Prediction Models for Classification

Predicting Autism Spectrum Disorder and Attention Deficit Hyperactive Disorder Diagnoses From the Danish Health Registers

Master Thesis

Nicolai S. Schjøtt & Simon Grøntved

Aalborg University
Department of Mathematical Sciences

Department of Mathematical Sciences
Aalborg University
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Prediction Models for Classification
Predicting Autism Spectrum Disorder and Attention Deficit Hyperactive Disorder Diagnoses From the Danish Health Registers

Theme:
Prediction models

Project Period:
Fall semester 2018
Spring semester 2019

Project Group:
5.219a

Participants:
Nicolai Søndergård Schjøtt
Simon Grøntved

Supervisors:
María Rodrigo Domingo
Anne Marie Svane

Date of Completion:
June 10, 2019

Abstract:

The aim of this master thesis is to build the best prediction model that classify which children and adolescents get ASD or ADHD, respectively. If the model predict sufficiently good, it can be used to support the theory within the fields of ASD and ADHD. If the model is very good at predicting, it can be used by clinicians to substantiate their suspicion of diagnosis.
We started out by writing a protocol, used to order the data set used in this master thesis. Since it takes a long time from us ordering the data until us receiving the data, we end up simulating a data set, which we expected had the same properties as the ordered. This has proved to be a great advantage, as we have learned to simulate, link theory and practice and it has prepared us for the ordered data set.
The master thesis focuses on the classification method logistic regression, where we use splines for our continuous variables and LASSO to select the most important variables. We also use other non-likelihood based classification methods such as classification trees, which also contributed to our variable selection. When we fit a prediction model, it is important to determine whether it predicts good at all and whether it predicts better than other models. To determine this, we have used various evaluation measures, but our main focus has been AUC. All our evaluation measures are 10-fold cross-validated.
We do not recommend using the models that we have reached at present time, but rather we recommend expanding our thoughts and ideas for further research. We experience problems with logistic regression in the form of a time-dependent response as well as informative censoring for the predictors. Furthermore, we believe that one of the most advantageous improvements would be to add more predictors for the prediction models to become sufficiently good.

Contents

Preface

This master thesis has been written by two students at mathematics, Aalborg University. To be able to read this project, it is expected that the reader has a basic understanding of statistics.

The thesis is structured in a theory section with examples followed by an analysis section, applying the theory. Figures, tables and equations are numbered consecutively by chapter. References are listed by a number in brackets and a complete bibliography is available at the end of the project. A section is always based on [1] or [2] unless otherwise stated in a footnote at the beginning of a section. The sign ■ indicates completion of a proof and □ indicates the end of an example. The program R [3] is used for all statistics calculations and analyzes.

Notation

Throughout this thesis vectors are presented in bold $\mathbf{x_i}, \mathbf{Y}, \boldsymbol{\beta}$ and are always column vectors. Scalars are unbolded $x_{ij}, y_i, \beta_j, c, k$ and estimates have a hat $\hat{\boldsymbol{\beta}}, \hat{y}_i$. Matrices are upper case X, Σ, which is also the case for random variables Y, X_j where the distinction should be self-evident from the context.

For the data matrix X the i'th row, denoted $\mathbf{x_i}$, represents all predictors for observation i. The jth column of the data matrix X are all observations within predictor j, denoted $\mathbf{x}_{\bullet j}$, making x_{ij} the predictor j value for observation i.

Nicolai Søndergård Schjøtt
<nschjo10@student.aau.dk>

Simon Grøntved
<sgront12@student.aau.dk>

Funding

The project has been supported by The Psychiatric Research Unit, North Denmark Region Psychiatry as they have financed both our supervisor María Rodrigo Domingo and the access to and hosting of the data at Danmarks Statistik's servers. The Research unit for Child and Adolescent Psychiatry has financed the actual data sets delivered from Danmarks Statistik and Sundhedsdatastyrrelsen. The Department of Mathematics, Aalborg University, has financed our supervisor Anne Marie Svane.

Acknowledgments

We would first of all like to thank our supervisors Anne Marie Svane and María Rodrigo Domingo for their competent and engaged supervision, we would not have been able to reach this level of our potential without. We thank Marlene Briciet Lauritsen for the insights into the fascinating field of psychiatry and for her department's financial support. A big thanks to Rasmus Wentzer Licht for allowing us to conduct the study at The Unit for Psychiatric Research. We further wish to tank Anne Sofie Hansen and Anna Sofie Hansen for shearing their knowledge, medical in general and psychiatric in particular; it has been of great value in our efforts to understand our data set. We also thank everybody else at The Unit for Psychiatric Research for making us feel welcome and for the many discussions and advice on how to conduct epidemiological research. A special thanks to Jan Brink Valentin for originally coming up with the idea of us conducting a predictive study. Thanks to all the other masters students at the Department og Mathematics for struggling alongside us, making it possible for us to maintain at least some social life (if only during the lunch breaks). We also thank our families. Nicolai particularly thanks Marie for her support throughout the study, especially tolerating the though exam periods with little time at home. Simon thanks his parents and in-laws for helping out with the kids on so many occasions. He thanks his kids Otto and Alberte for still recognizing him as their dad, but most of all he thanks Henriette for still being by his side after too long being second to school and exams.

Danish Summary (Dansk resumé)

Dette kandidatspeciale handler om prædiktionsmodeller og hvordan de kan anvendes til klassifikation. Vi ønsker at prædiktere hvilke børn og unge, der er i risikogruppe for at få hhv. ASD eller ADHD på baggrund af en række variable. Formålet med dette speciale er at lave den bedste prædiktionsmodel baseret på data trukket fra de danske registre. Hvis modellen prædikterer tilstrækkelig godt, kan den anvendes til at underbygge teori inden for felterne ASD og ADHD eller antyde nye ukendte sammenhænge. Hvis modellen prædikterer meget præcist, kan den anvendes af klinikere til at underbygge deres mistanke om diagnose.

Vi har selv skrevet protokol til at bestille registerdata fra kandidatspecialets start og har derfor måtte vente på at modtage vores datasæt. Da der går lang tid fra bestilling til modtagelse, valgte vi at simulere et datasæt, der kunne afspejle de egenskaber, vi kunne

forvente at se i det bestilte datasæt. Dette har vist sig at være en stor fordel, idet vi herved har lært at simulere, koble teori og praksis samt forberedt os på, hvordan vi skulle håndtere det bestilte datasæt. Projektet har fokus på klassifikationsmetoden logistisk regression, hvor vi blandt andet har anvendt splines for vores kontinuerte variable samt LASSO til at udvælge de vigtigste variable. I kandidatspecialet har vi også anvendt andre ikke-likelihood baserede klassifikationsmetoder såsom klassifikations træer, der også har bidraget til vores variabel udvælgelse.

Fra kandidatspecialets start var vores plan at lave væsentlig flere variable, men da vi modtog vores datasæt forholdsvis sent, kunne dette ikke realiseres. Vi havde ellers forberedt os på at skulle anvende metoder til at udføre en sonderende analyse af vores variable ved at anvende PCA samt klynge analyse metoderne K-means- og hierarchical clustering.

Når vi har tilpasset en prædiktionsmodel, er det naturligvis vigtigt at afgøre, om den overhovedet prædikterer godt og om den prædikterer bedre end andre modeller. Til at afgøre dette har vi anvendt diverse evalueringsmål, hvor der hovedsageligt er lagt vægt på AUC, eftersom der for vores datasæt er en meget lav forekomst af diagnosticeret personer med ASD og/eller ADHD.

Efter modtagelse af datasættet har vi udført datamanagement samt lavet en analyse. Analysen handler i høj grad om kalendertidsproblemer, fordi der i løbet af årene 1995 til 2017 bliver diagnosticeret markant flere personer med ASD og ADHD hvert år. Et andet problem, som vi støder på i analysen er imputering af manglende værdier samt, at der ikke er fuld opfølgning på mange personer i vores datasæt, eftersom vores population er børn og unge under 18 år.

Vi anbefaler ikke at anvende de modeller, som vi er nået frem til på nuværende tidspunkt, men derimod at udbygge vores tanker og ideer til videre forskning. Den bedste prædiktionsmodel for dette kandidatspeciale er bygget på baggrund af klassifikationsmetoden logistisk regression. Vi oplever dog problemer med logistisk regression i form af en tidsafhængig respons samt informativ censurering for nogle prædiktorer. Vi anbefaler, at der konstrueres flere variable, før prædiktionsmodellerne bliver tilstrækkelig gode. Ydermere anbefaler vi, at det undersøges nærmere, hvordan manglende værdier skal behandles.

Introduction

The reported prevalence of Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactivity Disorder (ADHD) has been increasing worldwide [4, 5], to a point where it is estimated that 1% of the Danish population have one of the diagnoses of ASD [6] and 5% one of the diagnoses of ADHD [7]. Various factors have been linked to ASD, some to ADHD, and some to both. These factors include prenatal risk factors, early-life infections, inherited and molecular factors, and other environmental factors [8–21].

Statistical prediction tools can be used to aid in the diagnostic and prognostic processes and in some cases might form the basis for public health recommendations. Such tools have been used on a number of somatic conditions including breast and ovarian cancer [22], myocardial infarction [23, 24], and lymphoma [25]. In Denmark, the models used for estimating the risk of chromosomal abnormalities during the 12-week pregnancy scan [26] are probably the best known application of prediction models. Among the statistical methods previously used to predict disease in other areas of health care are Cox Proportional Hazards [22, 24], logistic regression [23], and various machine learning classification algorithms [25]. Within the psychiatric field, one study predicted depression [27], while another study predicted ADHD in a Danish Cohort [28], and several others have tried to predict ADHD [29, 30] or ASD [31, 32] outside of Denmark. To our knowledge, no previous studies have statistically predicted both ADHD and ASD in a Danish cohort using a broad range of predictors.

We have thus defined the project:

> "Predicting Autism Spectrum Disorder and Attention Deficit Hyperactive Disorder in a Danish Cohort of Children and Adolescents".

The aim of this project is to develop statistical models to predict a diagnosis of ASD and ADHD, both individually and combined, based on the Danish registers. We want to develop different predictive models, both regarding the time to prediction and the selection of predictors, and to evaluate and compare such models.

The project is conducted in collaboration with the Research unit for Child and Adolescent Psychiatry and The Psychiatric Research Unit, North Denmark Region Psychiatry, Aalborg, Denmark, that will grant us access to several Danish registers. As part of this collaboration we plan to write several peer reviewed articles that will spring from decisions and recommendations produced by this master thesis.

In this thesis several supervised learning models will be specified and tested. Because of the nature of the outcome (presence/absence of ASD and/or ADHD), we will use methods for classification, mainly the methods described in [1, 2].

To conduct and present our work we will follow the recommendations in [33–35]. The initial selection of predictors includes risk factors known from the literature and are chosen in collaboration with an experienced child and adolescent psychiatrist, Marlene Briciet Lauritsen[1]. We will propose different forms of coding the predictors including re-categorization, splines, and interactions between predictors and investigate how this affects the models. Under variable selection and model specification we might include methods such as subset selection, PCA and shrinkage. Bootstrapping and/or cross-validation will be used to evaluate the model. The model performance will be assessed both visually via ROC curves and computationally with, among others, the evaluation F-score [35, 36].

In Chapter 1, we first describe how we designed the study and why. We round off the chapter by simulating a small data set for use in our examples throughout the thesis. This data set will be simulated in such a way that some of the properties expected in the real data are represented. In Chapter 2 we describe, how we intend to explore the data. The chapter contains theory and application of Principal Component Analysis, Hierarchical Clustering and K-Means Clustering, which are all unsupervised learning methods. Chapter 3 covers the main classification method of this thesis, logistic regression, both in theory and examples. For logistic regression we present methods for transforming predictors and methods for selecting predictors. We also present several other methods for classification and compare them to logistic regression. The methods we use for evaluation of our models are presented, theoretically, described and tested on simulated data in Chapter 4. In Chapter 5 we describe what data we use and the registers from which it originates. We then apply most of the methods presented in the chapters 2, 3 and 4 to this real data.

[1] Child and Adolescent Psychiatry, Region of Northern Jutland Psychiatry, Aalborg, Denmark

1. Study Design

Most of the decisions made regarding the design of this study are based on the article "To Explain or to Predict" by Galit Shmueli [34], hence this chapter starts out with a short summary of this article. After this we present a section on how we simulate a small data set for use in our examples throughout the thesis.

To Explain or to Predict

Shmueli discriminates between Explanatory studies and Prediction studies, and argues that several decisions throughout a study should take this distinction into account.

Explanatory modeling is conducted when the theory in a given field is studied and certain hypothesis are set up, usually about a causal mechanism. Statistics can then be used to support an association, and thus verify the causal mechanism set up by the field specific theory. Statistical models are also used to describe the magnitude of such an association. Explanatory statistical models are thus generally preferred to be comprehensible and interpretable.

Predictive modeling, which is the focus of this thesis, aims to best predict an outcome based on available data. In predictive modeling, the field specific theory is used to select the predictors, on which to base the statistical model, but other than that, a predictive model does not rely on the field's specific theory, there is no hypothesis to reject or approve. As the goal is to best predict, complicated and less interpretable models are accepted, as long as they produce better results than simpler ones. We will later define, what better is.

A central part of study design is study size. In explanatory studies, as the object is to explain, confirm or reject some field specific theoretical assumptions, there is a limit to how large a study is needed to reach a certain level of certainty. After that limit is reached, increasing the study size provides only small amounts of further information, thus making it possible to respect time and cost restrictions. Furthermore, the study size in an explanatory study can be lowered by selecting the participants in a way so that only the necessary information is gathered. This is not preferable in a predictive study. Here all data contains useful information, and "cleaning" of data can lead to neglecting unknown dependencies. Using noisy data restricts interpretability which is not as big an issue in predictive studies as in an explanatory one. The sample size in predictive studies

is chosen based on availability, here more data will always lead to better results, so the goal is to optimize sample size with respect to time, cost, and data quality. Whereas in an explanatory study the variables are preferred to be as close to the assumed underlying theoretical construct as possible. This would lead to interpretable results, and make sure that any finding actually does match the theories in question.

In a predictive study some underlying theory is indeed needed, as predictors should at least be associated, even if not causally, with the outcome one is trying to predict. But as no assumptions are to be described, the main considerations when choosing predictors are the quality and availability. Furthermore, as interpretability is not a goal in itself, one can deviate from the hierarchical principle and include interactions without including main effects. For a predictive study to be applicable to the real world, a predictor should be chronologically available prior to the response.

Thus, the model construction and variable selection in an explanatory study is done prior to evaluation, based on theory from the relevant discipline, whereas in a predictive study the model selection is done based on statistical evaluation of how well the model predicts, or is expected to predict, new data.

The study on which this masters thesis is based has access to data from the Danish registers, which have some of the world's widest covering and a highest data quality [37, 38]. We understand from "To Explain or to Predict" by Galit Shmueli [34] that the high data quality and wide coverage makes the Danish registers ideal for predictive studies.

1.1 Simulated Data

To investigate and explain the methods presented in this thesis, we have simulated a data set. This data set is designed to emulate some of the properties we expect in the real data set. The simulated data set consist of 1000 observations, where 10 observations of the response class `1percent` are drawn from the multivariate normal distribution $\mathcal{N}_2\left(\left[\begin{smallmatrix}180\\60\end{smallmatrix}\right], \left[\begin{smallmatrix}60 & 27\\27 & 40\end{smallmatrix}\right]\right)$, 50 observations of the class `5percent` drawn from $\mathcal{N}_2\left(\left[\begin{smallmatrix}165\\85\end{smallmatrix}\right], \left[\begin{smallmatrix}70 & -22\\-22 & 35\end{smallmatrix}\right]\right)$ and 940 observations of the class `94percent` drawn from $\mathcal{N}_2\left(\left[\begin{smallmatrix}160\\55\end{smallmatrix}\right], \left[\begin{smallmatrix}30 & -24\\-24 & 50\end{smallmatrix}\right]\right)$. The names of the classes (`1percent`, `5percent` and `94percent`) are based on the fact that the class sizes are the same as we expect ASD and ADHD to be in the real data, based on the literature [6, 7]. The means and variances are chosen, such that it is possible to distinguish the classes, the covariances are furthermore randomly generated, but restrained by the pre-chosen variances and the fact that a covariance matrix is positive semidefinite. Figure 1.1 shows a plot of the data set. The simulated data set will be tweaked and expanded in

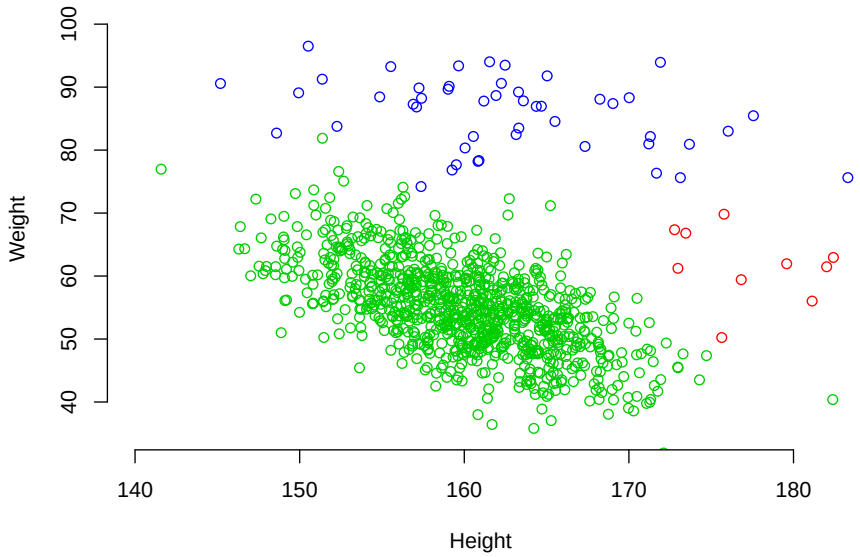


Figure 1.1: Three simulated classes based on the bi-variate normal distributions of `1percent` (red), `5percent` (blue) and `94percent` (green).

examples throughout the thesis. We call the values on the x-axis `Height` and the values on the y-axis `Weight`.

2. Exploratory Data Analysis

The aim of this chapter is to acquire knowledge about the data set, based only on the data itself. We thus consider unsupervised learning methods, where the interesting thing is not to predict a predetermined $\mathbf{y}$, but rather to study the relationship between the predictors $\mathbf{x}_{\bullet j}$ for $j = 1, \ldots, p$. An approach addressed is clustering, where it is examined which predictors are similar to each other and on the basis of this subgroups of predictors are found, where a subgroup has predictors similar to each other and predictors in different subgroups are quite different. Another approach known as Principal Component Analysis (PCA) creates new and fewer predictors by making linear combinations of the original predictors.

As the relationships between predictors are explored, this may give rise to exclusion or inclusion of predictors in the models. It should be noted that there is no way to evaluate unsupervised learning methods, and thus the performance of the methods is subjective, whereas models in a supervised learning setup can be evaluated by using evaluation measures, which is discussed in Chapter 4.

2.1 Clustering Methods

In our project, clustering is used to split the set of predictors into subgroups where each subgroup includes predictors that are quite similar, and thus are assumed to capture the same latent variable. Predictors in different subgroups are assumed to capture different latent variables. The purpose of clustering methods in our project is therefore to find structure in data that should be considered throughout the analysis, as this may give rise to including or excluding predictors in models or found a basis for choosing which predictors to aggregate. Two common clustering methods are K-means clustering and hierarchical clustering, as the rest of this section elaborates.

