| \\ &\quad -\frac{1}{2}\sum_{t=1}^{n}\left(\mathbf{y}_t-\mathbf{f}_t\right)^{\top}Q_t^{-1}\left(\mathbf{y}_t-\mathbf{f}_t\right), \quad (10.11) \end{aligned}$$

where $\mathbf{f}_t$ and $Q_t$ depend implicitly on $\boldsymbol{\psi}$.

In this way, for a given parameter vector $\boldsymbol{\psi}$, the loglikelihood can be obtained from the Kalman filter, theorem 10.2, and equation (10.11) can be maximized numerically w.r.t. $\boldsymbol{\psi}$ to obtain the MLE, denoted $\hat{\boldsymbol{\psi}}$. [40, section 4.1].

# Bibliography


[1] World Health Organization - Cardiovascular Diseases. `http://www.who.int/mediacentre/factsheets/fs317/en/`, visited 9.14.2011 at 17:00.

[2] N. K. Nissen and S. Rasmussen. *2008 Hjertestatistik - Fokus på Køn og Sociale Forskelle.* Hjerteforeningen, 2008.

[3] L. Frost, L. V. Andersen, L. S. Mortensen, and C. Dethlefsen. Seasonal Variation in Stroke and Stroke-Associated Mortality in Patients with a Hospital Diagnosis of Nonvalvular Atrial Fibrillation or Flutter. *Neuroepidemiology, vol. 26, pp. 220-225*, 2006.

[4] J. P. Ornato, M. A. Peberdy, N. C. Chandra, and D. E. Bush. Seasonal Pattern of Acute Myocardial Infarction in the National Registry of Myocardial Infarction. *Journal of the American College of Cardiology, vol. 28, pp. 684-1688*, 1965.

[5] A. L. Christensen. *Gradually Changing Seasonal Variation of Cardiovascular Diseases - A Danish Nationwide Cohort Study.* Project written at Department of Mathematical Sciences, Aalborg University in cooperation with Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital, 2009.

[6] A. L. Christensen. *Seasonal Variation of Cardiovascular Diseases - a Danish Nationwide Cohort Study.* Project written at Department of Mathematical Sciences, Aalborg University in cooperation with Center for Cardiovascular Research, Aalborg Hospital, Aarhus University Hospital, 2008.

[7] C. Ku, C. Yang, W. Lee, H. Chiang, C. Liu, and S. Lin. Absence of a Seasonal Variation in Myocardial Infarction Onset in a Region Without Temperature Extremes. *Cardiology, vol. 89 no. 4, pp. 277-282*, 1998.

[8] M. A. McGeehin and M. Mirabelli. The Potential Impacts of Climate Variability and Change on Temperature-Related

Morbidity and Mortality in the United States. *Environmental Health Perspectives, vol. 35, pp. 432-441*, 1974.

[9] L.S. Kalkstein and J.S. Greene. An Evaluation of Climate/Mortality Relationships in Large US Cities and the Possible Impacts of a Climate Change. *Environmental Health Perspectives, vol. 105, No. 1, pp. 84-93.*

[10] A.L.F. Braga, A. Zanobetti, and J. Schwartz. The Effect of Weather on Respiratory and Cardiovascular Deaths in 12 U.S. Cities. *Environmental Health Perspectives, vol. 110, No. 9, pp. 969-974*, 2002.

[11] J. Schwartz. The Distributed Lag Between Air Pollution and Daily Deaths. *Epidemiology, vol. 11, pp. 320-326*, 2000.

[12] S. Almon. The Distributed Lag Between Capital Appropriations and Expenditures. *Econometrica, vol. 33, pp. 178-196*, 1965.

[13] A.L.F. Braga, A. Zanobetti, and J. Schwartz. The Time Course of Weather Related Deaths: A Tridimensional Estimate in 12 US Cities. *Epidemiology, vol. 12, pp. 662-667*, 2001.

[14] D. A. Botter, B. Jørgensen, and A.A.Q. Peres. A Longitudinal Study of Mortality and Air Pollution for São Paulo, Brazil. *Journal of Exposure Analysis and Environmental Epidemiology, vol. 12, pp. 335-343*, 2002.

[15] Sundhedstyrelsen. `http://www.sst.dk/Indberetning%20og%20statistik/Landspatientregisteret.aspx`, visited 9.28.2011 at 09:34.

[16] WHO. `http://www.who.int/classifications/icd/en/`, visited 9.28.2011 at 09:43.

[17] T. F. Andersen, M. Madsen, J. Jørgensen, L. Mellemkjær, and J.H. Olsen. The Danish National Hospital Register - A Valuable Source of Data for Modern Health Sciences. *Danish Medical Bulletin, vol. 46, No. 3, pp. 263-268*, 1999.

[18] Det Centrale Person Register. `http://www.cpr.dk/cpr/site.aspx?p=154&ArticleID=4118`, visited 9.28.2011 at 09:39.

[19] DMI.
http://www.dmi.dk/dmi/index/om_dmi/i_korte_traek.htm,
visited 1.13.2012 at 10:04.

[20] F.H. Martini. *Fundamentals of Anatomy & Physiology - seventh edition.* Pearson Education - Benjamin Cummings, 2006.

[21] L.B. Mitchell. Merck Manuals - Atrial Fibrillation.
http://www.merckmanuals.com/professional/
cardiovascular_disorders/arrhythmias_and\
_conduction_disorders/atrial_fibrillation_af.html,
visited 9.22.2011 at 21:00.

[22] B. Alsbjørn, A. Bertelsen, O. Bonnevie, P. Bretlau, K. Brøndum-Nielsen, I.C. Bygbjerg, E. Dickmeiss, A. Dirksen, P. Fleckenstein, H. Fledelius, E. Gutschik, F. Gyntelberg, M.M. Hansen, P. Hertoft, J. Hilden, T. Horn, H. Høyer, M. Juhler, J.P. Kampmann, H. Kirk, F.U. Knudsen, A. Krasnik, S.U. Larsen, B. Lund, J.O. Lund, S. Madsbad, E. Magid, T. Morgensen, M. Norup, P.S. Olsen, B. Ottesen, M. Rørth, T.V. Schrøder, S. Schulze, I. Sewerin, P. Skinhøj, K. Stengaard-Pedersen, S. Strandgaard, K. Thestrup-Pedersen, J.L. Thomsen, J. Tranum-Jensen, S. Vorstrup, and S. Walter. *Klinisk Ordbog.* Munksgaard Denmark, 2004.

[23] A.L. Waldo. The Interrelationship Between Atrial Fibrillation and Atrial Flutter. *Progress in Cardiovascular Diseases*, 2005.

[24] A.M. Joensen, M K. Jensen, K. Overvad, C. Dethlefsen, E. Schmidt, L. Rasmussen, A. Tjønneland, and S. Johnsen. Predictive Values of Acute Coronary Syndrome Discharge Diagnoses Differed in the Danish National Patient Registry. *Journal of Clinical Epidemiology, vol. 62, pp. 188-194*, 2009.

[25] S. P. Johnsen, K. Overvad, H.T. Sørensen, A. Tjønneland, and S. E. Husted. Predictive Value of Stroke and Transient Ischemic Attack Discharge Diagnoses in the Danish National Registry of Patients. *Journal of Clinical Epidemiology, vol. 55, pp. 602-607*, 2002.

[26] M.T. Severinsen, S. Kristensen, K. Overvad, C. Dethlefsen, A. Tjønneland, and S.P. Johnsen. Venous Thromboembolism Discharge Diagnoses in the Danish National Patient Registry

Should be Used with Caution. *Journal of Clinical Epidemiology, vol. 63, pp. 223-228*, 2008.

[27] T.A. Rix, S. Riahi, K. Overvad, S. Lundbye-Christensen, E.B. Schmidt, and A.M. Joensen. Validity of the Diagnoses Atrial Fibrillation and Atrial Flutter in a Danish Patient Registry. *In preparation*, 2011.

[28] A. Tjønneland, K. Boll A. Olsen, C. Stripp, J. Christensen, G. Engholm, and K. Overvad. Study Design, Exposure Variables, and Socioeconomic Determinants of Participation in Diet, Cancer and Health: A Population-Based Prospective Cohort Study of 57,053 Men and Women in Denmark. *Scandinavian Journal of Public Health, vol. 35, pp. 432-441*, 2007.

[29] Statistics Denmark. `http://www.statistikbanken.dk`, visited 1.12.2012 at 16:45.

[30] K. J. Rothman. *Epidemiology - An introduction.* Oxford University press, 2002.

[31] K. J. Rothman, S. Greenland, and T. L. Lash. *Modern Epidemiology.* Lippincott Williams and Wilkins, 2008.

[32] H. Hotelling. Contribution to Discussion on Paper. *Journal of the Royal Statistical Society, Series B, pp. 229-230*, 1953.

[33] P. McCullagh and J.A. Nelder. *Generalized Linear Models.* Chapman and Hall, 1989.

[34] A.J. Dobson and A.G. Barnett. *An Introduction to Generalized Linear Models.* Chapman and Hall/CRC, 2008.

[35] A. Azzalini. *Statistical Inference - Based on the likelihood.* Chapman and Hall/CRC, 2002.

[36] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, 1990.

[37] T. Hastie and R. Tibshirani. Generalised Additive Models. *Statistical Science, vol. 1, No. 3, pp. 297-318*, 1986.

[38] A. Bujas, T.J. Hastie, and R.J. Tibshirani. Linear Smoothers and Additive Models. 1987.

[39] S.N. Wood. *Generalized Additive Models - An Introduction with R.* Chapman and Hall, 2006.

[40] G. Petris, S. Petrone, and P. Campagnoli. *Dynamic Linear Models with R.* Springer Science+Business Media, 2009.

[41] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models.* Springer, 1997.

# Appendices

# Nomenclature

Throughout the thesis the following mathematical notation and symbols are applied.

| | |
|---|---|
| $\mathbb{R}^n$ | Vector space of $n$ dimensional real vectors |
| $f(\cdot)$ | Notation for the density function of argument $\cdot$ |
| $f(\cdot\|\cdot)$ | Notation for the conditional density function |
| $f(\cdot,\cdot)$ | Notation for the joint density function |
| $p(\cdot)$ | Notation for the probability function of argument $\cdot$ |
| $p(\cdot\|\cdot)$ | Notation for the conditional probability function |
| $N(\mu_i,\sigma^2)$ | Normal distribution with mean vector $\mu_i$ and variance $\sigma^2$ |
| $N_m(\boldsymbol{\mu},\Sigma)$ | Multivariate normal distribution of dimension $m$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma$ |
| $\chi^2_n$ | Chi-square distribution with $n$ degrees of freedom |
| $L(\cdot)$ | Notation for the likelihood function |
| $l(\cdot)$ | Notation for the log-likelihood function |
| $\mathbb{E}[\cdot]$ | Expected value of argument $\cdot$ |
| $\mathbb{E}[\cdot\|\cdot]$ | Conditional expected value of arguments $\cdot$ and $\cdot$ |
| $\mathrm{Var}[\cdot]$ | Variance of argument $\cdot$ |
| $\mathrm{Var}[\cdot\|\cdot]$ | Conditional variance of arguments $\cdot$ and $\cdot$ |
| i.i.d. | Notation for independent and identically distributed random variables |
| $\propto$ | Notation for proportionality |
| 95%CI:$[\cdot;\cdot]$ | 95% confidence interval |
| s.d. | Standard deviation |

d.o.f. Degrees of freedom

w.r.t. With respect to

$U$ Score statistic

$\mathcal{J}$ Information

$\text{Po}(\lambda)$ Poisson distribution with intensity parameter $\lambda$

$I$ Identity matrix

$\sim$ Distributed as

**Generalized Linear and Additive Models**

$y_t$ Response to time $t$

$n_t$ Number of residents in Denmark at time $t$

$\mu_t$ Intensity per resident

$s(\cdot)$ Smoothing function of time $t$

$X$ $n \times p$ design matrix consisting of covariates

$\mathbf{x}_i$ The $i$'th row of $X$

$\mathbf{X}$ Matrix consisting of lagged meteorological variables

$\mathbf{X}^2$ Design matrix consisting of all entrances in $\mathbf{X}$ squared

$\boldsymbol{\theta}$ Vector of parameter estimates

$\boldsymbol{\beta}$ Vector of parameter estimates

$\boldsymbol{\gamma}$ Vector of parameter estimates

$\boldsymbol{\tau}$ Vector of parameter estimates. $\boldsymbol{\beta}$ is a linear combination of the entries in $\boldsymbol{\tau}$

$\boldsymbol{\varphi}$ Vector of parameter estimates. $\boldsymbol{\gamma}$ is a linear combination of the entries in $\boldsymbol{\varphi}$

$p3$ Third-degree polynomial

$\hat{y}$ Fitted values of $y$

| | |
|---|---|
| $(\alpha, \beta)$ | Parameter pair |
| $(a,b)$ | Parameter pair |
| $\eta$ | Linear predictor |
| $g(\mu_i)$ | Link function |
| $\hat{\boldsymbol{\beta}}^{(m)}$ | Estimation of $\boldsymbol{\beta}$ at the $(m)$'th iteration |
| $\hat{\boldsymbol{\beta}}_{max}$ | Maximum likelihood estimate |
| $z_i$ | Adjusted dependent variable |
| $\alpha$ | Intercept |
| $\epsilon_i$ | Error term |
| $N^s(x_i)$ | Symmetric nearest neighborhood |
| $W$ | tri-cube weight function |
| $\lambda$ | Likelihood ratio |
| $D$ | Notation for deviance |
| $r_i$ | Pearson residuals |
| $D_i$ | Cook's distance |

**State Space Models**

| | |
|---|---|
| $Y_t$ | Observation at time $t$ |
| $\theta_t$ | State at time $t$ |
| $F_t$ | Observation matrix at time $t$ |
| $G_t$ | State transfer matrix at time $t$ |
| $V_t$ | Observation variance at time $t$ |
| $v_t$ | Observation error at time $t$ |
| $W_t$ | State variance at at time $t$ |
| $w_t$ | State error at time $t$ |
| $\{F_t, G_t, V_t, W_t\}$ | Quadruple defining a state space model at time $t$ |