2.1.1 K-Means Clustering

Choose K as the number of clusters desired for the p predictors and let $C_1, C_2, \cdots, C_K$ denote the sets that contain the indices of the predictors in each cluster, that is $j \in C_k$ means that predictor $\mathbf{x}_{\bullet j}$ is contained in the k'th cluster. How to choose an optimal K in a given situation is treated in an example after this theoretical section. According to

K-means clustering, the K clusters must satisfy that all predictors are contained in one of the clusters and that each predictor is only contained in one of the clusters. In order to find out which cluster a predictor should belong to, a measure $W(C_k)$ is introduced, which indicates how different predictors within cluster k are. Often, squared Euclidean distance between standardized predictors is used to measure the variation within a cluster k given by

$$W(C_k) = \frac{1}{|C_k|} \sum_{j,j' \in C_k} \sum_{i=1}^{n} \left(x_{ij} - x_{ij'} \right)^2, \tag{2.1}$$

where $|C_k|$ denotes the number of predictors in the k'th cluster. It is desirable to have minimum variation within each cluster and therefore it is desired to solve the optimization problem

$$\underset{C_1, C_2, \cdots, C_K \in C}{\arg\min} \left(\sum_{k=1}^{K} W(C_k) \right), \tag{2.2}$$

where C is the set of all possible clusters. There are almost K^p ways to split the p predictors, thus the optimization problem is not straightforward to solve. Instead, Algorithm 2.1 is performed, which will provide a local minimum.

Algorithm 2.1 K-Means Clustering

1: Each predictor is randomly assigned a number from 1 to K.
2: **repeat**
3: For each cluster a so-called cluster centroid vector is calculated, where the average of each of the n observations values of the predictors contained in this cluster is found.
4: Each predictor is then assigned to the cluster where the distance between the predictor and the cluster centroid is smallest.
5: **until** all predictors stop changing clusters

Since the algorithm will find a local minimum, which depends to a large extent on step 1 of the algorithm, the algorithm should be run multiple times and the result minimizing (2.2) is chosen.

The algorithm converges towards a local minimum because (2.1) can be written as

$$\frac{1}{|C_k|} \sum_{j,j' \in C_k} \sum_{i=1}^{n} \left(x_{ij} - x_{ij'} \right)^2 = 2 \sum_{j \in C_k} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_{ik} \right)^2, \tag{2.3}$$

where $\bar{x}_{ik} = \frac{1}{|C_k|} \sum_{j \in C_k} x_{ij}$.

Line 3 in Algorithm 2.1 calculates such cluster centroids $\bar{x}_{ik}$, and relocating the predictors in line 4 can therefore only make (2.2) smaller because of (2.3).

The calculation in (2.3) holds as

$$\frac{1}{|C_k|}\sum_{j,j'\in C_k}\sum_{i=1}^{n}\left(x_{ij}-x_{ij'}\right)^2 \tag{2.4}$$

$$=\sum_{j\in C_k}\sum_{i=1}^{n}\sum_{j'\in C_k}\frac{1}{|C_k|}\left(x_{ij}-\bar{x}_{ik}+\bar{x}_{ik}-x_{ij'}\right)^2 \tag{2.5}$$

$$=\sum_{j\in C_k}\sum_{i=1}^{n}\sum_{j'\in C_k}\frac{1}{|C_k|}\left((x_{ij}-\bar{x}_{ik})^2+(\bar{x}_{ik}-x_{ij'})^2+2(x_{ij}-\bar{x}_{ik})(\bar{x}_{ik}-x_{ij'})\right), \tag{2.6}$$

where the last terms are equal to 0 because

$$\sum_{j'\in C_k}\frac{2}{|C_k|}(x_{ij}-\bar{x}_{ik})(\bar{x}_{ik}-x_{ij'}) \tag{2.7}$$

$$=\frac{2}{|C_k|}(x_{ij}-\bar{x}_{ik})\sum_{j'\in C_k}(\bar{x}_{ik}-x_{ij'}) \tag{2.8}$$

$$=\frac{2}{|C_k|}(x_{ij}-\bar{x}_{ik})\sum_{j'\in C_k}\left(\frac{1}{|C_k|}\sum_{j\in C_k}x_{ij}-x_{ij'}\right) \tag{2.9}$$

$$=\frac{2}{|C_k|}2(x_{ij}-\bar{x}_{ik})\left(\frac{1}{|C_k|}|C_k|\sum_{j\in C_k}x_{ij}-\sum_{j'\in C_k}x_{ij'}\right)=0. \tag{2.10}$$

The remaining part of (2.6) is thus

$$\sum_{j\in C_k}\sum_{i=1}^{n}\sum_{j'\in C_k}\frac{1}{|C_k|}\left((x_{ij}-\bar{x}_{ik})^2+(\bar{x}_{ik}-x_{ij'})^2\right)=\sum_{j\in C_k}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{ik}\right)^2+\sum_{j'\in C_k}\sum_{i=1}^{n}\left(\bar{x}_{ik}-x_{ij'}\right)^2 \tag{2.11}$$

$$=2\sum_{j\in C_k}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_{ik}\right)^2, \tag{2.12}$$

which is the right hand side of (2.3).

Example

As the idea in clustering for us is to cluster predictors, we first need more predictors than we currently have in the simulated data set. We construct predictors that we expect to cluster by transforming the ones we already have. One of these transformations is `Weight2`, which is created by multiplying `Weight` by 8 and adding a normally distributed noise $\mathcal{N}(0, 10)$. In this fashion we have created what we expect to be two clusters `(Height, Height1, Height2, Height3)` and `(Weight, Weight1, Weight2, Weight3)`. As `Height` and `Weight` are possibly correlated through the three classes `1percent`, `5percent` and `94percent`, they might end up clustering together due to this correlation. Because of this we generated two more possible clusters by drawing 1000 observations from $\mathcal{N}(50, 30)$ and 1000 observations from Unif$(0, 1)$. Based on these draws we made the clusters `(Normal, Normal1, Normal2, Normal3)` and `(Uniform, Uniform1, Uniform2, Uniform3)` the same way as the others, by transforming and adding noise. The exact transformations can be seen in Table A.1 in Appendix A.

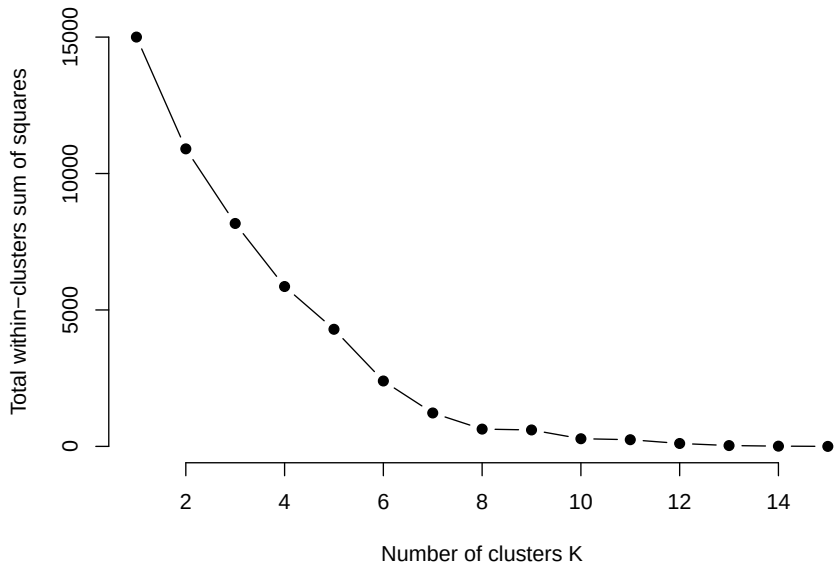


Figure 2.1: Elbow-plot of K-means on scaled data with 16 predictors of 1000 observations.

When performing the K-means analysis, K has to be chosen. We make this choice based on an elbow plot. One such plot is seen in Figure 2.1, here we see the total within cluster sum of squares plotted against K. We choose 7 clusters, as there seems to be only a little reduction in total within cluster sum of squares gained by increasing K to 8. The total within cluster sum of squares for the $K = 7$ model was 1225.079.

In Table 2.1, we see the seven clusters, and they each consist of predictors we already knew to be correlated. As we originally created four clusters, we performed K-means clustering with $K = 4$, this yields a total within cluster sum of squares of 5859.278. The resulting clusters seen in Table 2.2, unexpectedly, do not show the original four clusters. Changing the random seed in the generation of predictors shows similar results.

Clusters
Height, Height1, Height2, Height3
Weight, Weight2, Weight3
Weight1
Normal, Normal1
Normal2, Normal3
Uniform, Uniform1
Uniform2, Uniform3

Table 2.1: K-means clusters, when K is 7. K-means performed on scaled data with 16 predictors of 1000 observations.

Clusters
Height, Height1, Height2, Height3
Weight, Weight2, Weight3
Weight1, Uniform, Uniform1, Normal, Normal1
Normal2, Normal3, Uniform2, Uniform3

Table 2.2: K-means clusters, when K is 4. K-means performed on scaled data with 16 predictors of 1000 observations.

Note that when performing K-means clustering the initial assignment is random, thus we performed K-means clustering 20 times for the chosen K, and choose the clustering

with the least total within cluster sum of squares.

It is furthermore important to scale data prior to clustering. As an example we performed K-means clustering on unscaled data, with $K = 2$ based on an elbow plot with no visible change after 2. The result was a cluster containing only `Height3`, and another one containing the remaining variables. It is clear from Table 2.3 why `Height3` got clustered alone, as it is measured on a, compared to the others, extreme scale.

	`Height`	`Height1`	`Height2`	`Height3`	`Weight`	`Weight1`	`Weight2`	`Weight3`
Mean	160.36	481.00	22.88	4.52e+17	56.56	-7.00	452.30	16.08
	`Normal`	`Normal1`	`Normal2`	`Normal3`	`Uniform`	`Uniform1`	`Uniform2`	`Uniform3`
Mean	49.04	1471.40	-735.50	-3464.64	0.4984	9.951	-5.00	-0.49

Table 2.3: K-means clusters, when K is 4.

We conclude, that scaling is of the utmost importance. When data is scaled the predictors in each cluster are almost what we expected from the design of the predictors. □

K-means clustering could be useful in creating clusters of predictors in our real data. But as the elbow plot recommends seven clusters in this example, and we know there are four, the K-means method might be too conservative. On the positive side, each of the seven clusters did not contain predictors from more than one of the expected clusters, so K-means clustering seems to be conservative, but correct. To figure out whether K-means clustering is appropriate for our data we now investigate another clustering method for comparison.

2.1.2 Hierarchical Clustering

Hierarchical clustering does not require predetermining the choice of clusters K as was the case for K-means clustering. This section describes the most common type of hierarchical clustering called bottom-up clustering, which can be represented in the form of dendrograms. First, it is explained what a dendrogram is and how it is interpreted, after which it is explained how dendrograms are build.

At the bottom of the dendrogram, so-called leaves are seen, each representing a predictor, where each predictor has its own cluster. Examples of dendrograms can be seen in Figure 2.2. Above the leaves in the dendrogram it is seen that the leaves are merged and then called knots, which means that two predictors now belong to the same cluster. If predictors merge far down in the dendrogram, this means that these predictors are very similar and predictors merging their clusters at the top of the dendrogram may be quite different. Leaves and knots are fused until a single knot in the top of the dendrogram is achieved, where all predictors belong to the same knot/cluster. Since it is noninformative to have too many or too few clusters, the dendrogram has to be cut. A horizontal line in the dendrogram is made to indicate where the cut should be. Based on this cut, a number of clusters have been created.

In order to build a dendrogram, a dissimilarity measure between each pair of predictors is chosen and on the basis of this, clusters are merged. Typically, the Euclidean distance or correlation are chosen as dissimilarity measure. Let all predictors belong to their own cluster at the bottom of the dendrogram and merge the two clusters whose predictors are least dissimilar. This process can always be done for two clusters representing leaves, but in the case that one or both of the clusters represents a knot, then it is not obvious how to measure dissimilarity and therefore a so-called linkage measure is used instead. Two of the most common linkage measures are complete and average linkage. To use any of the two linkage measures, first all pairwise dissimilarities between a predictor in one cluster and a predictor in another cluster is measured. Complete linkage is the largest of the dissimilarities and average linkage is the average of all the dissimilarities.

Example

Applying hierarchical clustering with euclidean distance and complete linkage (and others) to the data set we previously used in the K-means example yields similar results as K-means. The elbow plots suggest seven clusters, and the seven clusters are the same as in Table 2.1. Setting the number of clusters to four, we again get the same clusters as in Table 2.2. A dendrogram of hierarchical clustering with euclidean distance and complete linkage can be seen in Figure 2.2 top left. As previously suggested, using correlation as dissimilarity measure could be beneficial as we are clustering variables. When we use correlation as dissimilarity measure, we get the four clusters that we originally designed, no matter the linkage method, a plot of a dendrogram based on hierarchical clustering with correlation as dissimilarity and complete linkage can be seen in Figure 2.2 top right. We expect there to be issues with both dissimilarity measure and linkage when we include binary predictors, thus we simulate two such. `Sex` is generated based on the already present variable `Height`, as we expect men to be taller than women, we have drawn the `Sex` predictor from a Bernoulli distribution with logit[`Height`] as the probabilities. We have also drawn the predictor `Smoke` from a Bernoulli distribution, this time with the probability of men smoking being 0.4 and the probability of women smoking being 0.6.

Now with the two binary predictors included they are vaguely suggested to be clustered alone when using euclidean distance. But when using correlation as a dissimilarity measure, the elbow plots do suggest that they each have their own cluster. Dendrograms of hierarchical clustering on the dataset including `Sex` and `Smoke` can be seen at the bottom of Figure 2.2.

When we set the number of clusters to four, `Sex` and `Smoke` are always clustered with the height variables no matter the choice of linkage method. One would expect the binary variables to be hard to cluster, as the observations have always (especially in euclidean distance) either the extreme lowest or extreme highest value. This is also what we saw in the example. Though correlation seems to handle the binary variables better, this could be due to the correlations over all better performance as dissimilarity measure. Note by the way that correlation is, as opposed to the other dissimilarity measures, invariant to scaling. □

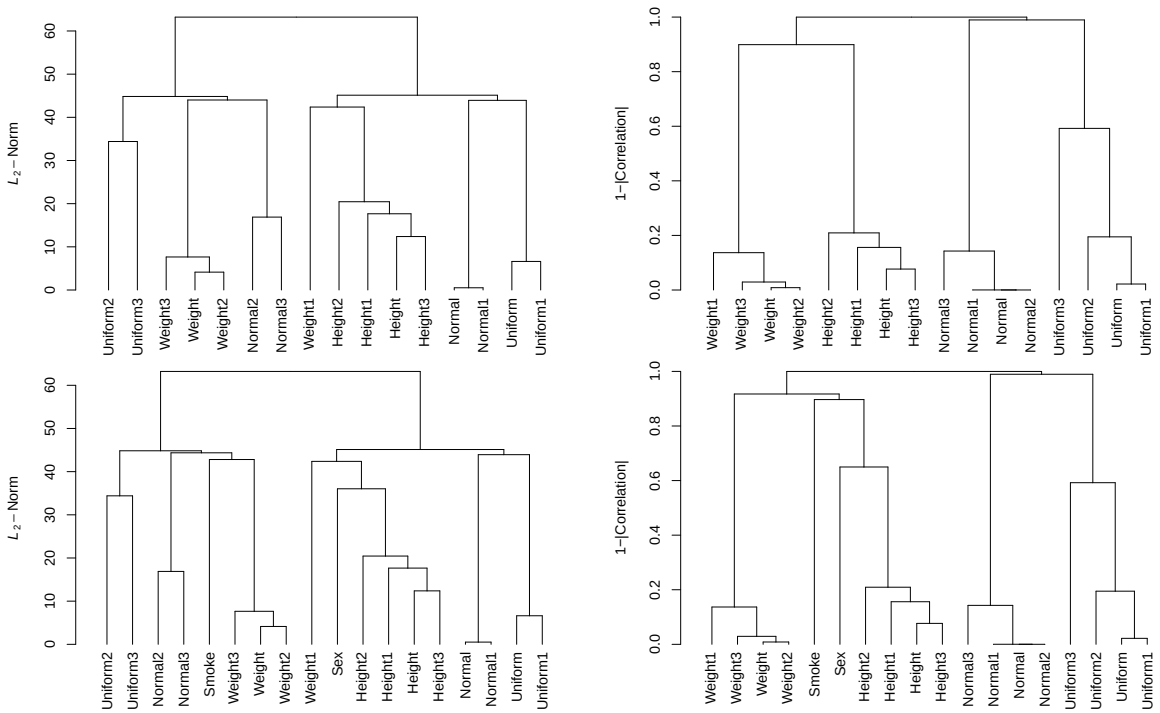


Figure 2.2: The lefts: Dendrograms produced by hierarchical clustering based on euclidean distance. **The rights:** Dendrograms produced by hierarchical clustering based on correlation. **The tops:** Dendrograms produced by hierarchical clustering of 16 continuous predictors. **The bottoms:** Dendrograms produced by hierarchical clustering of the 16 continuous plus two binary predictors.

We find that hierarchical clustering with correlation as dissimilarity measure outperforms K-means on our simulated data. This might not be the case if we found a way to use correlation as dissimilarity measure in K-means, but this would be outside the scope of this thesis. Though the choice of linkage seems unimportant, this could be due to both our limited number of predictors, or their highly correlated design, we thus expect to use different linkage methods on the real data. For the real data we expect to use hierarchical clustering over K-means clustering.

There are some problems with cluster methods. For the K-means clustering it is chosen how many clusters are desired and for hierarchical clustering there is a corresponding problem in selecting the number of clusters when the dendrogram is cut. In addition, one will likely get different dendrograms depending on the choice of dissimilarity measure as well as the type of linkage used. Several solutions should be considered as they may reveal interesting aspects of data. As mentioned earlier we expect to use the clustering information both in our descriptive analysis of data, but we are also interested in reducing the number of predictors, and a way to aggregate predictors in the same cluster could be principal component analysis.

2.2 Principal Component Analysis[1]

How our different classes relate to the covariates can be visually inspected by two way scatter plots, but this results in $\binom{p}{2}$ scatter plots, which will become unmanageable when p is large, as the method is subjective and each plot should be examined. Instead of having p predictors, some of which may be correlated, linear combinations of the original predictors can be made, such that we get fewer uncorrelated predictors. Principal component analysis deals with how these linear combinations are found such that the amount of the original variance retained from the data is maximized. Such a linear combination is called a principal component.

First, we clarify the notation. Let $\mathbf{X}$ be the random vector containing the random variables $X_1, X_2, \ldots, X_p$, of which $\mathbf{x}_{\bullet 1}, \mathbf{x}_{\bullet 2}, \ldots, \mathbf{x}_{\bullet p}$ are realizations. Furthermore let the data matrix X consist of realizations $\mathbf{x}_i = x_{i1}, x_{i2}, \ldots, x_{ip}$ for $i = 1, \ldots, n$ of $\mathbf{X}$.

Let the m'th principal component be denoted $\mathbf{z}_{\bullet m}$ and given by

$$\mathbf{z}_{\bullet m} = \phi_{1m}\mathbf{x}_{\bullet 1} + \phi_{2m}\mathbf{x}_{\bullet 2} + \cdots + \phi_{pm}\mathbf{x}_{\bullet p}, \tag{2.13}$$

where $\phi_{1m}, \phi_{2m}, \cdots \phi_{pm}$ are called the loadings, which are to be estimated. Note that the left subscript for the loadings indicate which original predictor $\mathbf{x}_{\bullet j}$ it belongs to and the right subscript refers to the principal component, in this case $\mathbf{z}_{\bullet m}$.