$x_{j,t}$ The $j$'th covariate at time $t$

$X_t$ Vector of lagged covariates

$\beta_{j,t}$ Vector of parameter estimates at time $t$ for the $j$'th covariate

$\mathbf{z}_t$ Binary indicator of day of the week at time $t$

$\boldsymbol{\rho}_t$ Vector containing 8 weekday levels

$\boldsymbol{\zeta}_i$ Parameter estimates for seasonal components

$\boldsymbol{\tau}$ Vector of parameter estimates. $\boldsymbol{\beta}$ is a linear combination of the entries in $\boldsymbol{\tau}$

$\boldsymbol{\varphi}$ Vector of parameter estimates. $\boldsymbol{\gamma}$ is a linear combination of the entries in $\boldsymbol{\varphi}$

$p3$ Third-degree polynomial

$\mathbf{a}_t$ Mean of the one-step-ahead predictive distribution of $\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}$ at time $t$

$R_t$ Variance of the one-step-ahead predictive distribution of $\boldsymbol{\theta}_t|\mathbf{y}_{1:t-1}$ at time $t$

$\mathbf{f}_t$ Mean of the one-step-ahead predictive distribution of $\mathbf{Y}_t|\mathbf{y}_{1:t-1}$ at time $t$

$Q_t$ Variance of the one-step-ahead predictive distribution of $\mathbf{Y}_t|\mathbf{y}_{1:t-1}$ at time $t$

$\mathbf{m}_t$ Mean of filtering distribution of $\boldsymbol{\theta}_t|\mathbf{y}_{1:t}$ at time $t$

$C_t$ Variance of filtering distribution of $\boldsymbol{\theta}_t|\mathbf{y}_{1:t}$ at time $t$

$\mathbf{e}_t$ The forecast error at time $t$

$\mathbf{s}_t$ Mean of the smoothing distribution at time $t$

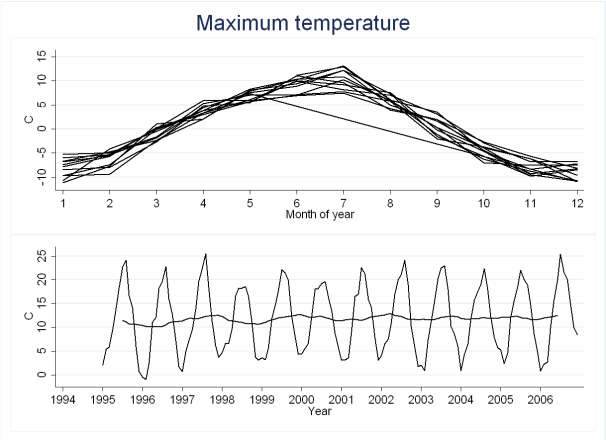
$S_t$ Variance of the smoothing distribution at time $t$

$\boldsymbol{\psi}_t$ Hyper parameter at time $t$

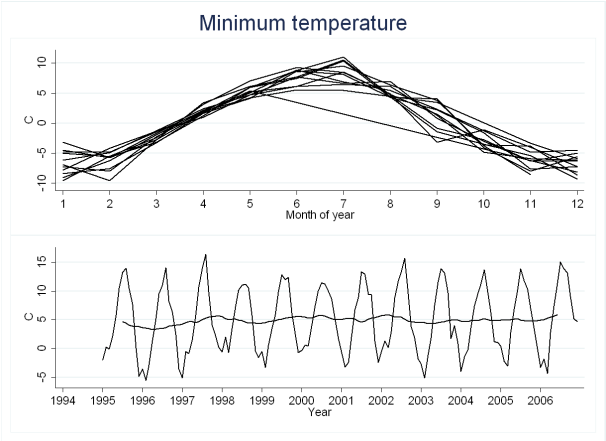
$\hat{\boldsymbol{\psi}}_t$ MLE of the hyper parameter at time $t$

# Plots of Meteorological Variables

The following figures show the maximum and minimum of all the meteorological variables, a spaghetti plot of the season for every year and a trend curve.

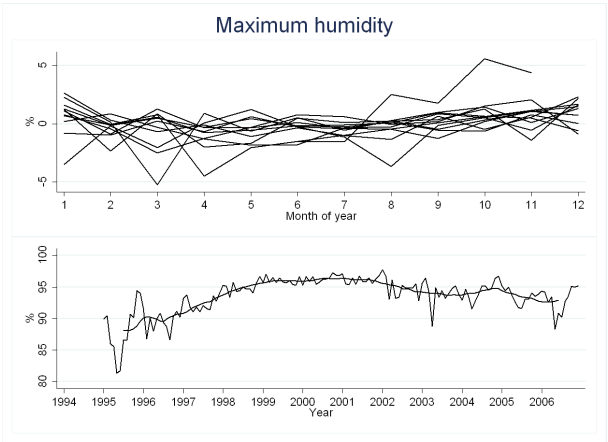



(a) Monthly maximum temperature.

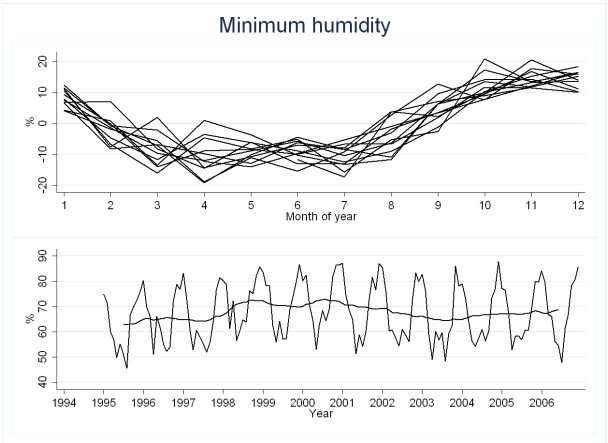



(b) Monthly minimum temperature.

**Figure B.1:** Monthly maximum and minimum temperature in the ten biggest municipalities in Denmark from 1.1.1995-12.31.2006. The x-axis shows the the number of the month. The first figure shows the temperature through the years. The second figure shows a spaghetti plot of the season through the years and the last figure shows the trend through the years.

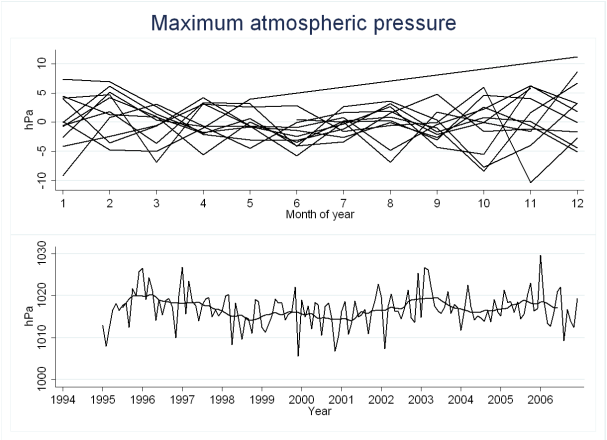



(a) Monthly maximum humidity.