The score of the m'th principal component for observation i is written as

$$z_{im} = \phi_{1m}x_{i1} + \phi_{2m}x_{i2} + \cdots + \phi_{pm}x_{ip} = \boldsymbol{\phi}_m^\top \mathbf{x}_i. \tag{2.14}$$

To estimate $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_M$ for the respective principal components, the first principal component $z_{i1} = \boldsymbol{\phi}_1^\top \mathbf{x}_i$ is considered first, where it is desired that it captures as much variance as possible. Thus we find the $\boldsymbol{\phi}_1$ that fulfills

$$\arg\max_{\boldsymbol{\phi}_1} \left(\mathrm{Var}[\boldsymbol{\phi}_1^\top \mathbf{X}] \right) = \arg\max_{\boldsymbol{\phi}_1} \left(\boldsymbol{\phi}_1^\top \mathrm{Var}[\mathbf{X}] \boldsymbol{\phi}_1 \right) = \arg\max_{\boldsymbol{\phi}_1} \left(\boldsymbol{\phi}_1^\top \Sigma \boldsymbol{\phi}_1 \right). \tag{2.15}$$

Note that a maximum will only exist if $\boldsymbol{\phi}_1$ is constrained, as $\boldsymbol{\phi}_1$ else could be chosen arbitrarily large in order to achieve larger variance. Therefore a normalization of $\boldsymbol{\phi}_1$ is needed, which we choose to be $\boldsymbol{\phi}_1^\top \boldsymbol{\phi}_1 = 1$. Thus we have the optimization problem

$$\arg\max_{\boldsymbol{\phi}_1} \left(\boldsymbol{\phi}_1^\top \Sigma \boldsymbol{\phi}_1 \right) \quad \textit{subject to} \quad \boldsymbol{\phi}_1^\top \boldsymbol{\phi}_1 = 1. \tag{2.16}$$

In practice, the covariance matrix for $\mathbf{X}$ is not known and therefore it has to be estimated. It is estimated by the sample covariance matrix. When the original predictors have mean set to 0, the sample covariance is given by

$$\hat{\Sigma} = \frac{1}{n} X^\top X. \tag{2.17}$$

For the covariance matrix Σ, the entry (j, j') is given by $\mathrm{Cov}[X_j, X_{j'}]$, when $j \neq j'$ and given by $\mathrm{Var}[X_j]$, when $j = j'$.

[1] This section is based on [1, 39–41].

To maximize $f(\boldsymbol{\phi}_1) = \boldsymbol{\phi}_1^\top \Sigma \boldsymbol{\phi}_1$, ie. the variance of the first principal component subject to $g(\boldsymbol{\phi}_1) = \boldsymbol{\phi}_1^\top \boldsymbol{\phi}_1 = c$, where c is a constant (in our case 1), the technique of the Lagrange multiplier is used. To solve this optimization problem we first find the critical points for the function h_1 defined by

$$h_1(\boldsymbol{\phi}_1, \lambda) = f(\boldsymbol{\phi}_1) - \lambda \left(g(\boldsymbol{\phi}_1) - c\right) \tag{2.18}$$

$$= \boldsymbol{\phi}_1^\top \Sigma \boldsymbol{\phi}_1 - \lambda \left(\boldsymbol{\phi}_1^\top \boldsymbol{\phi}_1 - 1\right). \tag{2.19}$$

The function h_1 is called the Lagrangian and the new variable λ is called the Lagrange multiplier. We differentiate h_1 with respect to $\boldsymbol{\phi}_1$ and λ and set the resulting gradient equal to zero in order to find critical points. As differentiating with respect to λ results in $\boldsymbol{\phi}_1^\top \boldsymbol{\phi}_1 = 1$, which is the condition we previously specified, we now differentiate with respect to $\boldsymbol{\phi}_1$

$$\nabla_{\boldsymbol{\phi}_1} h_1(\boldsymbol{\phi}_1; \lambda) = \mathbf{0} \tag{2.20}$$

$$\nabla_{\boldsymbol{\phi}_1} \left(\boldsymbol{\phi}_1^\top \Sigma \boldsymbol{\phi}_1 - \lambda(\boldsymbol{\phi}_1^\top \boldsymbol{\phi}_1 - 1)\right) = \mathbf{0} \tag{2.21}$$

$$2\Sigma \boldsymbol{\phi}_1 - 2\lambda \boldsymbol{\phi}_1 = \mathbf{0}. \tag{2.22}$$

$$\left(\Sigma - \lambda \mathrm{I_p}\right) \boldsymbol{\phi}_1 = \mathbf{0}, \tag{2.23}$$

where $\mathrm{I_p}$ is the $p \times p$ identity matrix. We want to solve (2.23) with respect to $\boldsymbol{\phi}_1$ and λ and since the equation is on this form, λ is an eigenvalue of Σ with corresponding eigenvector $\boldsymbol{\phi}_1$. In order to determine which eigenvector of Σ gives the largest variance of the first principal component (2.15), it is first noted that due to (2.23) and that $\boldsymbol{\phi}_1$ is assumed to be a non-zero vector, then the following applies

$$\boldsymbol{\phi}_1^\top \Sigma \boldsymbol{\phi}_1 = \boldsymbol{\phi}_1^\top \lambda \mathrm{I_p} \boldsymbol{\phi}_1 = \lambda \boldsymbol{\phi}_1^\top \boldsymbol{\phi}_1 = \lambda. \tag{2.24}$$

Thus, maximizing the variance $\mathrm{Var}[\boldsymbol{\phi}_1^\top \mathbf{X}]$ corresponds to finding the largest eigenvalue λ of Σ, which has corresponding eigenvector $\boldsymbol{\phi}_1$, where the estimated $\boldsymbol{\phi}_1$ is used as the loadings of the first principal component. Later, λ will be denoted as λ_1 to indicate that it belongs to the first principal component.

In general, it can be shown that for the m'th principal component, λ_m is the m'th largest eigenvalue of Σ, where $\boldsymbol{\phi}_m$ is the corresponding eigenvector. This is proved for $m = 2$, because the case where $m \geq 3$ is very similar, but more complicated. It is desired that the second principal component $z_{i2} = \boldsymbol{\phi}_2^\top \mathbf{x}_i$ explains as much variation in data as possible, which was also wanted for the first principal component. Thus we want to maximize $\mathrm{Var}[Z_{i2}] = \boldsymbol{\phi}_2^\top \Sigma \boldsymbol{\phi}_2$ under the condition that it is uncorrelated with the first principal component $z_{i1} = \boldsymbol{\phi}_1^\top \mathbf{x}_i$. Another way of formulating that the first two principal components are uncorrelated can be derived by considering the covariance between the two

$$\mathrm{Cov}[Z_{i1}, Z_{i2}] = \mathrm{Cov}[\boldsymbol{\phi}_1^\top \mathbf{X}, \boldsymbol{\phi}_2^\top \mathbf{X}] = \boldsymbol{\phi}_1^\top \mathrm{Cov}[\mathbf{X}, \mathbf{X}] \boldsymbol{\phi}_2 = \boldsymbol{\phi}_1^\top \Sigma \boldsymbol{\phi}_2 \tag{2.25}$$

$$= \boldsymbol{\phi}_2^\top \Sigma \boldsymbol{\phi}_1 = \boldsymbol{\phi}_2^\top \lambda \boldsymbol{\phi}_1 = \lambda \boldsymbol{\phi}_2^\top \boldsymbol{\phi}_1 = \lambda \boldsymbol{\phi}_1^\top \boldsymbol{\phi}_2. \tag{2.26}$$

Thus another way of formulating that z_{i1} and z_{i2} are uncorrelated is that any of these equations are true

$$\boldsymbol{\phi}_1^\top \Sigma \boldsymbol{\phi}_2 = 0, \tag{2.27}$$
$$\boldsymbol{\phi}_2^\top \Sigma \boldsymbol{\phi}_1 = 0, \tag{2.28}$$
$$\boldsymbol{\phi}_1^\top \boldsymbol{\phi}_2 = 0, \tag{2.29}$$
$$\boldsymbol{\phi}_2^\top \boldsymbol{\phi}_1 = 0. \tag{2.30}$$

Note that the case of $\lambda = 0$ would also result in z_{i1} and z_{i2} being uncorrelated, but as $\lambda = 0$ would mean that the largest eigenvalue is 0, thus all eigenvalues are 0, which indicates that the covariance matrix is equal to the zero-matrix, representing an uninteresting special case.

It is chosen to use the equation $\boldsymbol{\phi}_2^\top \boldsymbol{\phi}_1 = 0$ to provide no correlation between the first two principal components, which is an arbitrary choice and one might as well have chosen one of the other equations. Furthermore, it is noted that the maximum variance, as was the case for $\boldsymbol{\phi}_1$, cannot be found for the second principal component unless $\boldsymbol{\phi}_2$ is constrained. Thus it is chosen to normalize by $\boldsymbol{\phi}_2^\top \boldsymbol{\phi}_2 = 1$. To summarize, we will maximize $f(\boldsymbol{\phi}_2) = \boldsymbol{\phi}_2^\top \Sigma \boldsymbol{\phi}_2$ subject to $g(\boldsymbol{\phi}_2) = \boldsymbol{\phi}_2^\top \boldsymbol{\phi}_2 = c$ and $\tilde{g}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = \boldsymbol{\phi}_2^\top \boldsymbol{\phi}_1 = \tilde{c}$, where c and $\tilde{c}$ are constants. As with the first principal component, the technique of the Lagrange multiplier is used to find critical points for the function h_2 defined by

$$h_2(\boldsymbol{\phi}_2; \boldsymbol{\phi}_1, \bar{\lambda}, \tilde{\lambda}) = f(\boldsymbol{\phi}_2) - \bar{\lambda}\left(g(\boldsymbol{\phi}_2) - c\right) - \tilde{\lambda}\left(\tilde{g}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) - \tilde{c}\right) \tag{2.31}$$
$$= \boldsymbol{\phi}_2^\top \Sigma \boldsymbol{\phi}_2 - \bar{\lambda}\left(\boldsymbol{\phi}_2^\top \boldsymbol{\phi}_2 - 1\right) - \tilde{\lambda} \boldsymbol{\phi}_2^\top \boldsymbol{\phi}_1, \tag{2.32}$$

where $\bar{\lambda}$ and $\tilde{\lambda}$ are Lagrange multipliers. In order to find critical points for the Lagrangian h_2 with respect to $\boldsymbol{\phi}_2$, the gradient is found and set equal to the zero vector

$$\nabla_{\boldsymbol{\phi}_2} h_2(\boldsymbol{\phi}_2, \boldsymbol{\phi}_1, \bar{\lambda}, \tilde{\lambda}) = 2\Sigma \boldsymbol{\phi}_2 - 2\bar{\lambda} \boldsymbol{\phi}_2 - \tilde{\lambda} \boldsymbol{\phi}_1 = \mathbf{0}. \tag{2.33}$$

multiplying this by $\boldsymbol{\phi}_1^\top$ from the left yields

$$2\boldsymbol{\phi}_1^\top \Sigma \boldsymbol{\phi}_2 - 2\bar{\lambda} \boldsymbol{\phi}_1^\top \boldsymbol{\phi}_2 - \tilde{\lambda} \boldsymbol{\phi}_1^\top \boldsymbol{\phi}_1 = 0. \tag{2.34}$$

Since the first two terms are zero due to (2.27) and (2.29), and the fact that the normalization $\boldsymbol{\phi}_1^\top \boldsymbol{\phi}_1 = 1$, we have that $\tilde{\lambda} = 0$. Now consider (2.33), which is now reduced to $\Sigma \boldsymbol{\phi}_2 - \bar{\lambda} \boldsymbol{\phi}_2 = 0$. This can be rewritten as

$$\left(\Sigma - \bar{\lambda} \mathrm{I_p}\right) \boldsymbol{\phi}_2 = \mathbf{0}, \tag{2.35}$$

which means that $\bar{\lambda}$ is also an eigenvalue of Σ with corresponding eigenvector $\boldsymbol{\phi}_2$. Similarly, as shown for the first principal component in (2.24), $\bar{\lambda} = \boldsymbol{\phi}_2^\top \Sigma \boldsymbol{\phi}_2$ and $\bar{\lambda}$ must therefore be as large as possible to maximize the variance. The two eigenvalues $\bar{\lambda}$ and λ_1 can not be equal to each other because it will cause $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$, which is in violation with $\boldsymbol{\phi}_1^\top \boldsymbol{\phi}_2 = 0$ according to (2.29). Thus, $\bar{\lambda}$ must necessarily be the second largest eigenvalue of Σ, where $\boldsymbol{\phi}_2$ is the corresponding eigenvector and therefore $\bar{\lambda}$ is denoted as λ_2. As

previously mentioned, this can be extended to the m'th principal component having the m'th largest eigenvalue λ_m of Σ, where $\boldsymbol{\phi}_m$ is the corresponding eigenvector. Note that the m'th principal component is found under the constraint that it be uncorrelated with the previous $m-1$ principal components.

After the M principal components are found, the $\binom{M}{2}$ scatter plots can be considered to examine data through these fewer predictors and in addition, these principal components can be used instead of the original predictors to perform supervised learning in Chapter 3. To determine how many principal components to be made, it is investigated how much variance is lost in performing PCA. The total variance in data after centering is defined by

$$\sum_{j=1}^{p} \text{Var}[X_j] = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 \tag{2.36}$$

and the variance for the m'th principal component is given by

$$\text{Var}[Z_m] = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2 . \tag{2.37}$$

Thus, the m'th principal component represents the following proportion of the total variance

$$\frac{\text{Var}[Z_m]}{\sum_{j=1}^{p} \text{Var}[X_j]} = \frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}. \tag{2.38}$$

It can be investigated how much variance the first m principal components represent of the total variance by cumulating these values and thus it can be investigated how much more variance an extra principal component will contribute.

Example

We first conduct PCA on the simulated data set from the clustering examples, which currently consist of 1000 observations and 18 predictors. Consider the scree plot at the top left corner of Figure 2.3. A scree plot is a plot indicating how much variance is explained by each of the m principal components. We only show the first 10 principal components at the top left corner of Figure 2.3, since not much variance is explained by the remaining principal components. Note the big change from `PC4` to `PC5`. This indicates that four principal components could be enough to explain most of the variance in the data. This becomes further apparent from Table 2.4. Here we see that beyond `PC13`, no further variance is explained, and the cumulative proportion of variance is one, this is of course a consequence of rounding as a cumulative proportion of variance should not reach one before the last principal component (if none of the predictors in the data set are co-linear).

We conclude that four principal components seem to explain most of the variation, this is what we expected as we only have four underlying effects.

Choosing four principal components, we have examined the $\binom{4}{2}$ bi-plots. A bi-plot is a combination of two plots, the first being a scatter plot of the scores (2.14) from two

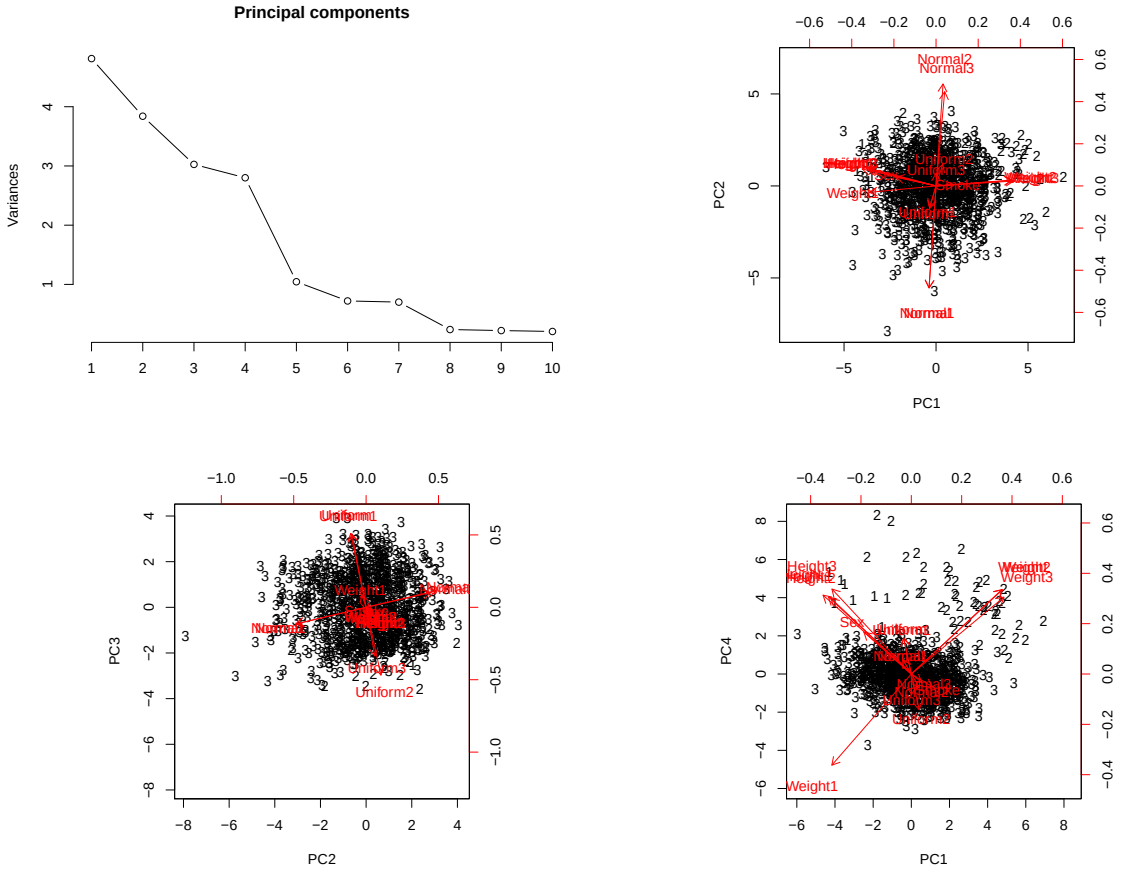


Figure 2.3: Scree-plot of the explained variance of each principal component and selected bi-plots from a principal component analysis of all 18 variables in the simulated data set. In the bi-plots the loadings are on the right and top borders and the scores on the left and lower borders. The red arrows represent the original variables respective loadings and the black numbers the transformed observations, where the number indicates the class to which the observation belongs.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Proportion of Variance	0.27	0.21	0.17	0.16	0.06	0.04	0.04	0.01	0.01
Cumulative Proportion	0.27	0.48	0.65	0.80	0.86	0.90	0.94	0.95	0.97
	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
Proportion of Variance	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Cumulative Proportion	0.98	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00

Table 2.4: The variance explained by the principal components, derived from a principal component analysis of all 18 variables in the simulated data set.

principal components and thus represents the transformed observations, the other plot is of the loadings from the same two principal components and thus represents the principal components relation to the original predictors. We present three of the $\binom{4}{2}$ bi-plots in Figure 2.3. Here we see in the top right plot that the height and weight variables align with `PC1`, the normal variables with `PC2` and the uniform variables seem unaffected by `PC1` and `PC2`. In the bottom left bi-plot of Figure 2.3 `PC2` and `PC3` seem to distinguish the uniforms and the normals with little explanation of the weights and heights. Furthermore note that the three classes are in the two first bi-plots not easily distinguished. Examining the lower right bi-plot of `PC1` and `PC4`, we see that `PC1` and `PC4` mainly explains the weight and height variables, and thus it is possible to distinguish the three classes.

Combining hierarchical clustering from Section 2.1.2 and principal component analysis, we can perform PCA on each of the four clusters: heights (including `Sex` and `Smoke`), weights, uniforms and normals.

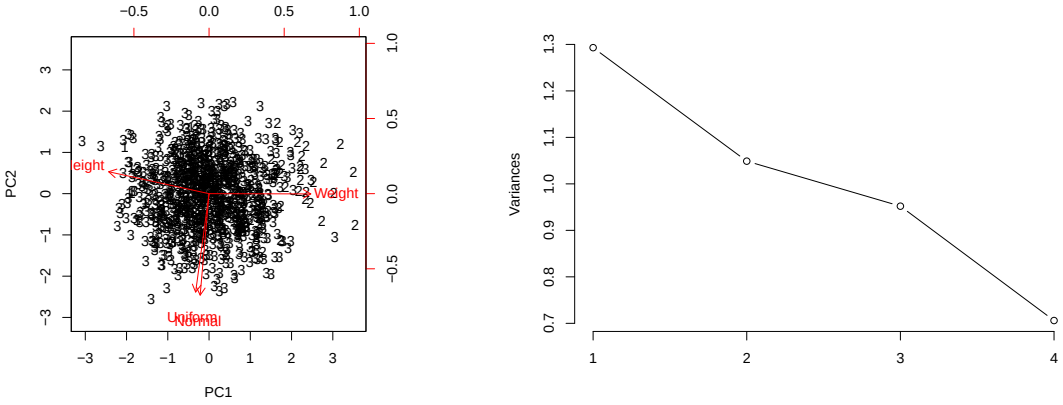


Figure 2.4: Bi-plot of the two first principal components and scree-plot, both based on only the latent variables `Height`, `Weight`, `Uniform` and `Normal`.

As we saw in Figure 2.3 the weights and heights seem to be explained by different principal components than the uniforms and normals. We have performed PCA on just the four main effects, and see in Figure 2.4, that the first pricipal component is closely related to `Height` and `Weight`, and the second principal component seems closely related to `Uniform` and `Normal`. As we know that the principal components are perpendicular, this might suggest that `Height` and `Weight` explain different aspects than `Uniform` and `Normal`.

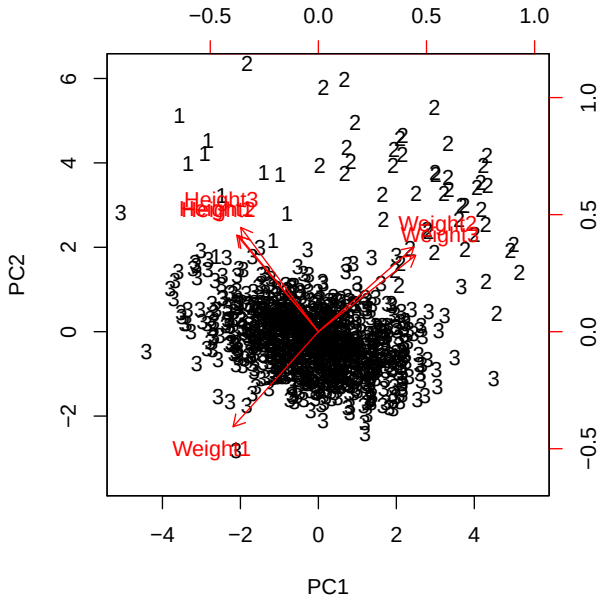


Figure 2.5: Bi-plot of the first two principal components from a principal component analysis of the weight and height variables (without `Weight` and `Height`).

Using only the generated variables `Height1`, `Height2`, `Height3`, `Weight1`, `Weight2` and `Weight3`, and not the "latent" variables `Height` and `Weight` themselves, we can separate the three classes quite well, thus representing the two latent variables by two principal components based on the generated variables. A bi-plot of this separation is seen in Figure 2.5.

As in previous examples we have tried to use unscaled data, and as previously it results in one principal component based primarily on `Height3` explaining so much variance that the remaining principal components become unimportant. □

As several of the predictors in our real data set are expected to explain different aspects of the same latent variable, we will use both hierarchical clustering and PCA to try and aggregate some of these. We expect a large number of predictors to induce overfitting, and thus the reduced number of variables could produce better predictions.

3. Classification Methods

As this project deals with a categorical response variable, this chapter presents various approaches to predicting such a response, called classification. Therefore, when an observation (individual) is predicted as belonging to a class, it is said that the observation is classified. A particular classification method is called a classifier and in this chapter some of the most common classifiers are presented. Special attention is given to logistic regression, as this classifier assigns to each individual a probability that the individual belongs to a certain class. The calculation of a probability expands our options when we have to choose an evaluation measure. It further holds information of the certainty of a given classification. We expect all these attributes to be useful as we have low incidences of ADHD and ASD.

A model is built on the basis of given training data. It is desired that a built model performs well on new data, that is, data which has not been used to build the model. In order to evaluate how good the performance is, evaluation measures are used, these are introduced in Chapter 4. When building a model based on training data, it is important that the model is not overfitted to the training data, as this can lead to perfect predictions for training data based on evaluation measures, but makes the model perform poorly on new data. In other words, a sensible bias-variance trade-off is sought. The first section in this chapter deals with this bias-variance trade-off before several classifiers are presented in the remainder of this chapter.

3.1 The Bias-Variance Trade-Off

Assume we are given some training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. Given that y_i is a realisation of the random variable Y_i, then the relation between $\mathbf{x}_i$, $i = 1, \ldots, n$ and Y_i can generally be expressed by

$$Y_i = f(\mathbf{x}_i) + \varepsilon_i, \tag{3.1}$$

where f is a fixed unknown function of $\mathbf{x_i}$ and ε_i is a random error term, which is independent of $\mathbf{x}_i$. It is assumed that the error term has mean 0 and variance σ^2 and thus the random variable Y_i can be estimated by

$$\hat{y}_i = \hat{f}(\mathbf{x}_i), \tag{3.2}$$

where $\hat{f}$ is the estimate of f and $\hat{y}_i$ is the estimation of y_i, where y_i is not itself used in the estimation. The purpose of $\hat{f}$ is to make accurate predictions for y_i in a predictive study. Note that Y_i is independent of the observations used to build $\hat{f}$, making Y_i and $\hat{f}$ independent too. When $\hat{f}$ is computed, the average prediction error for individual i is given by

$$\begin{aligned}
\mathbb{E}[(Y_i - \hat{f}(\mathbf{x}_i))^2] &= \mathbb{E}[Y_i^2] + \mathbb{E}[\hat{f}(\mathbf{x}_i)^2] - 2\mathbb{E}[Y_i\hat{f}(\mathbf{x}_i)] && (3.3)\\
&= \mathrm{Var}[Y_i] + \mathbb{E}[Y_i]^2 + \mathrm{Var}[\hat{f}(\mathbf{x}_i)] + \mathbb{E}[\hat{f}(\mathbf{x}_i)]^2 - 2f(\mathbf{x}_i)\mathbb{E}[\hat{f}(\mathbf{x}_i)] && (3.4)\\
&= \mathrm{Var}[Y_i] + \mathrm{Var}[\hat{f}(\mathbf{x}_i)] + \Big(f(\mathbf{x}_i)^2 + \mathbb{E}[\hat{f}(\mathbf{x}_i)]^2 - 2f(\mathbf{x}_i)\mathbb{E}[\hat{f}(\mathbf{x}_i)]\Big) && (3.5)\\
&= \mathrm{Var}[Y_i] + \mathrm{Var}[\hat{f}(\mathbf{x}_i)] + \Big(f(\mathbf{x}_i) - \mathbb{E}[\hat{f}(\mathbf{x}_i)]\Big)^2 && (3.6)\\
&= \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2] + \mathrm{Var}[\hat{f}(\mathbf{x}_i)] + \Big(f(\mathbf{x}_i) - \mathbb{E}[\hat{f}(\mathbf{x}_i)]\Big)^2 && (3.7)\\
&= \sigma^2 + \mathrm{Var}[\hat{f}(\mathbf{x}_i)] + \mathrm{Bias}[\hat{f}(\mathbf{x}_i)]^2, && (3.8)
\end{aligned}$$

where the term σ^2 in (3.8) is called the irreducible term as it is the error that cannot be reduced when estimating f. The last two terms are called the reducible terms as they can be reduced based on the statistical method used to make the estimation. It is desired to decrease both of the reducible terms and since one of these increases when the other decreases and vice versa, this is known as the bias-variance trade-off. The variance of $\hat{f}$ indicates how much $\hat{f}$ will change if another training data set is used to estimate it. Thus, more flexible statistical models, e.g. models with more parameters , will have high variance, as the model could be fit to the noise. Bias describes how well the fitted model actually matches the data. For example, high bias can be obtained by linear models if the data is not linear. Thus, flexible models generally have little bias. All in all, a sufficiently flexible model must be found on the basis of the bias-variance trade-off in order to get the best prediction for the test data. With bias-variance trade-off in mind, we are now ready to present different classifiers.

3.2 Bayes Classifier

The goal of a classifier is to assign an individual to a certain class $c \in \{1, \ldots, K\}$, for K disjoint classes.

A simple way to classify an observation is through the Bayes classifier, which assigns y_i to the class c for which $P(Y_i = c \mid \mathbf{x}_i)$ i largest

$$\hat{y}_i = \underset{c\in\{1,\ldots,K\}}{\arg\max}\, P(Y_i = c \mid \mathbf{X} = \mathbf{x}_i). \tag{3.9}$$

The reason we do not simply use the Bayes classifier is that in order to calculate the conditional probability of $P(Y_i = c \mid \mathbf{x}_i)$, we need to know the distributions of all variables in $Y \mid X$, and since the underlying distributions are unknown for the real data, we go through several classifiers. The distribution is not unknown for the simulated data and therefore the probability of belonging to each class can be calculated using Bayes' theorem

$$P(Y_i = c \mid \mathbf{x}_i) = \frac{P(Y_i = c)}{P(\mathbf{x}_i)} P(\mathbf{x}_i \mid Y_i = c), \tag{3.10}$$

where $P(\mathbf{x}_i \mid Y_i = c) \equiv f_c(\mathbf{x}_i)$ is the density function of X for an observation that comes from the c'th class.

Example

The simulated data set with the two predictors `Weight` and `Height` is considered in an example. As we know the distribution of `1percent`, `5percent`, and `94percent`, it is possible to calculate the density functions explicitly and thus utilize the Bayes classifier. Figure 3.1 shows four plots, where the two at the top and the bottom left show the binary classification problem, where the probability functions are set equal to each other to make the decision boundaries, that is

$$P(Y_i = c \mid \mathbf{x}_i) = P(Y_i = c' \mid \mathbf{x}_i) \tag{3.11}$$

$$P(Y_i = c) f_c(\mathbf{x}_i) = P(Y_i = c') f'_c(\mathbf{x}_i), \tag{3.12}$$

where $P(Y_i = c)$ and $P(Y_i = c')$ are estimated based on the proportion drawn from each class. for example for the `1percent` class the probability is equal to $\frac{10}{1000}$ in Figure 3.1. The plot at the bottom right shows the same decisions lines, but where the lines are approximated based on where it is inconclusive which probability function gives the highest value.

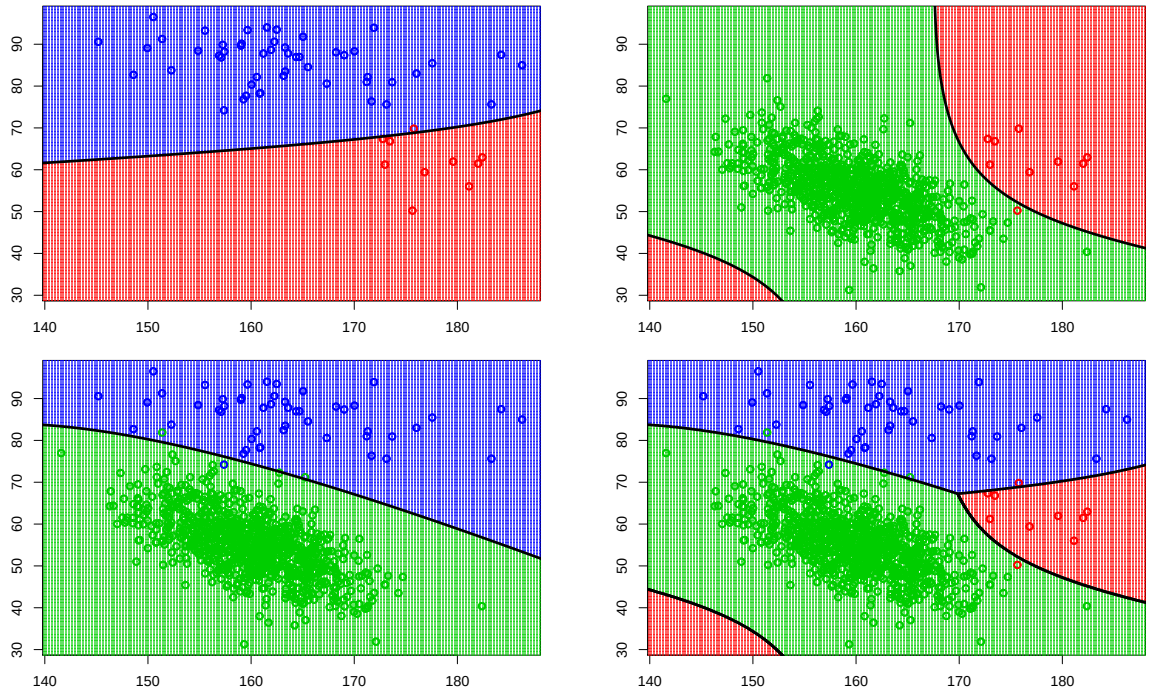


Figure 3.1: Top left, top right and bottom left are the Bayes classifier applied to each of our classes against each other respectively. The bottom right is all classes classified simultaneously. The circles are from our simulated data set, and drawn from the distributions on which the Bayes decision areas are defined.

The top right plot in figure 3.1 has a red area in the lower left corner, but none of the points in the simulated data are near this area. Therefore, an example is made where the

number of observations for the red class are increased to 1000. In Figure 3.2, it is seen more clearly that the distribution for red class is drawn closer to the red corner. This makes sense as the ten originally drawn observations seem to come near the mean of the distribution.

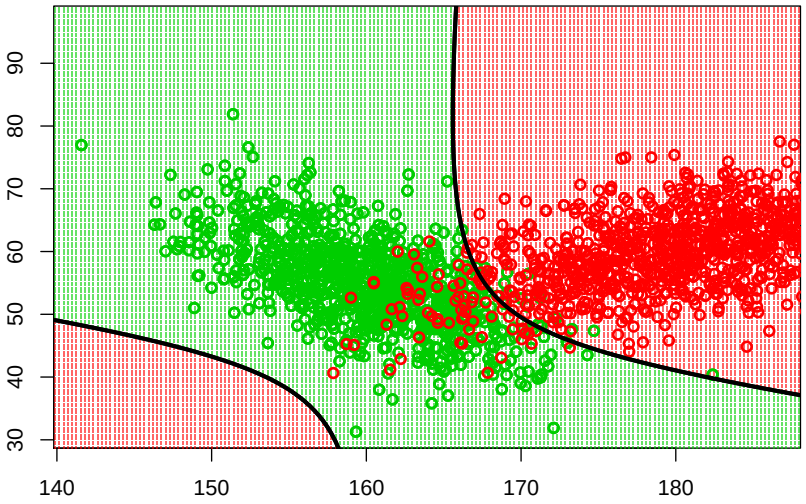


Figure 3.2: Our simulated data set with 990 further observations drawn from the `1percent` class.

□

3.3 Generalized Linear Models[1]

As mentioned in the introduction to this chapter, extra attention will be paid to logistic regression in this master thesis. A logistic regression model is a specific generalized linear model (GLM). Thus, GLM is presented before specifying logistic regression.. We define GLMs in Section 3.3.5, one of several requirements for a GLM is that the response should have a density from the so-called exponential family.

3.3.1 Exponential Family

Let $\mathbf{Y}$ be the random vector containing the independent identically distributed (i.i.d.) random responses Y_i for $i = 1, 2, \ldots, n$, of which y_i is a realization. If the Y_i's have marginal densities with respect to the Lebesgue measure of the form

$$f\left(y_i; \theta_i, \phi_i\right) = h\left(y_i; \phi_i\right) \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right), \tag{3.13}$$

then it is said that the Y_i's have a distribution from the exponential family. The parameter θ_i is called the canonical parameter, $b(\cdot)$ and $h(\cdot)$ are functions, which are specific for a certain exponential family (e.g. normal, Bernoulli, ...), while $\phi_i > 0$ is a dispersion parameter that can be known or unknown.

[1] This section is based on [42, 43].

Example

As this master thesis has a categorical response, it is shown that the Binomial and the Bernoulli distributions belong to the exponential family. The probability function for a binomially distributed random variable $X \sim \texttt{Bin}(n, p)$ may be written as

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{3.14}$$

$$= \exp\left(\log\left(\binom{n}{x} p^x (1-p)^{n-x}\right)\right) \tag{3.15}$$

$$= \exp\left(\log\binom{n}{x} + \log\left(p^x\right) + \log\left((1-p)^{n-x}\right)\right) \tag{3.16}$$

$$= \binom{n}{x} \exp\left(x\log(p) + (n-x)\log(1-p)\right) \tag{3.17}$$

$$= \binom{n}{x} \exp\left(x\left(\log(p) - \log(1-p)\right) + n\log(1-p)\right) \tag{3.18}$$

$$= \binom{n}{x} \exp\left(x\log\left(\frac{p}{1-p}\right) + n\log(1-p)\right), \tag{3.19}$$

where $x = 1, 2, \ldots, n$ indicates the number of successes. If we set $\theta = \log\left(\frac{p}{1-p}\right)$ then isolating p yields

$$\exp(\theta) = \frac{p}{1-p} \tag{3.20}$$

$$\exp(\theta) - p\exp(\theta) = p \tag{3.21}$$

$$\frac{\exp(\theta)}{p} = 1 + \exp(\theta) \tag{3.22}$$

$$p = \frac{\exp(\theta)}{1+\exp(\theta)}. \tag{3.23}$$

Thus, it is seen that the binomial distribution belongs to the exponential family as (3.19) can be written as (3.13) by setting $h(x) = \binom{n}{x}$, $\theta = \log\left(\frac{p}{1-p}\right)$, $b(\theta) = -n\log(1-p)$, $\phi_i = 1$.

When $n = 1$ only one person is considered, which means that the binomial distribution reduces to the Bernoulli distribution and thus the Bernoulli distribution also belongs to the exponential family. □

3.3.2 Mean and Variance of the Exponential Family

Let Y_i be a random variable with distribution from the exponential family. Subject to certain regularity conditions, the mean and variance of Y_i are given by

$$\mathbb{E}[Y_i] = \mu_i = b'(\theta_i) \tag{3.24}$$

$$\text{Var}[Y_i] = b''(\theta_i)\phi_i. \tag{3.25}$$

Proof. We start by proving (3.24) and use that the mean of a random variable is defined by $\mathbb{E}[Y_i] = \int y_i f(y_i) dy_i$ and that the integral of a density function equals 1, that is $1 = \int f(y_i; \theta_i, \phi_i) dy_i$. Differentiating the latter with respect to θ_i yields

$$
\begin{aligned}
0 &= \frac{d}{d\theta_i} \int f(y_i; \theta_i, \phi_i) dy_i && (3.26)\\
&= \int \frac{d}{d\theta_i} f(y_i; \theta_i, \phi_i) dy_i && (3.27)\\
&= \int \frac{d}{d\theta_i} h(y_i, \phi_i) \exp\left(\frac{y_i \theta_i - b(\theta_i)}{\phi_i} \right) dy_i && (3.28)\\
&= \int h(y_i, \phi_i) \exp\left(\frac{y_i \theta_i - b(\theta_i)}{\phi_i} \right) \left(\frac{y_i - b'(\theta_i)}{\phi_i} \right) dy_i && (3.29)\\
&= \int f(y_i; \theta_i, \phi_i) \left(\frac{y_i - b'(\theta_i)}{\phi_i} \right) dy_i && (3.30)\\
&= \frac{1}{\phi_i} \left(\int y_i f(y_i; \theta_i, \phi_i) dy_i - b'(\theta_i) \int f(y_i; \theta_i, \phi_i) dy_i \right) && (3.31)\\
&= \frac{1}{\phi_i} \left(\mathbb{E}[Y_i] - b'(\theta_i) \right), && (3.32)
\end{aligned}
$$

which proves (3.24). We now prove (3.25) by using that the variance of a random variable Y_i is defined by $\mathrm{Var}[Y_i] = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2]$. Once again it is used that the integral of a density function is 1, but this time the expression is differentiated twice with respect to θ_i.

$$
\begin{aligned}
0 &= \frac{d^2}{d\theta_i^2} \int f(y_i; \theta_i, \phi_i) dy_i && (3.33)\\
&= \int \frac{d}{d\theta_i} \left[\frac{y_i - b'(\theta_i)}{\phi_i} f(y_i; \theta_i, \phi_i) \right] dy_i && (3.34)\\
&= \int \frac{-b''(\theta_i)}{\phi_i} f(y_i; \theta_i, \phi_i) + \left(\frac{y_i - b'(\theta_i)}{\phi_i} \right)^2 f(y_i; \theta_i, \phi_i) dy_i && (3.35)\\
&= \frac{-b''(\theta_i)}{\phi_i} \int f(y_i; \theta_i, \phi_i) dy_i + \frac{1}{\phi_i^2} \int (y_i - \mu_i)^2 f(y_i; \theta_i, \phi_i) dy_i && (3.36)\\
&= \frac{-b''(\theta_i)}{\phi_i} + \frac{1}{\phi_i^2} \mathrm{Var}[Y_i] && (3.37)\\
\Rightarrow 0 &= -b''(\theta_i) + \frac{1}{\phi_i} \mathrm{Var}[Y_i], && (3.38)
\end{aligned}
$$

which proves (3.25). ■

Note that the mean and variance of a random variable with a distribution from the exponential family are determined by the function $b(\cdot)$. As the variance is always non-negative, $b'(\theta_i)$ is a monotone function and thus there exists an inverse on its image. Isolating θ_i in (3.24) yields

$$
\theta_i = b'^{-1}(\mu_i) \tag{3.39}
$$

and inserting this term in (3.25) leads to the variance becoming a function of the mean μ_i

$$\text{Var}[Y_i] = \phi_i b''(\theta_i) = \phi_i b''(b'^{-1}(\mu_i)) = \phi_i \mathcal{V}(\mu_i), \tag{3.40}$$

where $\mathcal{V}(\mu_i) = b''(b'^{-1}(\mu_i))$ is called the variance function.