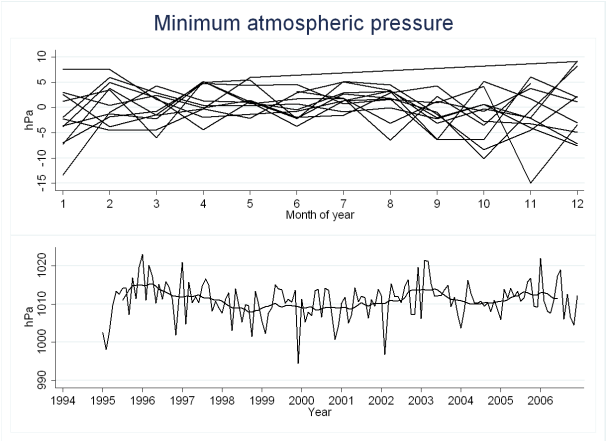



(b) Monthly minimum humidity.

**Figure B.2:** Monthly maximum and minimum humidity in the ten biggest municipalities in Denmark from 1.1.1995-12.31.2006. The x-axis shows the the number of the month. The first figure shows the temperature through the years. The second figure shows a spaghetti plot of the season through the years and the last figure shows the trend through the years.




(a) Monthly maximum air pressure.




(b) Monthly minimum air pressure.

**Figure B.3:** Monthly maximum and minimum air pressure in the ten biggest municipalities in Denmark from 1.1.1995-12.31.2006. The x-axis shows the the number of the month. The first figure shows the temperature through the years. The second figure shows a spaghetti plot of the season through the years and the last figure shows the trend through the years.

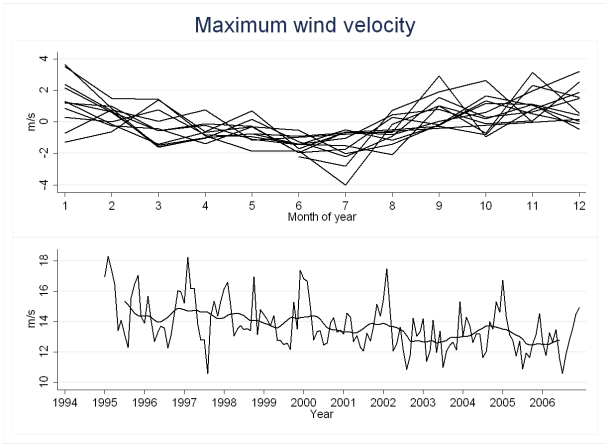



**Figure B.4:** Monthly maximum and minimum wind velocity in the ten biggest municipalities in Denmark from 1.1.1995-12.31.2006. The x-axis shows the the number of the month. The first figure shows the temperature through the years. The second figure shows a spaghetti plot of the season through the years and the last figure shows the trend through the years.

# Further Results Using DLMs

## ACS

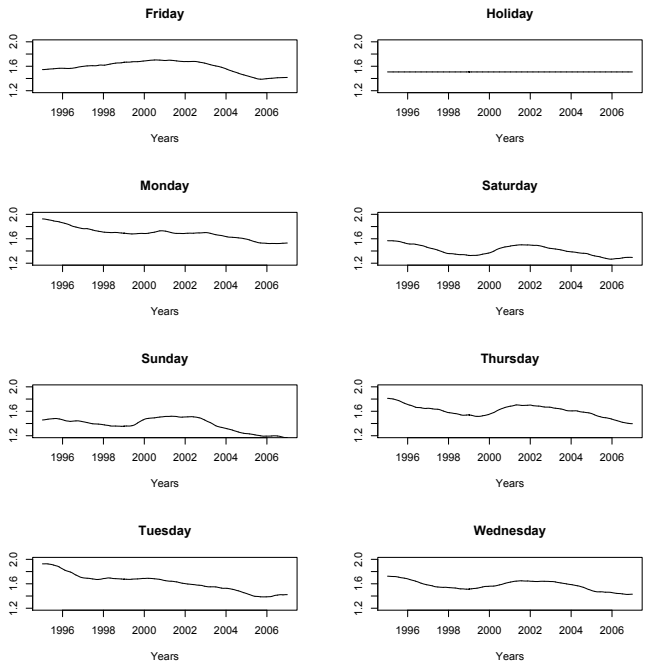



**Figure C.1:** The effect of the day of the week on daily counts of ACS for females. The $y$-axis shows the absolute change in the daily counts of incidences.

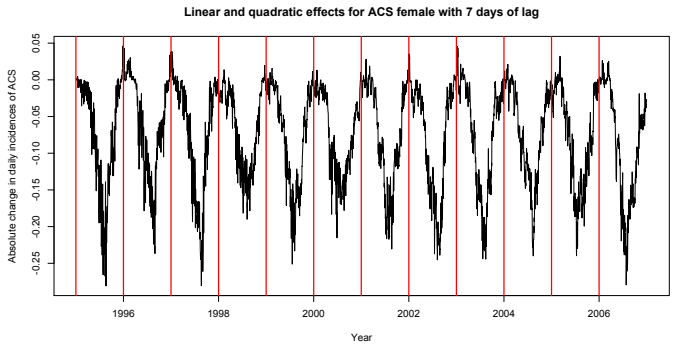



**Figure C.2:** The effect of the temperature on daily counts of incidences of ACS for females.

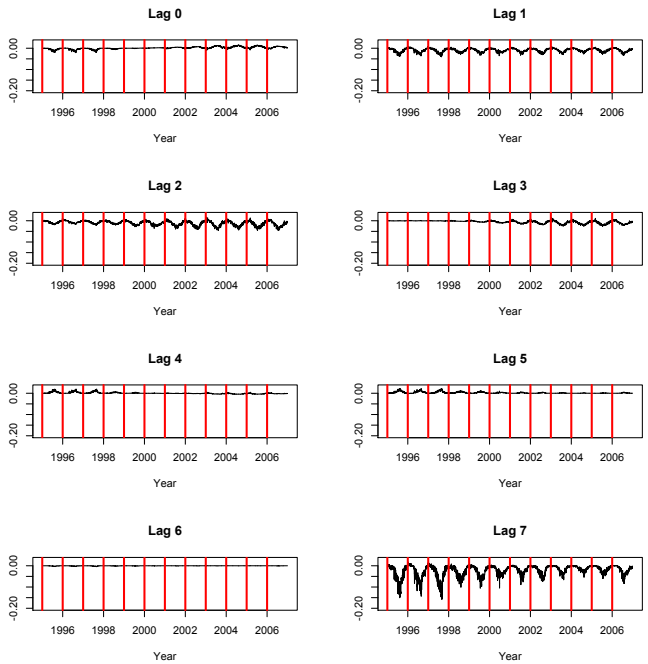



**Figure C.3:** The effect of the temperature on daily counts of incidences of ACS for females for each lag. The $y$-axis shows the absolute change in daily incidences of ACS for each lag.