Example

In this example, we show that (3.24) and (3.25) hold for the Bernoulli distribution. Let $Y_i \sim \texttt{Bern}(p_i)$. Then the mean of the Bernoulli distribution is given by $\mathbb{E}[Y_i] = p_i$ and the variance is given by $\text{Var}[Y_i] = p_i(1 - p_i)$. Thus it is shown that $b'(\theta_i) = p_i$ and $\phi_i b''(\theta_i) = p_i(1 - p_i)$. We first show that $b'(\theta_i) = p_i$, where we use that $b(\theta_i) = \log(1 - p_i)$ and $p_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$ from the previous Bernoulli distribution example

$$\begin{aligned}
b'(\theta_i) &= -\frac{d}{d\theta_i} \log(1 - p_i) && (3.41)\\
&= -\frac{d}{d\theta_i} \log\left(1 - \frac{\exp(\theta_i)}{1 + \exp(\theta_i)}\right) && (3.42)\\
&= -\frac{d}{d\theta_i} \log\left(\frac{1}{1 + \exp(\theta_i)}\right) && (3.43)\\
&= \frac{d}{d\theta_i} \log\left(1 + \exp\left(\theta_i\right)\right) && (3.44)\\
&= \frac{1}{1 + \exp(\theta_i)} \exp(\theta_i) && (3.45)\\
&= p_i. && (3.46)
\end{aligned}$$

It is now shown that $\phi_i b''(\theta_i) = \text{Var}[Y_i]$, where $\phi_i = 1$ for the Bernoulli distribution

$$\begin{aligned}
b''(\theta_i) &= \frac{d}{d\theta_i} b'(\theta_i) && (3.47)\\
&= \frac{d}{d\theta_i} \frac{1}{1 + \exp(\theta_i)} \exp(\theta_i) && (3.48)\\
&= \frac{-\exp(\theta_i)}{\left(1 + \exp(\theta_i)\right)^2} \exp(\theta_i) + \frac{1}{1 + \exp(\theta_i)} \exp(\theta_i) && (3.49)\\
&= \frac{-\exp(\theta_i)^2 + \exp(\theta_i)\left(1 + \exp(\theta_i)\right)}{\left(1 + \exp(\theta_i)\right)^2} && (3.50)\\
&= \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} \left(\frac{-\exp(\theta_i)}{1 + \exp(\theta_i)} + \frac{1 + \exp(\theta_i)}{1 + \exp(\theta_i)}\right) && (3.51)\\
&= p_i(1 - p_i). && (3.52)
\end{aligned}$$

Thus (3.24) and (3.25) hold for the Bernoulli distribution. Finally, the variance function of the Bernoulli distribution is given by

$$\mathcal{V}(\mu_i) = \text{Var}[Y_i] = p_i(1 - p_i) = \mu_i(1 - \mu_i). \tag{3.53}$$

□

To summarize all distributions from the exponential family have their own specific form of $b(\theta_i)$, which determines the mean, variance and variance function $\mathcal{V}(\mu_i)$. To define a GLM we need a variance function, a linear predictor and a link function. Thus, the latter two are presented in the next section.

3.3.3 The Linear Predictor and the Link Function

The linear predictor η_i for observation i is a linear function of the covariates $\mathbf{x}_i$

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}. \tag{3.54}$$

The link function g describes how the linear predictor η_i is associated with the mean μ_i

$$g(\mu_i) = \eta_i. \tag{3.55}$$

Thus the mean is not necessarily a linear combination of the covariates, as we are used to in the linear normal models where $\eta_i = \mu_i$. Assume that g is bijective such that g^{-1} exists. Then the mean can be seen as a function of the linear predictor, that is

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right). \tag{3.56}$$

We now have two expressions containing the mean; $\mu_i = b'(\theta_i)$ from (3.24) and $\eta_i = g(\mu_i)$ from (3.55) which combined lead to

$$\eta_i = g\left(b'(\theta_i)\right). \tag{3.57}$$

The link function $g(\cdot)$ that causes the canonical parameter θ_i to be equal to the linear predictor η_i is called the canonical link function, that is $g(\mu_i) = \eta_i = \theta_i$.

Example

In this example it is shown that the canonical link function for the Bernoulli distribution is the logit function. It is known that $\mathbb{E}[Y_i] = \mu_i = p_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$ from (3.23) and since the following applies, the logit transformation is the canonical link for the Bernoulli distribution

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{\frac{\exp(\theta_i)}{1+\exp(\theta_i)}}{1-\frac{\exp(\theta_i)}{1+\exp(\theta_i)}}\right) = \log\left(\frac{\frac{\exp(\theta_i)}{1+\exp(\theta_i)}}{\frac{1+\exp(\theta_i)-\exp(\theta_i)}{1+\exp(\theta_i)}}\right) = \theta_i. \tag{3.58}$$

Note that the canonical link function is not necessarily the one that best captures the mean structure for a given data set, it is just the link function that fulfills $\eta_i = \theta_i$. □

3.3.4 The Exponential Family Density Parameterized Relative to μ_i

For GLMs, $\boldsymbol{\beta}$ is estimated from the linear predictor $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ based on data. We wish to apply maximum likelihood to estimate $\boldsymbol{\beta}$. Since $\boldsymbol{\beta}$ is not an obvious part of the density

in (3.13), we now parameterize the density with respect to $\mu_i = g^{-1}\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)$.

The observations y_i may be compared to the parameter values μ_i by using the unit deviance defined by

$$d(y_i, \mu_i) = 2\int_{\mu_i}^{y_i} \frac{y_i - u}{\mathcal{V}(u)} du, \tag{3.59}$$

where $\mathcal{V}(\cdot)$ is the variance function.

It is now proved that the density of the exponential family can be parameterized in relation to μ_i using the unit deviance in

$$f(y_i; \mu_i, \phi_i) = a(y_i, \phi_i) \exp\left(-\frac{1}{2\phi_i} d(y_i, \mu_i)\right), \tag{3.60}$$

where $a(y_i, \phi_i)$ is a different function from h in (3.13).

Proof. Recall from (3.40) that the variance function is of the form

$$\mathcal{V}(u) = b''(b'^{-1}(u)) \tag{3.61}$$

Furthermore let $v = b'^{-1}(u)$, which yields $u = b'(v)$. Differentiating u with respect to v yields

$$\frac{d}{dv} u = \frac{d}{dv} b'(v) \tag{3.62}$$

$$du = b''(v) dv. \tag{3.63}$$

We now use integration by substitution to rewrite the unit deviance from (3.59)

$$\begin{aligned}
d(y_i, \mu_i) &= 2\int_{\mu_i}^{y_i} \frac{y_i - u}{\mathcal{V}(u)} du && (3.64)\\
&= 2\int_{\mu_i}^{y_i} \frac{y_i - u}{b''(b'^{-1}(u))} du && (3.65)\\
&= 2\int_{b'^{-1}(\mu_i)}^{b'^{-1}(y_i)} \frac{y_i - b'(v)}{b''(v)} b''(v) dv && (3.66)\\
&= 2\int_{b'^{-1}(\mu_i)}^{b'^{-1}(y_i)} (y_i - b'(v)) dv && (3.67)\\
&= 2\left[y_i v - b(v)\right]_{v=b'^{-1}(\mu_i)}^{b'^{-1}(y_i)} && (3.68)\\
&= 2\left(y_i\left(b'^{-1}(y_i) - b'^{-1}(\mu_i)\right) - \left(b(b'^{-1}(y_i)) - b(b'^{-1}(\mu_i))\right)\right). && (3.69)
\end{aligned}$$

We now insert this expression in (3.60)

$$\begin{aligned}
f(y_i; \mu_i, \phi_i) &= a(y_i, \phi_i) \exp\left(-\frac{1}{2\phi_i} d(y_i, \mu_i)\right) && (3.70)\\
&= a(y_i, \phi_i) \exp\left(-\frac{1}{\phi_i}\left(y_i\left(b'^{-1}(y_i) - b'^{-1}(\mu_i)\right) - \left(b(b'^{-1}(y_i)) - b(b'^{-1}(\mu_i))\right)\right)\right). && (3.71)
\end{aligned}$$

From (3.39) we have that $b'^{-1}(\mu_i) = \theta_i$, which yields

$$f(y_i;\theta_i,\phi_i) = a(y_i,\phi_i)\exp\left(-\frac{1}{\phi_i}\left(y_i\left(b'^{-1}(y_i)-\theta_i\right)-\left(b(b'^{-1}(y_i))-b(\theta_i)\right)\right)\right) \tag{3.72}$$

$$= a(y_i,\phi_i)\exp\left(\frac{1}{\phi_i}\left(-y_i b'^{-1}(y_i)+y_i\theta_i+b(b'^{-1}(y_i))-b(\theta_i)\right)\right). \tag{3.73}$$

Let $a(y_i,\phi_i) = h(y_i,\phi_i)\exp\left(\frac{1}{\phi_i}\left(y_i b'^{-1}(y_i) - b\left(b'^{-1}(y_i)\right)\right)\right)$, then we get the density of an exponential family with respect to the canonical parameter θ_i

$$f(y_i;\theta_i,\phi_i) = h\left(y_i;\phi_i\right)\exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i}\right). \tag{3.74}$$

■

To summarize, we now have two ways of presenting a density of an exponential family, one by its definition (3.13) with respect to the canonical parameter θ_i and another one as the parameterization relative to μ_i, as seen in (3.60).

3.3.5 Generalized Linear Models

A model is considered a GLM if the following assumptions hold.

- Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ be the random vector containing the i.i.d. random responses Y_i for $i = 1, 2, \ldots, n$, of which y_i is the realization.
- Let the Y_i's have densities of the form (3.60), with the same variance function $\mathcal{V}$, and assume that each response y_i has covariates $\mathbf{x}_i$.
- The covariates influence the distribution of the response through the linear predictor η_i given by $\eta_i = \mathbf{x}_i^\top\boldsymbol{\beta}$.
- The mean value $\mathbb{E}[Y_i] = \mu_i$ is a smooth and invertible function of the linear predictor $\mu_i = g^{-1}(\eta_i)$, where $g(\cdot)$ is the link function.

It is not necessary to estimate ϕ_i in this master thesis, as it is known that $\phi_i = 1$ for the binomial distribution, and we are thus left with the task of estimating $\boldsymbol{\beta}$, which we will aim to do utilizing maximum likelihood estimation in the remainder of this section, where it is assumed, that the weights are known.

3.3.6 The Joint Density of the Exponential Family

We now find the density function, log likelihood, score function, observed information and the Fisher information for the exponential family, as they are all needed when we seek to

estimate $\boldsymbol{\beta}$ in the next section. The joint density of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ using the parameterization $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ and dispersion parameter vector $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_n)$ is due to independence given by

$$f(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\phi}) = \prod_{i=1}^{n} f(y_i; \mu_i, \phi_i) = \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{1}{\phi_i} d(y_i, \mu_i)\right) \prod_{i=1}^{n} a(y_i, \phi_i). \tag{3.75}$$

The log-likelihood is thus given by

$$\mathcal{L}(\boldsymbol{\mu}; \mathbf{y}) = -\frac{1}{2}\sum_{i=1}^{n} \frac{1}{\phi_i} d(y_i, \mu_i) + \sum_{i=1}^{n} \log\left(a(y_i, \phi_i)\right). \tag{3.76}$$

Differentiating this with respect to μ_i yields

$$\frac{d}{d\mu_i}\mathcal{L}(\boldsymbol{\mu}; \mathbf{y}) = \frac{d}{d\mu_i}\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{1}{\phi_i} 2\int_{\mu_i}^{y_i} \frac{y_i - u}{\mathcal{V}(u)} du\right) = \frac{1}{\phi_i}\frac{y_i - \mu_i}{\mathcal{V}(\mu_i)}. \tag{3.77}$$

Thus, the score function $\mathcal{S}$ with respect to $\boldsymbol{\mu}$ is given by

$$\mathcal{S}_{\boldsymbol{\mu}}(\boldsymbol{\mu}; \mathbf{y}) = \frac{\partial}{\partial \boldsymbol{\mu}}\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{1}{\phi_i} d(y_i, \mu_i)\right) = diag\left(\frac{1}{\phi_i \mathcal{V}(\mu_i)}\right)(\mathbf{y} - \boldsymbol{\mu}), \tag{3.78}$$

where $diag\left(\frac{1}{\phi_i \mathcal{V}(\mu_i)}\right)$ is an $n \times n$ diagonal matrix wherein the value of the i'th diagonal entry is $\left(\frac{1}{\phi_i \mathcal{V}(\mu_i)}\right)$.

The observed information is given by

$$\begin{aligned} j_{\boldsymbol{\mu}}(\boldsymbol{\mu}; \mathbf{y}) &= -\frac{\partial}{\partial \boldsymbol{\mu}^\top}\mathcal{S}_{\boldsymbol{\mu}}(\boldsymbol{\mu}; \mathbf{y}) && (3.79)\\ &= -\frac{\partial}{\partial \boldsymbol{\mu}^\top} diag\left(\frac{1}{\phi_i \mathcal{V}(\mu_i)}\right)(\mathbf{y} - \boldsymbol{\mu}) && (3.80)\\ &= diag\left(\frac{d}{d\mu_i}\frac{1}{\phi_i \mathcal{V}(\mu_i)}(\mu_i - y_i)\right) && (3.81)\\ &= diag\left(\frac{1}{\phi_i \mathcal{V}(\mu_i)} + \frac{1}{\phi_i}(y_i - \mu_i)\frac{\mathcal{V}'(\mu_i)}{\mathcal{V}^2(\mu_i)}\right). && (3.82) \end{aligned}$$

And thus the Fisher information matrix is

$$i_{\boldsymbol{\mu}}(\boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\mu}}\left[j_{\boldsymbol{\mu}}(\boldsymbol{\mu}; \mathbf{Y})\right] = diag\left(\frac{1}{\phi_i \mathcal{V}(\mu_i)}\right). \tag{3.83}$$

3.3.7 Maximum Likelihood Estimation

Recall from (3.56) and (3.54) that $\mu_i = g^{-1}(\eta_i)$ and $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$, with the data matrix X from Section 2.2. We then define the local design matrix corresponding to $\boldsymbol{\beta}$ by

$$X(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^\intercal}, \tag{3.84}$$

which can then be rewritten as

$$X(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}^\intercal}\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}^\intercal} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}^\intercal}X = \frac{\partial g^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^\intercal}X = diag\left(\frac{1}{g'(g^{-1}(\eta_i))}\right)X = diag\left(\frac{1}{g'(\mu_i)}\right)X, \tag{3.85}$$

where $diag\left(\frac{1}{g'(\mu_i)}\right)$ is a $n \times n$ diagonal matrix, X is a given data matrix and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^\top$.

Now consider a GLM for $\mathbf{Y}$ *as described in Section (3.3.5) with linear predictor* $\boldsymbol{\eta} = X\boldsymbol{\beta}$*. Then the maximum likelihood estimate* $\hat{\boldsymbol{\beta}}$ *for* $\boldsymbol{\beta}$ *is found as the solution to*

$$X(\boldsymbol{\beta})^\top i_{\boldsymbol{\mu}}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}, \tag{3.86}$$

where $X(\boldsymbol{\beta})$ *is the local design matrix and* $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$*.*

Proof. Recall from (3.78) the score function with respect to $\boldsymbol{\mu}$. If this is differentiated with respect to $\boldsymbol{\beta}$ and set equal to $\mathbf{0}$, the maximum likelihood estimate is found with respect to $\boldsymbol{\beta}$ by solving the resulting equation

$$\mathcal{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}; \mathbf{y}) = \frac{\partial \boldsymbol{\mu}^\top}{\partial \boldsymbol{\beta}}\mathcal{S}_{\boldsymbol{\mu}}(\boldsymbol{\mu}; \mathbf{y}) \tag{3.87}$$

$$= X(\boldsymbol{\beta})^\top diag\left(\frac{1}{\phi_i \mathcal{V}(\mu_i)}\right)(\mathbf{y} - \boldsymbol{\mu}) \tag{3.88}$$

$$= X(\boldsymbol{\beta})^\top i_{\boldsymbol{\mu}}(\boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu}), \tag{3.89}$$

where $i_{\boldsymbol{\mu}}$ is the found Fisher information from (3.83). ■

Equation (3.86) can not be solved directly, as it is not linear in $\boldsymbol{\beta}$. Instead, the maximum likelihood estimate is found by an iterative method called *iteratively reweighted least squares* (IRWLS), which is based on the Newton Raphson method. For IRWLS, it is assumed that the link function is the canonical link function, that is, $g(\boldsymbol{\mu}) = \boldsymbol{\theta}$, which due to (3.39) can be written as

$$g(\boldsymbol{\mu}) = b'^{-1}(\boldsymbol{\mu}). \tag{3.90}$$

It is first noted that the differentiated canonical link function with respect to $\boldsymbol{\mu}$ is due to differentiation of inverse function given by

$$g'(\boldsymbol{\mu}) = \frac{\partial}{\partial \boldsymbol{\mu}}b'^{-1}(\boldsymbol{\mu}) = \frac{1}{b''(b'^{-1}(\boldsymbol{\mu}))}. \tag{3.91}$$

The denominator of (3.91) is by definition the variance function (3.40), thus

$$g'(\boldsymbol{\mu}) = \frac{1}{\mathcal{V}(\boldsymbol{\mu})}. \tag{3.92}$$

Therefore, in the case of the link function being the canonical link function, the local design matrix is given by

$$X(\boldsymbol{\beta}) = diag\left(\frac{1}{g'(\mu_i)}\right)X = diag(\mathcal{V}(\mu_i))X. \tag{3.93}$$

Inserting this in (3.86) yields

$$\left(diag(\mathcal{V}(\mu_i))\, X\right)^\top i_{\boldsymbol{\mu}}(\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}))=\mathbf{0}, \tag{3.94}$$

which we thus want to solve with respect to $\boldsymbol{\beta}$. This can also be written as

$$X^\top diag\left(\mathcal{V}(\mu_i)\right) diag\left(\frac{1}{\phi_i \mathcal{V}(\mu_i)}\right)(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}))=\mathbf{0} \tag{3.95}$$

$$X^\top diag\left(\frac{1}{\phi_i}\right)(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta}))=\mathbf{0}, \tag{3.96}$$

where the left hand side of this equation is the score function for $\boldsymbol{\beta}$. For the Newton Raphson method, the differentiated score function with respect to $\boldsymbol{\beta}$ must also be used. It is given by

$$\mathcal{S}'_{\boldsymbol{\beta}}(\boldsymbol{\beta};\mathbf{y})=\frac{\partial}{\partial \boldsymbol{\beta}^\top} X^\top diag\left(\frac{1}{\phi_i}\right)(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta})) \tag{3.97}$$

$$=-X^\top diag\left(\frac{1}{\phi_i}\right) X(\boldsymbol{\beta}) \tag{3.98}$$

$$=-X^\top diag\left(\frac{1}{\phi_i}\right) diag\left(\mathcal{V}(\mu_i)\right) X \tag{3.99}$$

$$=-X^\top diag\left(\frac{\mathcal{V}(\mu_i)}{\phi_i}\right) X, \tag{3.100}$$

where the second equal sign is due to 3.84 and the third equal sign is due to 3.93. The Newton Raphson method is an iterative process for determining roots, in this case for $\mathcal{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta};\mathbf{y})=\mathbf{0}$. Newton Raphson's formula is

$$\hat{\boldsymbol{\beta}}_{v+1}=\hat{\boldsymbol{\beta}}_v-\left[\mathcal{S}'_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_v;\mathbf{y})\right]^{-1}\mathcal{S}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_v;\mathbf{y}). \tag{3.101}$$

The formula therefore requires us to come up with a start guess $\hat{\boldsymbol{\beta}}_0$. This iterative formula is repeated until the difference of $\hat{\boldsymbol{\beta}}_v$ and $\hat{\boldsymbol{\beta}}_{v+1}$ is sufficiently small. The iterative process can be written as

$$\hat{\boldsymbol{\beta}}_{v+1}=\hat{\boldsymbol{\beta}}_v-\left[\mathcal{S}'_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_v;\mathbf{y})\right]^{-1}\mathcal{S}_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}_v;\mathbf{y}) \tag{3.102}$$

$$=\hat{\boldsymbol{\beta}}_v+\left(X^\top diag\left(\frac{\mathcal{V}(\mu_i)}{\phi_i}\right)X\right)^{-1}X^\top diag\left(\frac{1}{\phi_i}\right)(\mathbf{y}-\boldsymbol{\mu}(\boldsymbol{\beta})). \tag{3.103}$$

The method is called the iteratively reweighted least squares method because 3.103 is a weighted least square estimate, where $diag\left(\frac{\mathcal{V}(\mu_i)}{\phi_i}\right)$. When it is not the canonical link function that is used, then the process is called the Fisher scoring algorithm. Note that the inverse of $X^\top diag\left(\frac{\mathcal{V}(\mu_i)}{\phi_i}\right)X$ is used, which is only possible if X has full rank as $diag\left(\frac{\mathcal{V}(\mu_i)}{\phi_i}\right)$ is a diagonal matrix. Thus, it is a requirement for GLM's that $n>p$, that is, that there are more observations than parameters.