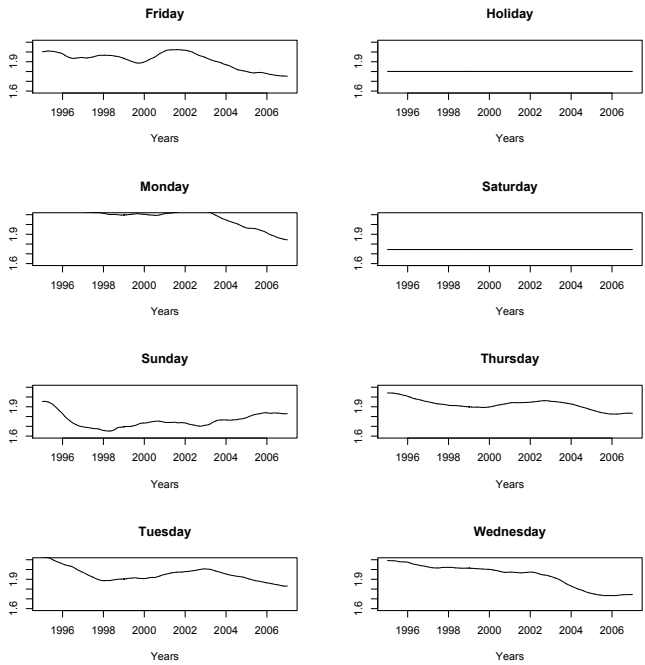



**Figure C.4:** The effect of the day of the week on daily counts of ACS for males. The $y$-axis shows the absolute change in the daily counts of incidences.

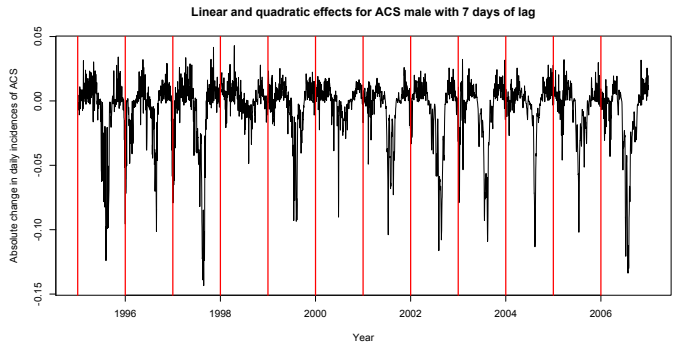



**Figure C.5:** The effect of the temperature on daily counts of incidences of ACS for males.

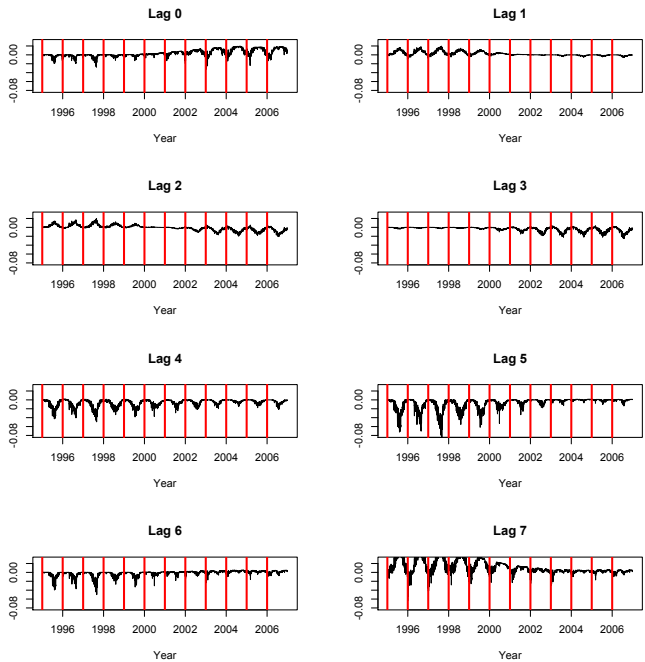



**Figure C.6:** The effect of the temperature on daily counts of incidences of ACS for males for each lag. The $y$-axis shows the absolute change in daily incidences of ACS for each lag.

# APO

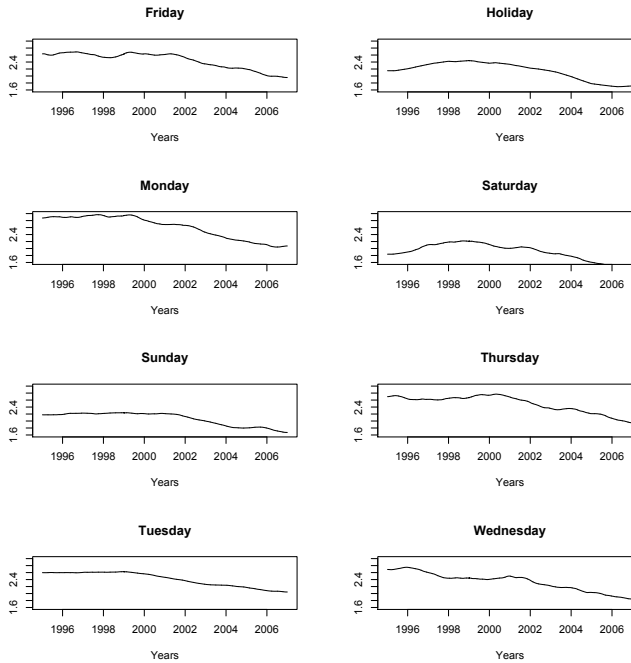



**Figure C.7:** The effect of the day of the week on daily counts of APO for females. The $y$-axis show the absolute change in the daily counts of incidences.