3.4 Logistic Regression

Some classifiers calculate the probability that an individual belongs to a specific group and on the basis of a threshold value (the simplest being 50/50). Logistic regression is one such classifier. For the case of only two classes $K = 2$, let $\mathbf{Y}$ be the random vector containing the Bernoulli distributed responses $Y_i \sim \texttt{Bern}(P(Y_i = 1 \mid \mathbf{x}_i))$ for individual $i = 1, \ldots, n$. We will later in this section look at an extension for over two classes.

Let X be the $n \times p$ data matrix for p covariates and let $\mathbf{x}_i$ be the vector containing all covariates for individual i.
In order to classify an individual, the probability $P(Y_i = 1 \mid \mathbf{x}_i)$ is considered. Since probabilities have values between 0 and 1, functions with values within this range must be used. Many functions fulfill this, but logistic regression uses the logistic function which is given by

$$P(Y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})}, \tag{3.104}$$

where $\boldsymbol{\beta}$ is a vector of p parameters to estimate. To ease interpretation of the parameters we manipulate equation (3.104) into log-odds by

$$\log\left(\frac{P(Y_i = 1 \mid \mathbf{x}_i)}{1 - P(Y_i = 1 \mid \mathbf{x}_i)}\right) = \log\left(\frac{\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}}{1 - \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}}\right) = \log\left(\frac{\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}}{\frac{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})-\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}. \tag{3.105}$$

The log-odds of the logistic function are linear and therefore the parameters are easier to interpret than for the logistic function. We use this throughout the project and denote the log-odds of the logistic function by $\text{logit}\{P(Y_i = 1 \mid \mathbf{x}_i)\}$.

Note that logistic regression is a special case of a GLM, where the Y_i's are Bernoulli distributed and the link function is the logit function

$$g(\mu_i) = \text{logit}\{\mu_i\} = \log\left(\frac{P(Y_i = 1 \mid \mathbf{x}_i)}{1 - P(Y_i = 1 \mid \mathbf{x}_i)}\right) = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_i. \tag{3.106}$$

We already saw that the logit function also is the canonical link function for the Bernoulli distribution in (3.58).

The parameter vector $\boldsymbol{\beta}$ in (3.104) is estimated using maximum likelihood estimation. Let $\mathbf{y}$ be the realization of $\mathbf{Y}$ and assume Y_i and Y_j are independent for $i \neq j$. Consider the density function for Y_i given $\mathbf{x}_i$

$$f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) = P\left(Y_i = 1 \mid \mathbf{x}_i\right)^{y_i} \left(1 - P(Y_i = 1 \mid \mathbf{x}_i)\right)^{1-y_i}. \tag{3.107}$$

The joint density function for all responses $Y_1, Y_2, \ldots, Y_n$ is the product of these densities for $i = 1, \ldots, n$ due to independence. That is

$$f(\mathbf{y} \mid X; \boldsymbol{\beta}) = \prod_{i=1}^{n} P\left(Y_i = 1 \mid \mathbf{x}_i\right)^{y_i} \left(1 - P(Y_i = 1 \mid \mathbf{x}_i)\right)^{1-y_i}. \tag{3.108}$$

Inserting the logistic function (3.104) yields

$$f(\mathbf{y} \mid X; \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right)^{y_i} \left(1 - \left(\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right) \right)^{1-y_i} \tag{3.109}$$

$$= \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right)^{y_i} \left(\frac{1}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right)^{1-y_i}. \tag{3.110}$$

The log-likelihood is then given by

$$\mathcal{L}(\boldsymbol{\beta}; \mathbf{y} \mid X) = \sum_{i=1}^{n} \log \left(\left(\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right)^{y_i} \left(\frac{1}{1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})} \right)^{1-y_i} \right) \tag{3.111}$$

$$= \sum_{i=1}^{n} \left(y_i \left(\mathbf{x}_i^\top \boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})) \right) - (1 - y_i) \log(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})) \right) \tag{3.112}$$

$$= \sum_{i=1}^{n} \left(y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})) \right). \tag{3.113}$$

The maximum likelihood estimator for $\boldsymbol{\beta}$ can often not be expressed explicitly. In such cases it can be solved iteratively with tools like the Newton-Raphson method [2]. Once the estimate $\hat{\boldsymbol{\beta}}$ is found, it can be used for estimating the logistic function in (3.104) and thereby classify an observation based on the predicted probability. An example for logistic regression in the binary case is presented after logistic regression for multiple response classes in order to get fewer plots showing the same principle.

3.4.1 Logistic Regression for More Than Two Classes

For two classes, the log-odds are calculated by (3.105), where $1 - P(Y_i = 1 \mid \mathbf{x}_i)$ is of course the same as $P(Y_i = 0 \mid \mathbf{x}_i)$. When there are K classes, then the $K - 1$ log-odds are calculated instead

$$\log \left(\frac{P(Y_i = 1 \mid \mathbf{x}_i)}{P(Y_i = K \mid \mathbf{x}_i)} \right) = \mathbf{x}_i^\top \boldsymbol{\beta}_1 \tag{3.114}$$

$$\log \left(\frac{P(Y_i = 2 \mid \mathbf{x}_i)}{P(Y_i = K \mid \mathbf{x}_i)} \right) = \mathbf{x}_i^\top \boldsymbol{\beta}_2 \tag{3.115}$$

$$\vdots$$

$$\log \left(\frac{P(Y_i = K - 1 \mid \mathbf{x}_i)}{P(Y_i = K \mid \mathbf{x}_i)} \right) = \mathbf{x}_i^\top \boldsymbol{\beta}_{K-1}, \tag{3.116}$$

where the choice of denominator is arbitrary. If the exponential function is applied to both sides and the numerator is isolated, the probability of belonging to a particular class is obtained as

$$P(Y_i = 1 \mid \mathbf{x}_i) = P(Y_i = K \mid \mathbf{x}_i) \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_1) \tag{3.117}$$

$$P(Y_i = 2 \mid \mathbf{x}_i) = P(Y_i = K \mid \mathbf{x}_i) \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_2) \tag{3.118}$$

$$\vdots$$

$$P(Y_i = K - 1 \mid \mathbf{x}_i) = P(Y_i = K \mid \mathbf{x}_i) \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_{K-1}). \tag{3.119}$$

Since all probabilities must sum to 1, $P(Y_i = K \mid \mathbf{x}_i)$ can be found by

$$P(Y_i = K \mid \mathbf{x}_i) = 1 - \sum_{c=1}^{K-1} P(Y_i = c \mid \mathbf{x}_i) = 1 - \sum_{c=1}^{K-1} P(Y_i = K \mid \mathbf{x}_i) \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_c), \quad (3.120)$$

which can be rewritten as

$$-\frac{P(Y_i = K \mid \mathbf{x}_i) - 1}{P(Y_i = K \mid \mathbf{x}_i)} = \sum_{c=1}^{K-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_c) \quad (3.121)$$

$$-1 + \frac{1}{P(Y_i = K \mid \mathbf{x}_i)} = \sum_{c=1}^{K-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_c) \quad (3.122)$$

$$P(Y_i = K \mid \mathbf{x}_i) = \frac{1}{1 + \sum_{c=1}^{K-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_c)}. \quad (3.123)$$

If (3.123) is inserted in (3.117)-(3.119), the other probabilities can thus be found

$$P(Y_i = 1 \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_1)}{1 + \sum_{c=1}^{K-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_c)} \quad (3.124)$$

$$P(Y_i = 2 \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_2)}{1 + \sum_{c=1}^{K-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_c)} \quad (3.125)$$

$$\vdots$$

$$P(Y_i = K - 1 \mid \mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta}_{K-1})}{1 + \sum_{c=1}^{K-1} \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_c)}. \quad (3.126)$$

In order to calculate the probabilities for an observation belonging to a particular class, the parameters must thus be estimated, which, like the case with two classes, is estimated by maximum likelihood, which can again be approximated by Newton-Raphson.

Example

The simulated data from Section 1.1 is used in this example. Logistic regression for the binary case is presented first and then expanded to multiple classes.

Applying logistic regression we get an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ for (3.105). From this estimate, when $P(Y_i = 1 \mid \mathbf{x}_i) = 0.5$, the left hand side of (3.105) is zero, and thus the decision line for logistic regression is found by isolating x_{i2} and calculating the intercept and the slope for the two dimensional case. That is

$$x_{i2} = \frac{-\beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} x_{i1}, \quad (3.127)$$

where $\frac{-\beta_0}{\beta_2}$ is the intercept and $-\frac{\beta_1}{\beta_2}$ is the slope. Figure 3.3 shows the simulated data, where data is split into two categories based on these decision boundaries, where the top left plot shows the case where an observation belongs to `1percent` or not. Similarly, the top right plot is whether the observation belongs to `5percent` or not and the bottom left plot shows whether the observation belongs to the `94percent` class or not. These three binary decision boundaries are assembled in the last plot at the bottom right, marked

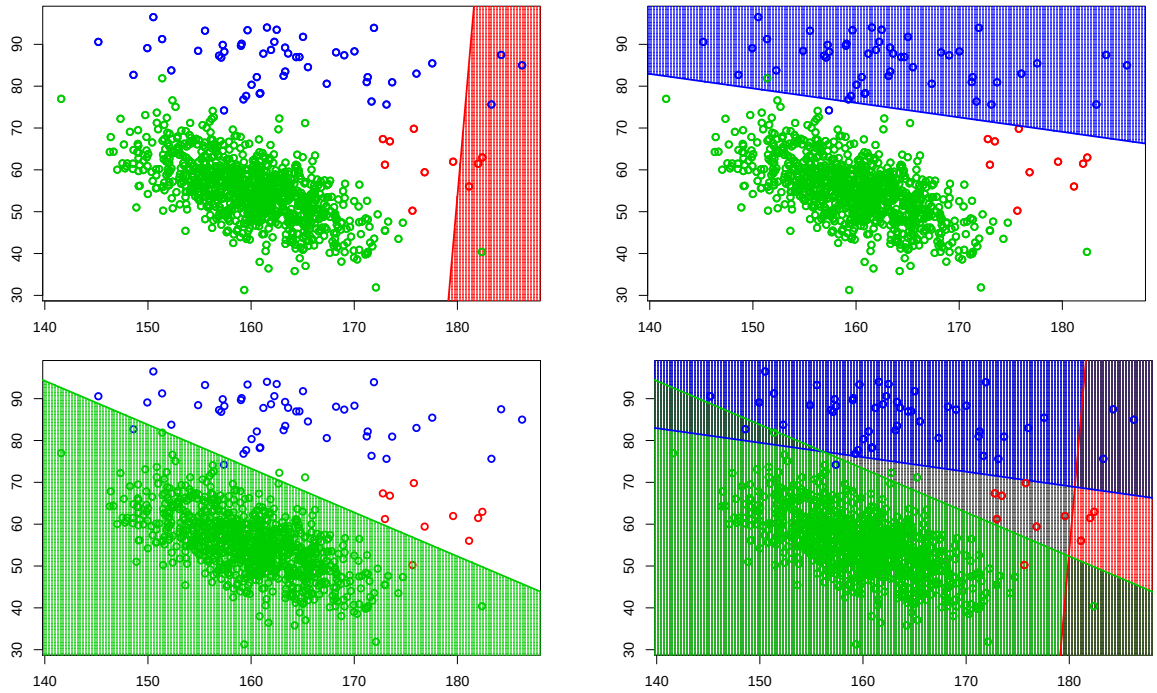


Figure 3.3: Logistic regression binary classification with a threshold of 50% based on three logistic regressions of `class`∼`Height+Weight`, where `class` is `1percent`, `5percent` and `94percent` in the three logistic regressions respectively. Gray indicates inconclusive areas.

with the same lines. The gray areas indicate initially inconclusive areas. We thus review methods that can be used when there are more than two classes.

One way to avoid the gray decision areas is by using one-vs-rest logistic regression. The method is based on the fact that with the three $\hat{\boldsymbol{\beta}}$ estimates we have just calculated, the black decision lines on the left plot of Figure 3.4 can be found by setting the probability functions equal to each other, one of them is $P(Y_i = 1 \mid \mathbf{x}_i) = P(Y_i = 2 \mid \mathbf{x}_i)$, and thus the plot is the result of one-vs-rest logistic regression for three classes. The plot to the right in Figure 3.4 also shows the method one-vs-rest logistic regression, but where the number of observations in each class are the same. Note that the number of observations thus affects where the decision boundary is located for one-vs-rest logistic regression. We conclude that given either unequal or equal n in the three classes the Bayes decision line is nicely approximated with straight lines.

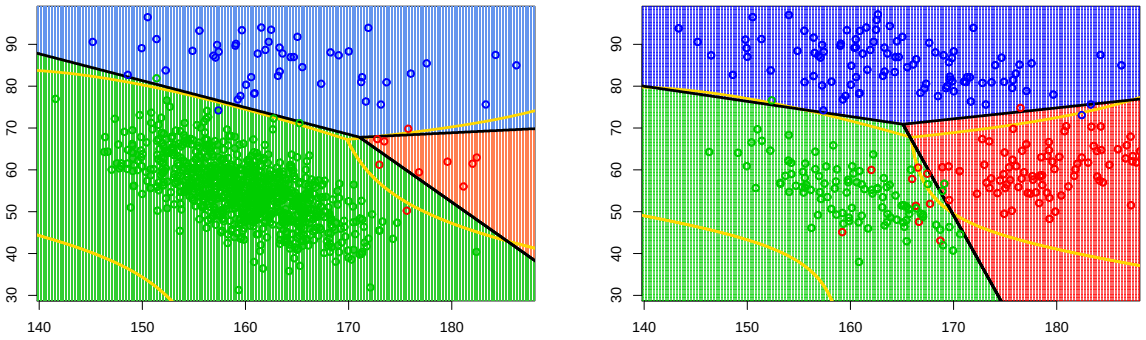


Figure 3.4: One-vs-rest Logistic regression performed on the simulated data set on the left and one-vs-rest Logistic regression performed on equally sized groups on the right. The Bayes decision line is shown in gold.

Another method that can be used when there are more than two classes is multinomial logistic regression. This method has been reviewed in theory in section 3.4.1. In the case of three classes, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are thus estimated based on (3.114) and (3.115). Then probabilities can be found for (3.124) and (3.125) and finally, the probability of the last class can be found by

$$P(Y_i = 3 \mid \mathbf{x}_i) = 1 - P(Y_i = 1 \mid \mathbf{x}_i) - P(Y_i = 2 \mid \mathbf{x}_i). \tag{3.128}$$

Similar to the one-vs-rest example, the probability functions are set equal to each other in order to make the decision boundaries shown in Figure 3.5. The plots show that on the simulated $n = 1000$ data set the multinomial logistic regression performs poorly, as this might be due to the low occurrence of class `1percent`. We thus try with equal class sizes, and see slightly better results. The splitting of the area is still unstable under change in reference class, we thus increase n to 3000 such that each class consist of 1000 observations. This increase in n led to more stable results, and we thus tried with uneven class sizes. We conclude that the number of observations thus again has influence on decision boundaries.

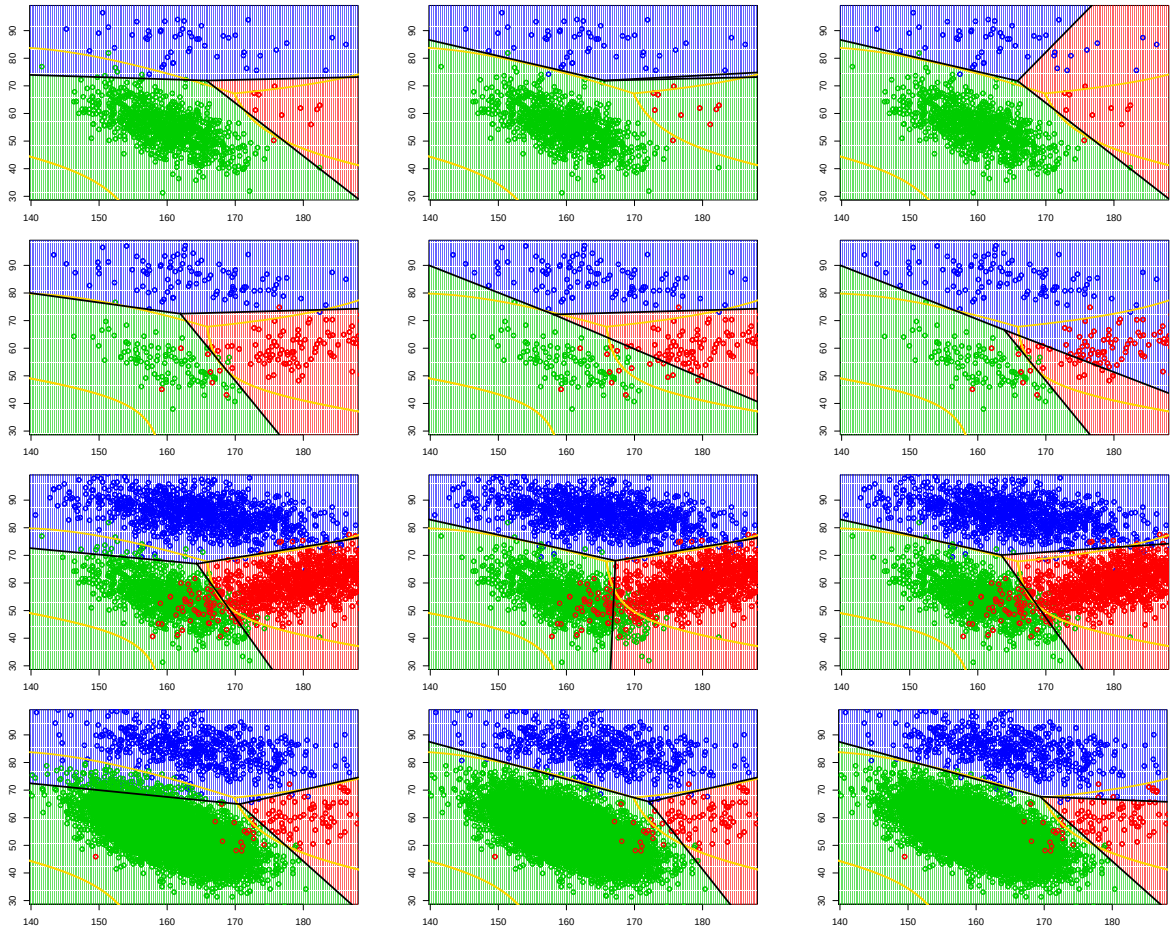


Figure 3.5: Multinomial logistic regression with the red class as reference in the first column, blue class as reference in the second column and green class as reference in the third column. All three classes are drawn from the usual distributions from Section 1.1. In the first row we have the usual dataset, in the second row there are 100 observations of each class, in the third row there are 1000 observations from each class and in the bottom row we have the classes `1percent`, `5percent` and `94percent` as usual, but there are 10 times as many observations as usually such that $n = 10000$.