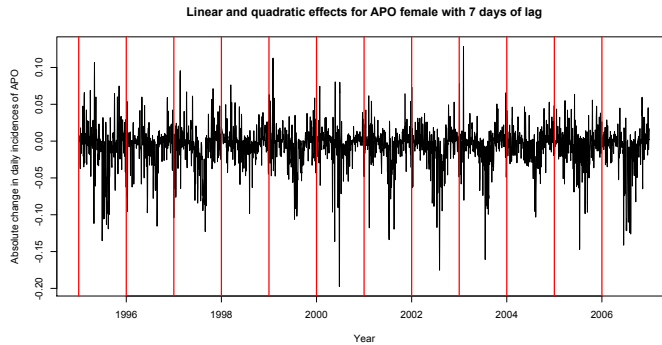



**Figure C.8:** The effect of the temperature on daily counts of incidences of APO for females.

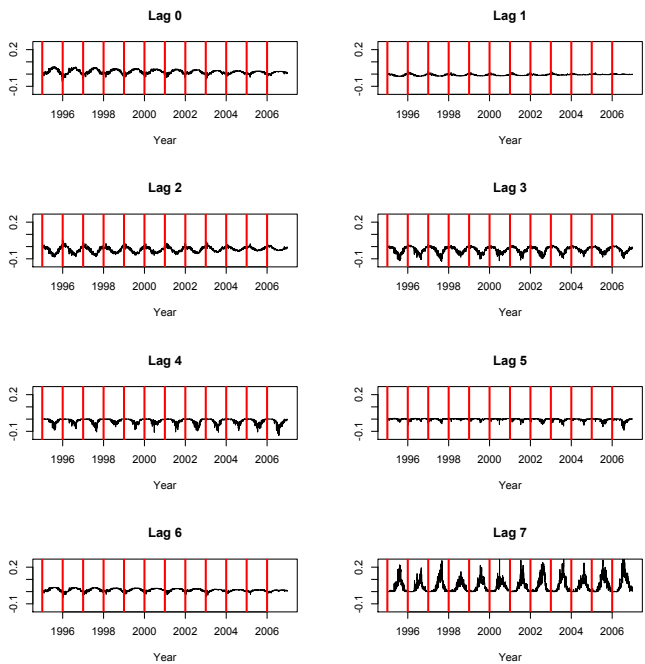



**Figure C.9:** The effect of the temperature on daily counts of incidences of APO for females for each lag. The $y$-axis show the absolute change in daily incidences of APO for each lag.

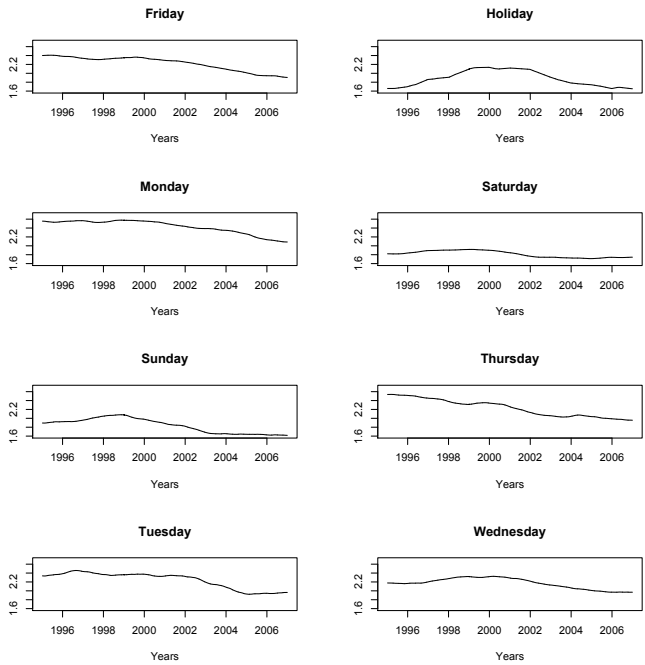



**Figure C.10:** The effect of the day of the week on daily counts of APO for males. The $y$-axis show the absolute change in the daily counts of incidences.

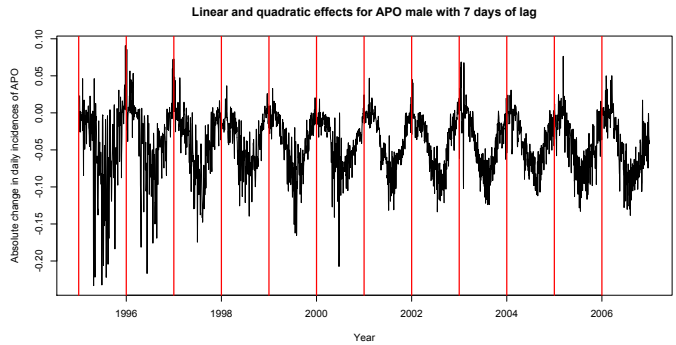



**Figure C.11:** The effect of the temperature on daily counts of incidences of APO for males.

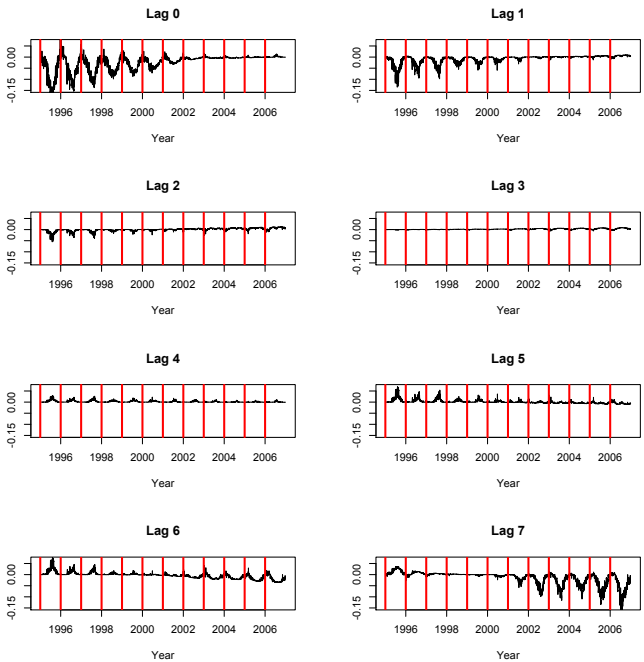



**Figure C.12:** The effect of the temperature on daily counts of incidences of APO for males for each lag. The $y$-axis show the absolute change in daily incidences of APO for each lag.

# VTE

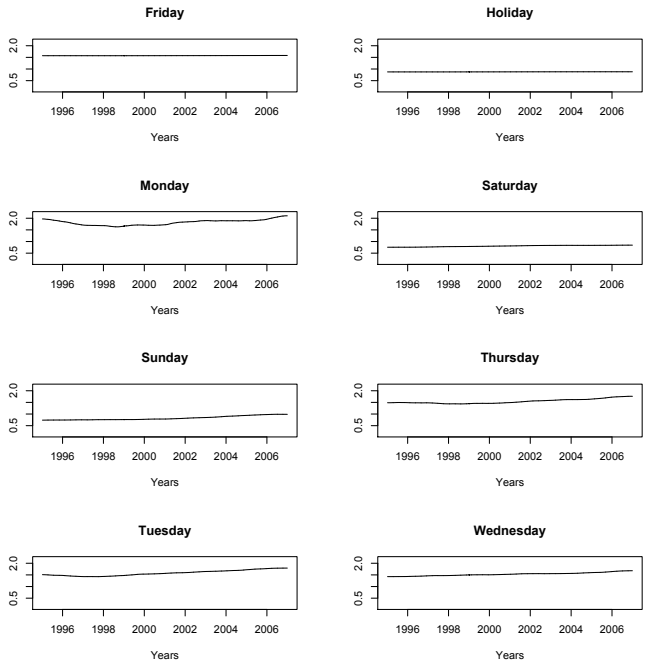



**Figure C.13:** The effect of the day of the week on daily counts of VTE for females. The $y$-axis show the absolute change in the daily counts of incidences.

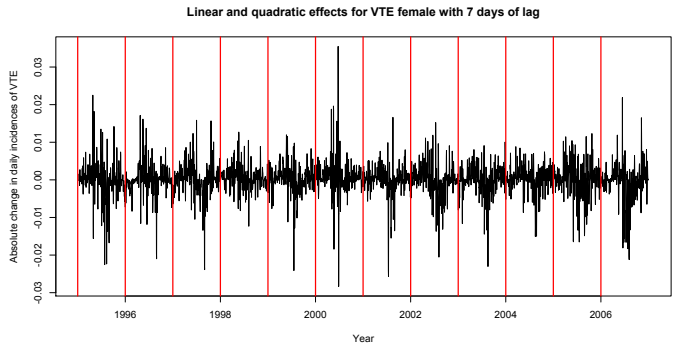



**Figure C.14:** The effect of the temperature on daily counts of incidences of VTE for females.

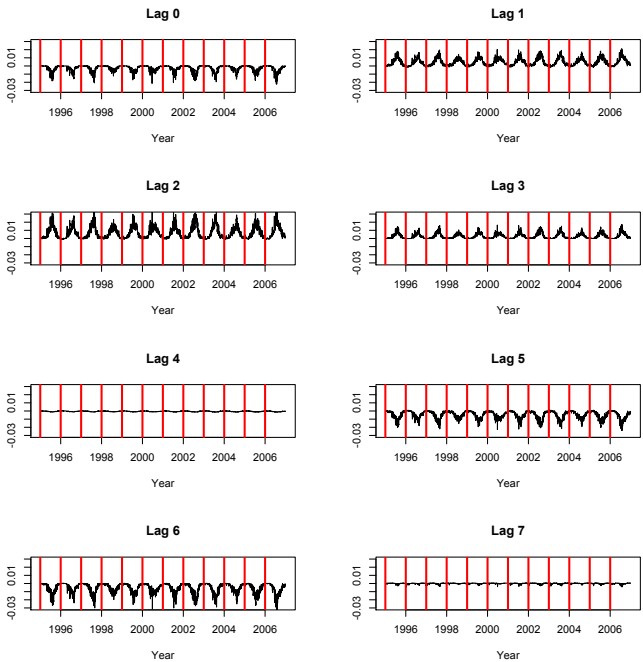



**Figure C.15:** The effect of the temperature on daily counts of incidences of VTE for females for each lag. The $y$-axis show the absolute change in daily incidences of VTE for each lag.

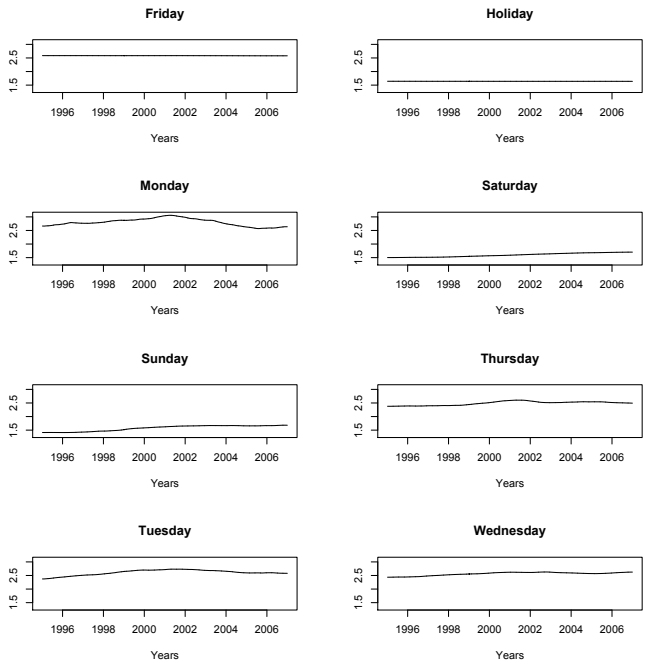



**Figure C.16:** The effect of the day of the week on daily counts of VTE for males. The $y$-axis show the absolute change in the daily counts of incidences.

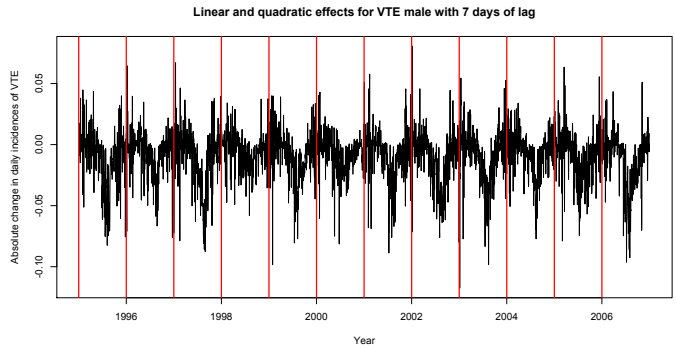



**Figure C.17:** The effect of the temperature on daily counts of incidences of VTE for males.

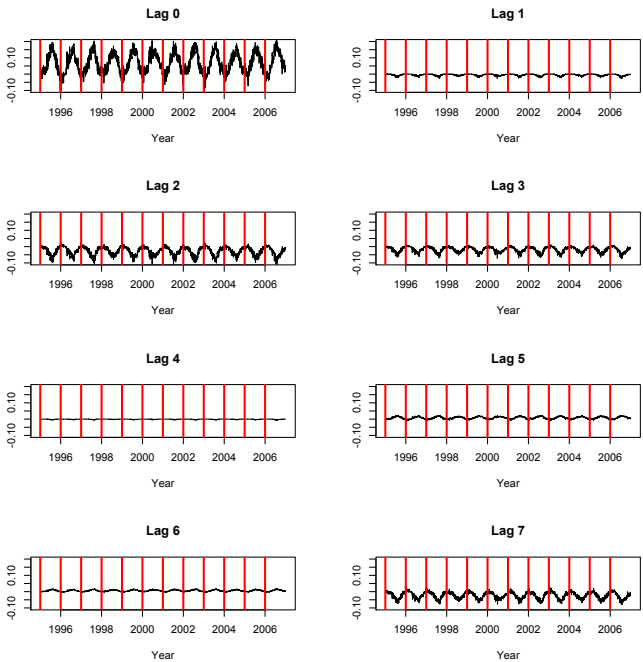



**Figure C.18:** The effect of the temperature on daily counts of incidences of VTE for males for each lag. The $y$-axis show the absolute change in daily incidences of VTE for each lag.

# AF

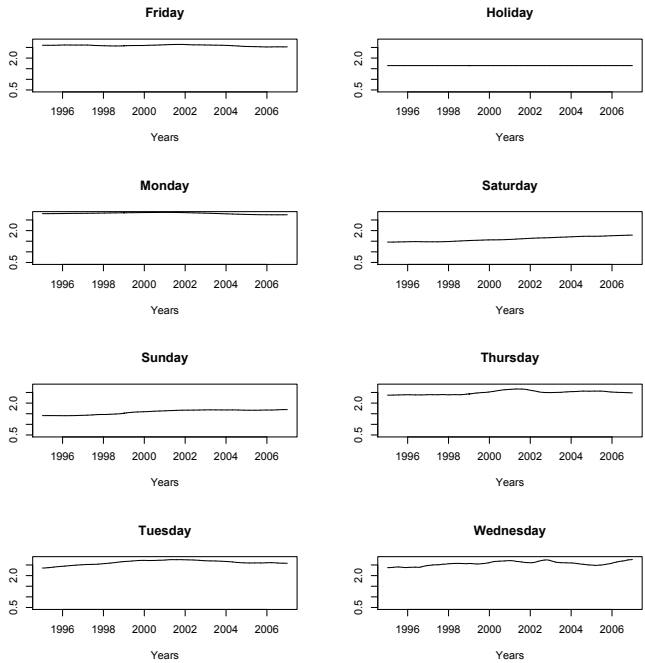



**Figure C.19:** The effect of the day of the week on daily counts of AF for females. The $y$-axis show the absolute change in the daily counts of incidences.