One of the benefits of using logistic regression for us is that we can calculate the probability that an observation belongs to a particular class. In addition, decision boundaries can be made based on different tresholds. In all of the above, a threshold of 0.5 is chosen. Figure 3.6 shows two different thresholds. The solid line is the decision boundary for belonging to the class `94percent` with the 0.5 threshold. The dashed line shows a threshold of 0.9, such that there is considerably more certainty that one belongs to the green class `94percent`. It is calculated by isolating x_{i2} in (3.105).

$$x_{i2} = \frac{\log\left(\frac{0.9}{0.1}\right) - \beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} x_{i1}. \tag{3.129}$$

The threshold thus has no influence on the slope, but the intercept is changed based on the threshold. □

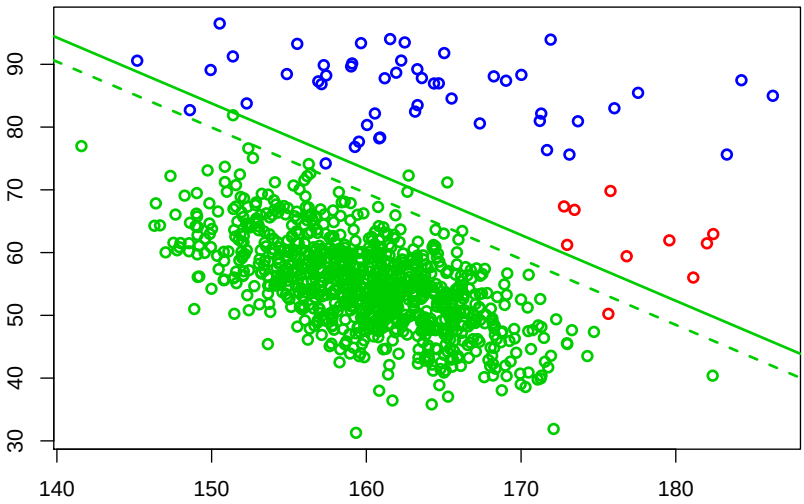


Figure 3.6: Logistic regression for the binary case of green or not green. The solid line indicates a threshold of 0.5 and the dashed line a threshold of 0.9.

One main advantage of logistic regression for us is the fact that logistic regression produces a probability. Even though we aim to actually predict ADHD and ASD, the certainty of such prediction is also important. Furthermore, the logistic regression model is a simple classifier in the sense that it produces linear decision lines, and thus is a high bias classifier, but with low variance. In the next section we try and alleviate some of the bias by moving beyond linearity with ex. splines.

3.4.2 Moving Beyond Linearity

Recall the logit function (3.105) for the log-odds of a binary response. We wish to extend this to a model that allows for non-linear relationships between $\mathbf{y}$ and some or all of the predictors $\mathbf{x}_{\bullet j}$. Such a model is called a *Generalized Additive Model (GAM)*, where instead of estimating a single parameter β_j to each predictor $x_{\bullet j}$, we fit a function $f_j(x_{ij})$ while still maintaining additivity, that is

$$\text{logit}\left\{P(Y_i = 1 \mid \mathbf{x}_i)\right\} = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}). \tag{3.130}$$

A GAM is more flexible than the simpler linear models on the form $y_i = \sum_{i=1}^{n} x_i\beta_i$ which can be graphed as a straight line, this flexibility can lead to better predictions, especially if a predictor $\mathbf{x}_{\bullet j}$ is related non-linearly to $\mathbf{y}$. Previously we have considered flexibility in terms of how many parameters are to be fit, and thus the *degrees of freedom* have been equal to the number of predictors. Depending on the choice of f, this is no longer the case, and we will from here on discuss flexibility in terms of degrees of freedom.

Extending a model with additional degrees of freedom, and manipulating the predictors leads to complex and hard-to-interpret models, which is an acceptable drawback in a predictive study.

A function $f_j(x_{ij})$ can be constructed in many ways, the remainder of this section outlines a number of different ways to construct such a function for one given predictor j henceforth denoted $f(x_{i0})$.

Step Function

Given a continuous variable $\mathbf{x}_{\bullet 0}$, in some cases it could be preferable to assign specific values to different ranges of that variable. To create a *step function* we make K cut points $c_1, c_2, \ldots, c_K$ in the range of $\mathbf{x}_{\bullet 0}$ and generate $K + 1$ functions

$$\begin{aligned}
C_0(x_{i0}) &= \mathbb{1}[x_{i0} < c_1]\,, \\
C_1(x_{i0}) &= \mathbb{1}[c_1 \leq x_{i0} < c_2]\,, \\
C_2(x_{i0}) &= \mathbb{1}[c_2 \leq x_{i0} < c_3]\,, \\
&\vdots \\
C_{K-1}(x_{i0}) &= \mathbb{1}[c_{K-1} \leq x_{i0} < c_K]\,, \\
C_K(x_{i0}) &= \mathbb{1}[c_K \leq x_{i0}]\,.
\end{aligned}$$

Generating a step function corresponds to creating an ordered categorical variable, and thus fitting a model

$$\text{logit}\left\{P(Y_i = 1 \mid x_{i0})\right\} = \beta_0 + f(x_{i0}) = \beta_0 + \beta_1 C_1(x_{i0}) + \beta_2 C_2(x_{i0}) + \ldots + \beta_K C_K(x_{i0}), \tag{3.131}$$

corresponds to a GAM with dummy variables and $C_1, \ldots, C_K$ as added functions. The function C_0 is omitted as it is covered by the case of all other functions set to zero: $C_1 = C_2 = \cdots = C_K = 0$. A step function uses K degrees of freedom as the original variable is represented by K dummy variables.

Polynomial Functions

Extending the linear model to a non-linear setting is straightforward with a polynomial function. It is possible to approximate data extremely well with an n-degree polynomial, this would of course introduce high variance, and we thus make do with polynomials of a much lower degree than n. Even though a polynomial is straight line, we can still use the linear setup to approximate the coefficients

$$\text{logit}\left\{P(Y_i = 1 \mid x_{i0})\right\} = \beta_0 + \beta_1 x_{i0} + \beta_2 x_{i0}^2 + \beta_3 x_{i0}^3 + \ldots + \beta_d x_{i0}^d. \tag{3.132}$$

A higher order polynomial can as mentioned fit arbitrarily well to the data, but high order polynomials can adopt strange shapes and, especially near the boundaries, take extreme values. Because of the high bias and unpredictable shape of higher order polynomials, polynomials are usually not used with a degree of more than 3 or 4. A polynomial function uses d degrees of freedom, besides the intercept, as a coefficient is fit for each degree of the polynomial.

Combining a polynomial function with a step function allows for separate polynomials on different ranges of a variable. This is useful if a variable has e.g. a change point or outliers near the edges of the range of $\mathbf{x}_{\bullet 0}$. Such a model is called a piece-wise polynomial. For example, a d'th degree polynomial with one cutpoint c is given by

$$\text{logit}\left\{P(Y_i = 1 \mid x_{i0})\right\} = \begin{cases} \beta_{0,1} + \beta_{1,1} x_{i0} + \beta_{2,1} x_{i0}^2 + \beta_{3,1} x_{i0}^3 + \ldots + \beta_{d,1} x_{i0}^d & if\ x_{i0} < c \\ \beta_{0,2} + \beta_{1,2} x_{i0} + \beta_{2,2} x_{i0}^2 + \beta_{3,2} x_{i0}^3 + \ldots + \beta_{d,2} x_{i0}^d & if\ x_{i0} \geq c, \end{cases} \tag{3.133}$$

where the coefficients are estimated as for logistic regression, where $\beta_{0,1}, \beta_{1,1}, \ldots, \beta_{d,1}$ are estimated for $x_{i0} < c$ and $\beta_{0,2}, \beta_{1,2}, \ldots, \beta_{d,2}$ are estimated for $x_{i0} \geq c$.

A piece-wise polynomial uses $(K+1) \times (d+1)$ degrees of freedom, as a d-dimensional polynomial is to be fit to each step of the variable.

Generally, instead of just using the step function and the polynomial function separately or together, whole functions, containing all kinds of transformations can be constructed for a predictor before conducting logistic regression. We call these functions basis functions b_k

$$\text{logit}\left\{P(Y_i = 1 \mid x_{i0})\right\} = \beta_0 + \beta_1 b_1(x_{i0}) + \beta_2 b_2(x_{i0}) + \beta_3 b_3(x_{i0}) + \ldots + \beta_K b_K(x_{i0}), \tag{3.134}$$

where, for example, the basic function of the polynomial function is $b_j(x_{i0}) = x_{i0}^j$.

Splines

The definition of a d-degree spline is a piecewise polynomial of degree d that is continuous in the cut points and which $1, 2, \ldots, d-1$ derivatives are also continuous. A commonly used spline is the cubic spline, which involves third degree polynomial functions, as they look smooth to the human eye at the cut points. It turns out that a cubic spline with K cutpoints can be modeled using the basis function model (3.134)

$$\text{logit}\left\{P(Y_i = 1 \mid x_{i0})\right\} = \beta_0 + \beta_1 b_1(x_{i0}) + \beta_2 b_2(x_{i0}) + \ldots + \beta_{K+3} b_{K+3}(x_{i0}) \tag{3.135}$$

and the model, for an appropriate choice of basic functions, can thus be fitted by logistic regression. A way to represent a cubic spline is to start out with a cubic polynomial and then add a *truncated power basis* function $h(x_{i0}, \xi)$ for each knot ξ.

$$h(x_{i0}, \xi) = (x_{i0} - \xi)_+^3 = \begin{cases} (x_{i0} - \xi)^3 & if\ x_{i0} > \xi \\ 0 & otherwise, \end{cases} \tag{3.136}$$

where " $+$ " indicates that the truncated power basis function is 0 for all values of x_{i0} less than or equal to the value of the cut point ξ. Adding a truncated power basis function to a cubic polynomial function yields

$$\text{logit}\left\{P(Y_i = 1 \mid x_{i0})\right\} = \beta_0 + \beta_1 x_{i0} + \beta_2 x_{i0}^2 + \beta_3 x_{i0}^3 + \beta_4 h(x_{i0}, \xi). \tag{3.137}$$

We will show that this function is continuous and that the same applies to its first and second derivatives. The functions $g_1(x_{i0}) = x_{i0}, g_2(x_{i0}) = x_{i0}^2, g_3(x_{i0}) = x_{i0}^3$ are all continuous and the same applies to their derivatives. Thus, the function $h(x_{i0}, \xi)$ is considered, which is continuous at the point ξ if

$$\lim_{x_{i0} \to \xi^+} (x_{i0} - \xi)^3 = \lim_{x_{i0} \to \xi^-} (0). \tag{3.138}$$

Obviously, the function $h(x_{i0}, \xi)$ is continuous at the point ξ. The first derivative of the function is $h'(x_{i0}, \xi) = 0$ for $x_{i0} \leq \xi$ and for $x_{i0} > \xi$ it is

$$((x_{i0} - \xi)^3)' = (x_{i0}^3 - 3x_{i0}^2\xi + 3x_{i0}\xi^2 - \xi^3)' = 3x_{i0}^2 - 6x_{i0}\xi + 3\xi^2. \tag{3.139}$$

The limit is taken correspondingly on both sides of the derivative function at the point ξ to show continuity for the derivative function. The function is differentiated again to get the second derivative

$$((x_{i0} - \xi)^3)'' = 6x_{i0} - 6\xi, \quad for\ x_{i0} > \xi, \tag{3.140}$$

which means that the second derivative function is also continuous at the point ξ. Thus, the function $h(x_{i0,\xi})$ is continuous at ξ and the same applies to its first and second derivatives. The third derivative of the function is not continuous because

$$((x_{i0} - \xi)^3)''' = 6, \quad for\ x_{i0} > \xi \tag{3.141}$$

and thus $\lim_{x_{i0} \to \xi^+} h(x_{i0}, \xi) \neq \lim_{x_{i0} \to \xi^-} h(x_{i0}, \xi)$, because the limits are 6 and 0 respectively. If K cut-points for a cubic spline are chosen, then the terms $\beta_4 h(x_{i0}, \xi_1), \beta_5 h(x_{i0}, \xi_2), \ldots, \beta_{K+3} h(x_{i0}, \xi_K)$ are added instead of $\beta_4 h(x_{i0}, \xi)$ in (3.137). Thus, a cubic spline with K cut points has $K + 4$ degrees of freedom, where $\xi_1, \xi_2, \ldots, \xi_K$ are the cut points.

Restricting a cubic spline even further, such that the first and last "pieces" are linear is called a *natural spline*. The cut-points can be placed uniformly on the interval for a variable and the number of cut-points is typically selected based on cross-validation, which is described in more detail in Chapter 4. Instead, we can use smoothing splines where the

number and location of the cut points should not be considered since a smoothing spline uses a maximum number of points, one for each observation. Generally, it is desired to maximize the log-likelihood function for logistic regression given by (3.113), which is done by determining the function

$$f(x_{i0}) = \text{logit}\left\{P(Y_i = 1 \mid x_{i0})\right\} = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{3.142}$$

that provides a maximum for the log-likelihood function. It implies that $P(Y_i = 1 \mid x_{i0})$ can be determined by

$$P(Y_i = 1 \mid x_{i0}) = \frac{\exp(f(x_{i0}))}{1 + \exp(f(x_{i0}))}. \tag{3.143}$$

Since f can be chosen such that the function goes through all observations, which of course will cause over-fitting, a restriction on the curvature is needed. Thus it is desired to maximize the penalized log-likelihood given by

$$\mathcal{L}(f; \lambda) = \sum_{i=1}^{n} \left[y_i f(x_{i0}) - \log(1 + \exp(f(x_{i0})))\right] - \frac{1}{2}\lambda \int f''(t)dt, \tag{3.144}$$

where $\lambda \geq 0$ is a tuning parameter that determines how much curvature is allowed for the function f. If $\lambda = 0$, it will result in no restrictions and $\lambda = \infty$ will cause a straight line. Thus, a value for λ must be found which provides the best bias-variance trade-off. The function that provides maximum for the penalized log-likelihood is called a smoothing spline.

When expanding a model with splines and thus increasing the degrees of freedom, the bias is reduced. Higher degrees of freedom is not necessarily a problem in predictive studies, but overfitting might still be an issue. We thus utilize the fact that logistic regression is likelihood-based and apply shrinkage methods.

3.4.3 Shrinkage Methods

Previously, models have been built with all p predictors. This can lead to unnecessary noise from the parameters and therefore the model may be less likely to predict new data correctly. Shrinkage methods shrink the parameter estimates towards 0, or even set them to 0. Thus, Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) are presented, which are two common shrinkage methods. These two shrinkage methods are likelihood-based and can therefore be used for logistic regression.

Ridge Regression

For logistic regression, the parameter vector $\boldsymbol{\beta}$ is estimated by Maximum Likelihood. In Ridge regression the sizes of the parameters are penalized such that we seek to optimize

$$\hat{\boldsymbol{\beta}}^{\text{R}} = \arg\max_{\boldsymbol{\beta}} \left(\mathcal{L}(\boldsymbol{\beta}; \mathbf{y} \mid X) - \lambda \sum_{j=1}^{p} \beta_j^2 \right), \tag{3.145}$$

where $\lambda \geq 0$ is a tuning parameter and the term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called the shrinkage penalty as it reduces the parameter estimates. The estimated parameter vector for Ridge regression is denoted $\hat{\boldsymbol{\beta}}^{\mathrm{R}}$. Typically, the predictors for Ridge regression are standardized and centered before estimation. This allows the parameters β_j, $j = 1, 2, \ldots, p$ to be penalized on the same scale. Ridge regression has the disadvantage that the method still includes all predictors, where it simply reduces the estimates, and hence makes the model less interpretable. A better prediction model may be made by setting some parameter estimates to 0 exactly as LASSO does.

LASSO

LASSO solves an optimization problem similar to that of Ridge regression

$$\hat{\boldsymbol{\beta}}^{\mathrm{L}} = \arg\max_{\boldsymbol{\beta}} \left(\mathcal{L}(\boldsymbol{\beta}; \mathbf{y} \mid X) - \lambda \sum_{j=1}^{p} |\beta_j| \right), \tag{3.146}$$

where the shrinkage penalty is $\lambda \sum_{j=1}^{p} |\beta_j|$ and λ is again a tuning parameter. Thus, the $\mathcal{L}_1$-norm is used in the shrinkage penalty

$$||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p} |\beta_j|. \tag{3.147}$$

The shrinkage penalty in this setting has the effect that when λ is sufficiently large, then some parameter estimates will be 0 exactly and thus LASSO can be used for variable selection. The parameter estimates from LASSO are denoted $\hat{\boldsymbol{\beta}}^{\mathrm{L}}$. LASSO estimates parameters and selects predictors at the same time since some estimates are most likely 0. Note that centering and standardization are also performed for LASSO as was the case for Ridge regression.

The log-likelihood for logistic regression in the case of two classes was found by (3.113) and if this likelihood is used in (3.146), the following optimization problem is obtained

$$\hat{\boldsymbol{\beta}}^{\mathrm{L}} = \arg\max_{\boldsymbol{\beta}} \left(\sum_{i=1}^{n} \left[y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta})) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right), \tag{3.148}$$

which can be similarly done for Ridge regression. The optimization problem can again be solved by maximum likelihood estimation using approximation methods such as Newton-Raphson.

Comparison of Ridge Regression and LASSO

Using a different notation, Ridge regression (3.145) and LASSO (3.146) can be reformulated as optimization problems, respectively solving

$$\max_{\boldsymbol{\beta}} \left(\mathcal{L}(\boldsymbol{\beta}; \mathbf{y} \mid X) \right), \textit{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s \tag{3.149}$$

$$\max_{\boldsymbol{\beta}} \left(\mathcal{L}(\boldsymbol{\beta}; \mathbf{y} \mid X) \right), \textit{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s. \tag{3.150}$$

Thus, for every value of λ there exists an s such that the optimization problem has the same solution. Consider for simplicity the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ for $p = 2$, that is, only 2 predictors are considered. In the case of LASSO, the $\boldsymbol{\beta}$ must be found within the diamond

$$|\beta_1| + |\beta_2| \leq s. \tag{3.151}$$

Similarly, the Ridge regression gives a parameter estimate within the circle

$$\beta_1^2 + \beta_2^2 \leq s. \tag{3.152}$$

Figure 3.7 illustrates what happens when LASSO and Ridge regression are used. In all four plots we try to predict the class `5percent` based on logistic regression with `Height1` and `Weight1` as predictors. In all four plots, the maximum likelihood estimate is indicated by $\hat{\boldsymbol{\beta}} = [0.46, -1.49]$. The top left plot shows the LASSO decision area with $\lambda = 0.06$ in blue, and the black contours representing estimates resulting in smaller likelihoods than the maximum on which $\hat{\boldsymbol{\beta}}$ is based. The estimate of the parameter vector for LASSO is the red point, where the contours meet the restricted area marked with blue. This gives $\hat{\boldsymbol{\beta}}^L = [0, -0.59]$. Similarly the top right plot shows the case of using Ridge regression, where $\hat{\boldsymbol{\beta}}^R = [0.22, -0.79]$ with $\lambda = 0.06$. That is, $\hat{\beta}_1^{\mathrm{L}} = 0$ for LASSO exactly where for Ridge regression $\hat{\beta}_1^{\mathrm{R}}$ is only shrunken towards 0. This will be the case for λ sufficiently large. The bottom left plot shows what happens if less penalty is applied using LASSO. When $\lambda = 0.02$, indicated by the blue solid line, then $\hat{\boldsymbol{\beta}}_1^L \neq 0$. The dashed line indicates what happens, if even less penalty is applied. The plot bottom right shows the situation where $\lambda = 0$, ie. where no penalty is imposed. This corresponds to maximizing the logistic regression log-likelihood. Note that if $\lambda = 0$ then there are infinitely many values for s that satisfy this. Here we just outlined one of them, represented by the blue solid line. In higher dimensions, that is, when more than two predictors are included, the contours could hit surfaces and corners for LASSO and thus set more parameter estimates to 0. The value of λ that gives the best prediction model can be found based on cross-validation as done in the example, for more on cross-validation see Section 4.2.1.