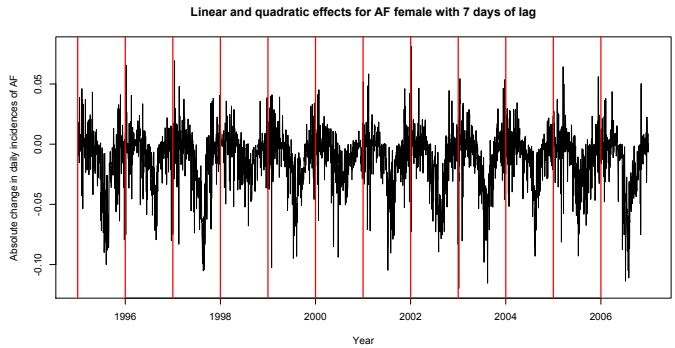



**Figure C.20:** The effect of the temperature on daily counts of incidences of AF for females.

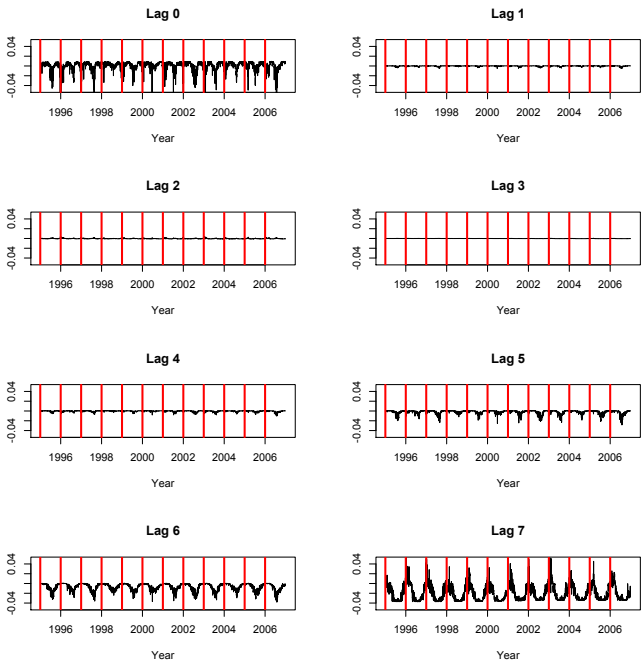



**Figure C.21:** The effect of the temperature on daily counts of incidences of AF for females for each lag. The $y$-axis show the absolute change in daily incidences of AF for each lag.

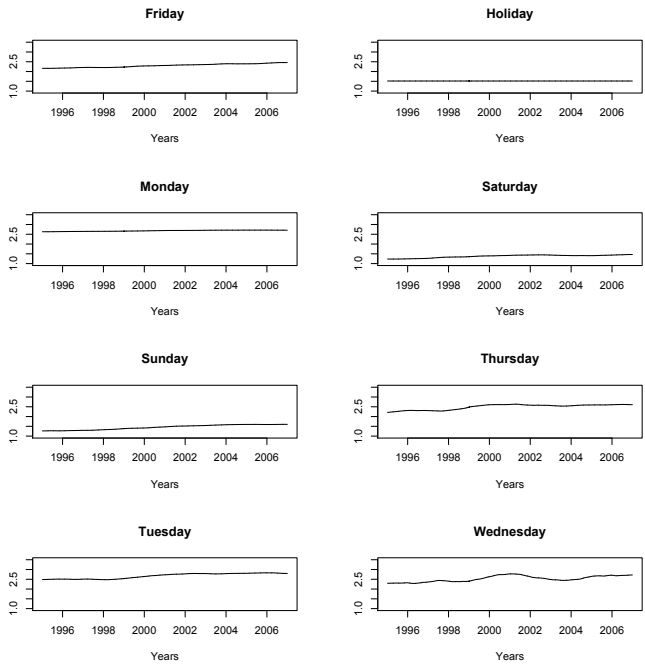



**Figure C.22:** The effect of the day of the week on daily counts of AF for males. The $y$-axis show the absolute change in the daily counts of incidences.

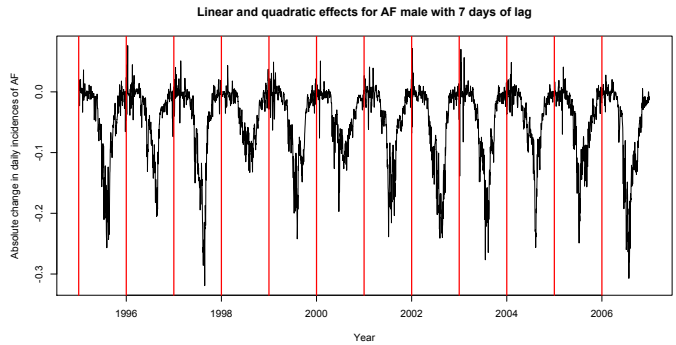



**Figure C.23:** The effect of the temperature on daily counts of incidences of AF for males.

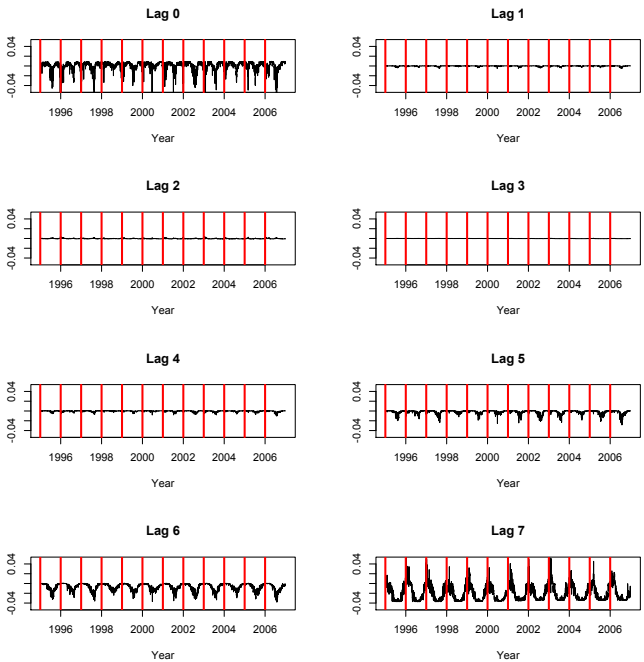



**Figure C.24:** The effect of the temperature on daily counts of incidences of AF for males for each lag. The $y$-axis show the absolute change in daily incidences of AF for each lag.