Example

First, a logistic model is fitted without a shrinkage penalty to be able to compare what happens to the parameter estimates when using LASSO and Ridge regression. We fit a logistic regression model with all height and weight variables defined in Appendix A, which are designed to influence the response. Finally, in this example, all normal and uniform variables that do not affect the response are included and therefore LASSO is expected to set their parameter estimates to 0. Furthermore all predictors are standardized in this example in order to compare the influences of the $\boldsymbol{\beta}$ estimates. The estimates for the logistic regression model are shown in the top row in Table 3.1.

It is now desired to find a LASSO and Ridge regression model. In order to create a LASSO or Ridge regression model, a value for λ must be selected first. Figure 3.8 shows a plot of the $\boldsymbol{\beta}$ estimates for respectively LASSO to the left and Ridge to the right relative

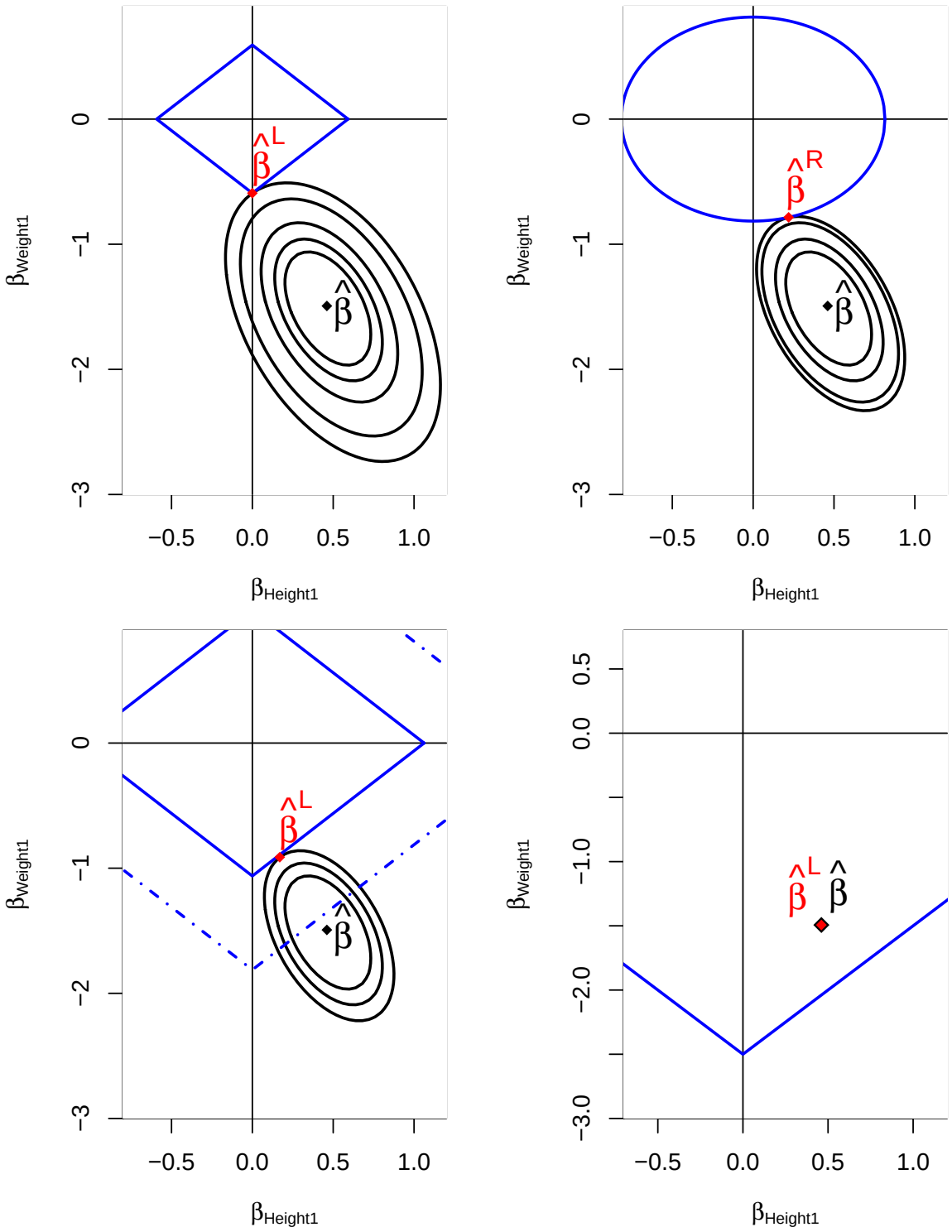


Figure 3.7: In all four plots we try to predict class `5percent` based on logistic regression using `Height1` and `Weight1` as predictors. In all four plots, the maximum likelihood estimate is indicated by $\hat{\boldsymbol{\beta}} = [0.46, -1.49]$.

Tops: On the left the LASSO decision area with $\lambda = 0.06$ in blue, and the black contours represents estimates resulting in smaller likelihoods than the maximum on which of $\hat{\boldsymbol{\beta}}$ is based. This gives $\hat{\boldsymbol{\beta}}^L = [0, -0.59]$. Similarly for Ridge regression $\hat{\boldsymbol{\beta}}^R = [0.22, -0.79]$ on the right with $\lambda = 0.06$.

Bottoms: The plot on the left shows what happens if a lower penalty is applied using LASSO. When $\lambda = 0.02$, indicated by the blue solid line, then $\hat{\boldsymbol{\beta}}_1^L \neq 0$. The dashed line indicates what happens, if even less penalty is applied. The plot to the right shows the situation where $\lambda = 0$, ie where no penalty is imposed. This corresponds to maximizing the logistic regression log-likelihood.

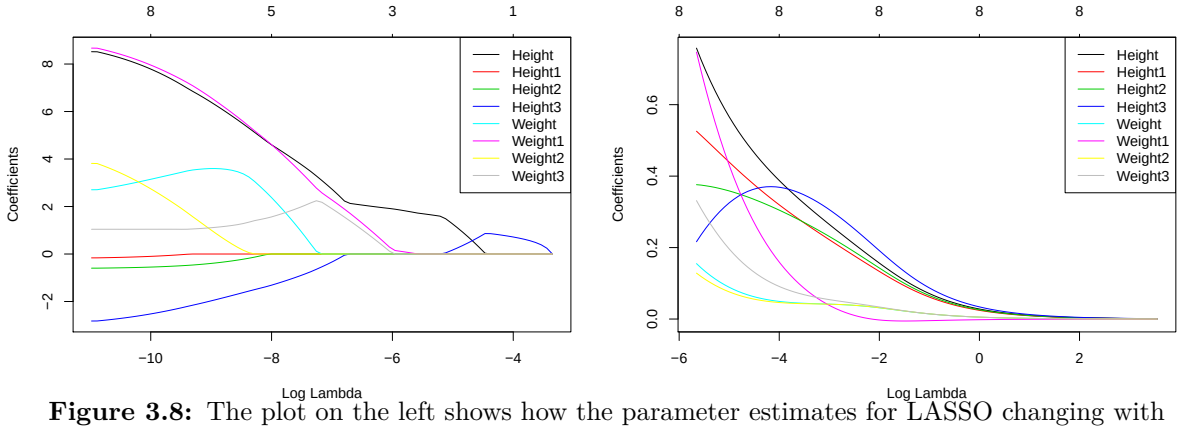


Figure 3.8: The plot on the left shows how the parameter estimates for LASSO changing with different choices of λ. Similarly, for Ridge regression, which is seen on the plot to the right. The logarithm of the λ values can be seen in the bottom x-axis of each plot. Large values of λ cause more shrinkage of the parameters. For Ridge regression, it is seen that the estimates are shrunk towards 0 and for LASSO they become 0 exactly. This is also indicated at the top of each of the plots, showing how many parameter estimates differ from 0 for a particular value of λ.

to different choices of λ. Thus, for LASSO, a λ value can be chosen which gives a desired number of predictors. Figure 3.8 also shows that the estimates for both LASSO and Ridge shrink towards zero when λ is sufficiently large.

For both LASSO and Ridge, we now select λ on the basis of 10-fold cross-validation, where the AUC evaluation measure is used, as logistic regression returns probabilities and thus is threshold dependent. For more on AUC see Section 4.1 and for more on cross-validation see Section 4.2.1. The λ value chosen is the largest λ within one standard deviation of the best λ, with the best λ value being the one that gives the largest AUC score. This resulted in a $\log(\lambda)$ value for LASSO of $\log(\lambda^{\mathrm{L}}) = -6.79$ and for Ridge regression, $\log(\lambda^{\mathrm{R}}) = -3.23$. The estimates for $\boldsymbol{\beta}$ based on these choices of λ are shown in Table 3.1.

	`Intercept`	`Height`	`Height1`	`Height2`	`Height3`	`Weight`	`Weight1`	`Weight2`	`Weight3`
LogReg	-14.977	9.311	-0.209	-0.603	-3.175	1.803	9.465	5.344	1.111
Ridge	-5.358	0.292	0.244	0.25	0.328	0.043	0.058	0.042	0.06
LASSO	-7.219	2.231	0	0	-0.048	0	1.89	0	1.553

Table 3.1: Parameter estimates for the model `1percent`$\sim$`Height+Height1+Height2+Height3+Weight+Weight1+Weight2+Weight3` using logistic regression, LASSO (with $\log(\lambda^{L}) = -6.79$) and Ridge regression (with $\log(\lambda^{R}) = -3.23$).

In Table 3.1 it can be seen that Ridge regression just shrinks the parameter estimates towards 0 compared to what the estimates for the logistic regression are. LASSO sets some parameter estimates to 0 exactly. It would have been expected that `Height` and `Weight`, as the latent variables, would have been the ones selected but LASSO set `Weight` to 0 and included `Weight1` and `Weight3` instead. Note that not all parameter estimates become smaller as λ becomes larger. For example, looking at Table 3.1 we see that the parameter estimate for `Weight3` has increased for LASSO by using $\log(\lambda^{L}) = -6.79$

compared with logistic regression. This is also seen in Figure 3.8, where the estimate value remains relatively high for this choice of λ.

Finally, we tried including all uniform and normal variables and correspondingly make a logistic regression, a LASSO and a Ridge regression model. As expected, LASSO set all the parameter estimates for the different uniform and normal variables to 0 and Ridge regression shrunk them towards 0. We did not start out the example with all these predictors, as eg. the plots in Figure 3.8, would be even harder to visually interpret with 16 variables. □

It may be that unnecessary noise has been removed by using LASSO or Ridge regression, and thus a better prediction model may be achieved. But whether the reduction of parameters from LASSO is better than just shrinking them with Ridge regression remains unexplored, and both will be applied to the real data.

The remainder of this chapter presents other classifiers, that might be applied to the real data set, as to compare them with our main method, logistic regression.

3.5 Linear Discriminant Analysis

Like logistic regression, linear discriminant analysis (LDA) is also a linear classifier. In Section 3.4, $P(Y_i = 1 \mid \mathbf{x}_i)$ was estimated directly using the logistic function. In linear discriminant analysis , the purpose is to estimate the same probability. The idea is to utilize Bayes Theorem to approximate the Bayes classifier (3.9). We get from Bayes theorem that

$$P(Y_i = c \mid \mathbf{X} = \mathbf{x}_i) = \frac{P(\mathbf{X} = \mathbf{x}_i \mid Y_i = c)P(Y_i = c)}{P(\mathbf{X} = \mathbf{x}_i)} \tag{3.153}$$

where $P(Y_i = c \mid \mathbf{X} = \mathbf{x}_i)$ is called the posterior probability. Let the density function for an observation, which belongs to class c be given by $f_c(\mathbf{x}_i) = P(\mathbf{X} = \mathbf{x}_i \mid Y_i = c)$. Furthermore, let $\pi_c = P(Y_i = c)$ be the prior probability that an individual belongs to class c, which means that $\sum_{c=1}^{K} \pi_c = 1$. Then the law of total probability implies

$$P(Y_i = c \mid \mathbf{X} = \mathbf{x}_i) = \frac{P(\mathbf{X} = \mathbf{x}_i \mid Y_i = c)P(Y_i = c)}{\sum_{l=1}^{K} P(\mathbf{X} = \mathbf{x}_i \mid Y_i = l)P(Y_i = l)} = \frac{f_c(\mathbf{x}_i)\pi_c}{\sum_{l=1}^{K} f_l(\mathbf{x}_i)\pi_l}. \tag{3.154}$$

The choice of c which maximizes (3.154) is the Bayes classifier. The prior probability π_c is easily found by taking the number of observations with response $Y_i = c$ and dividing it with the total number of observations. However, estimating $f_c(\mathbf{x}_i)$ is not as straightforward without some assumptions. LDA assumes that $f_c(\mathbf{x}_i)$ are c densities and each of these follows a p-dimensional normal distribution and assuming that the covariance matrix is the same for all K classes yields

$$f_c(\mathbf{x}_i) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c)\right), \tag{3.155}$$

where $\boldsymbol{\mu}_c$ is a mean vector for class c and Σ is a covariance matrix of size $p \times p$, which is thus assumed to be the same for all K classes. Inserting the density function in (3.154),

we get

$$P(Y_i = c \mid \mathbf{X} = \mathbf{x}_i) = \frac{\pi_c \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c)\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_l)^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_l)\right)} \tag{3.156}$$

$$= \frac{\pi_c \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c)\right)}{\sum_{l=1}^{K} \pi_l \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_l)^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_l)\right)}. \tag{3.157}$$

Maximizing this equation corresponds to maximizing the numerator herein, since the denominator is the same for all K classes. Moreover, since the logarithm is a monotone function, maximizing (3.157) corresponds to maximizing

$$\log\left(\pi_c \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c)\right)\right) \tag{3.158}$$

$$= \log(\pi_c) - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_c) \tag{3.159}$$

$$= \log(\pi_c) + \mathbf{x}_i^\top \Sigma^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\boldsymbol{\mu}_c^\top \Sigma^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\mathbf{x}_i^\top \Sigma^{-1}\mathbf{x}_i. \tag{3.160}$$

Since $\mathbf{x}_i$ does not depend on c, under maximization (3.160) reduces to

$$\delta_c(\mathbf{x}_i) = \log(\pi_c) + \mathbf{x}_i^\top \Sigma^{-1}\boldsymbol{\mu}_c - \frac{1}{2}\boldsymbol{\mu}_c^\top \Sigma^{-1}\boldsymbol{\mu}_c. \tag{3.161}$$

The function $\delta_c(\mathbf{x}_i)$ is called the linear discriminant function because it is linear for $\mathbf{x}_i$ and used to choose how to discriminate Y_i. In order to maximize (3.161) then $\delta_1(\mathbf{x}_i), \delta_2(\mathbf{x}_i), \ldots, \delta_K(\mathbf{x}_i)$ should be estimated, which is done by estimating the unknown parameters $\mu_1, \mu_2, \ldots, \mu_K$ and $\pi_1, \pi_2, \ldots, \pi_K$ as well as Σ and then inserting them into (3.161). These are given by the maximum likelihood estimates

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{n_c} \sum_{i:y_i=c} \mathbf{x}_i \tag{3.162}$$

$$\hat{\pi}_c = \frac{n_c}{n} \tag{3.163}$$

$$\hat{\Sigma} = \frac{1}{n-K} \sum_{c=1}^{K} \sum_{i:y_i=c} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)^\top (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c), \tag{3.164}$$

where n_c is the number of observations in class c and n is the total number of observations. Thus LDA approximates the Bayes classifier by assigning Y_i to the class for which the discriminant function δ_c is largest. Furthermore, $P(Y_i = c \mid \mathbf{X} = \mathbf{x}_i)$ can be estimated by inserting $\hat{\boldsymbol{\mu}}_c$, $\hat{\pi}_c$ and $\hat{\Sigma}$ into (3.155) and then (3.154).

Quadratic Discriminant Analysis

LDA can be extended to Quadratic Discriminant Analysis (QDA), here we still try to approximate the Bayes classifier with c density functions following p-dimensional normal distributions, but now we also have c different covariance matrices Σ_c.

Once again we maximize (3.157), now with Σ_c, which corresponds to maximizing $\delta_c(\mathbf{x}_i)$

$$\delta_c(\mathbf{x}_i) = -\frac{1}{2}\left(\mathbf{x}_i - \mu_c\right)^\top \Sigma_c^{-1}\left(\mathbf{x}_i - \mu_c\right) - \frac{1}{2}\log\left(|\Sigma_c|\right) + \log\left(\pi_c\right) \tag{3.165}$$

$$= -\frac{1}{2}\mathbf{x}_i^\top \Sigma_c^{-1}\mathbf{x}_i + \mathbf{x}_i^\top \Sigma_c^{-1}\mu_c - \frac{1}{2}\mu_c^\top \Sigma_c^{-1}\mu_c - \frac{1}{2}\log\left(|\Sigma_c|\right) + \log\left(\pi_c\right), \tag{3.166}$$

which is a quadratic function in $\mathbf{x}_i$ and thus called the quadratic discriminant function.

When the number of training observations is high QDA is preferred over LDA, due to the bias-variance trade-off [1].

Example

In this example, the simulated data set described in Section 1.1 is used. This data set is used because there are only two predictors `Height` and `Weight`. This is an advantage because we can thereby illustrate decision boundaries between the 3 classes in our plots. For the LDA, the three discriminant functions in equation (3.161) are set equal to each other pairwise, which results in the black linear decision boundaries seen in all the plots to the left in Figure 3.9. The top plot on the left is the unbalanced data set described in Section 1.1. The middle plot on the left is the balanced case where for each of the three classes there are 100 observations. The bottom plot on the left is the unbalanced case where there are 100, 500 and 9400 observations respectively from the three classes `1percent`, `5percent` and `94percent`. To the right, similar examples are made for QDA. In this case, the quadratic discriminant functions from equation (3.166) are set equal to each other pairwise, resulting in quadratic decision boundaries. The two classifiers result in decision lines that are close to Bayes decision lines. The plots in the middle and bottom show that we do not need to take extra account of our unbalanced data, if only n is sufficiently large, thereby gaining decision boundaries near Bayes decision lines. This makes sense because the mean values and covariance matrices are estimated assuming that data is drawn from a multivariate normal distribution. Since the simulated data has been drawn from the multivariate normal distribution, estimation will result in values that are close to those of the simulated data when n is sufficiently large. Also, note that QDA is preferred over LDA when n is large as mentioned in the theory. In general because it is more flexible, furthermore in case of this specific example, because it corresponds to the underlying model.

□

The decision lines of LDA look similar to those of logistic regression, this is further supported by [2] which claims that there is little difference in LDA and logistic regression in practice. We thus prefer logistic regression over LDA, as it is based on fewer assumptions. Should our real data be multivariate normally distributed for the predictors and should the classes have different covariance matrices, quadratic discriminant analysis would be an appropriate approach, as the introducing of non-linearity to logistic regression that is needed to model this situation is more complicated.

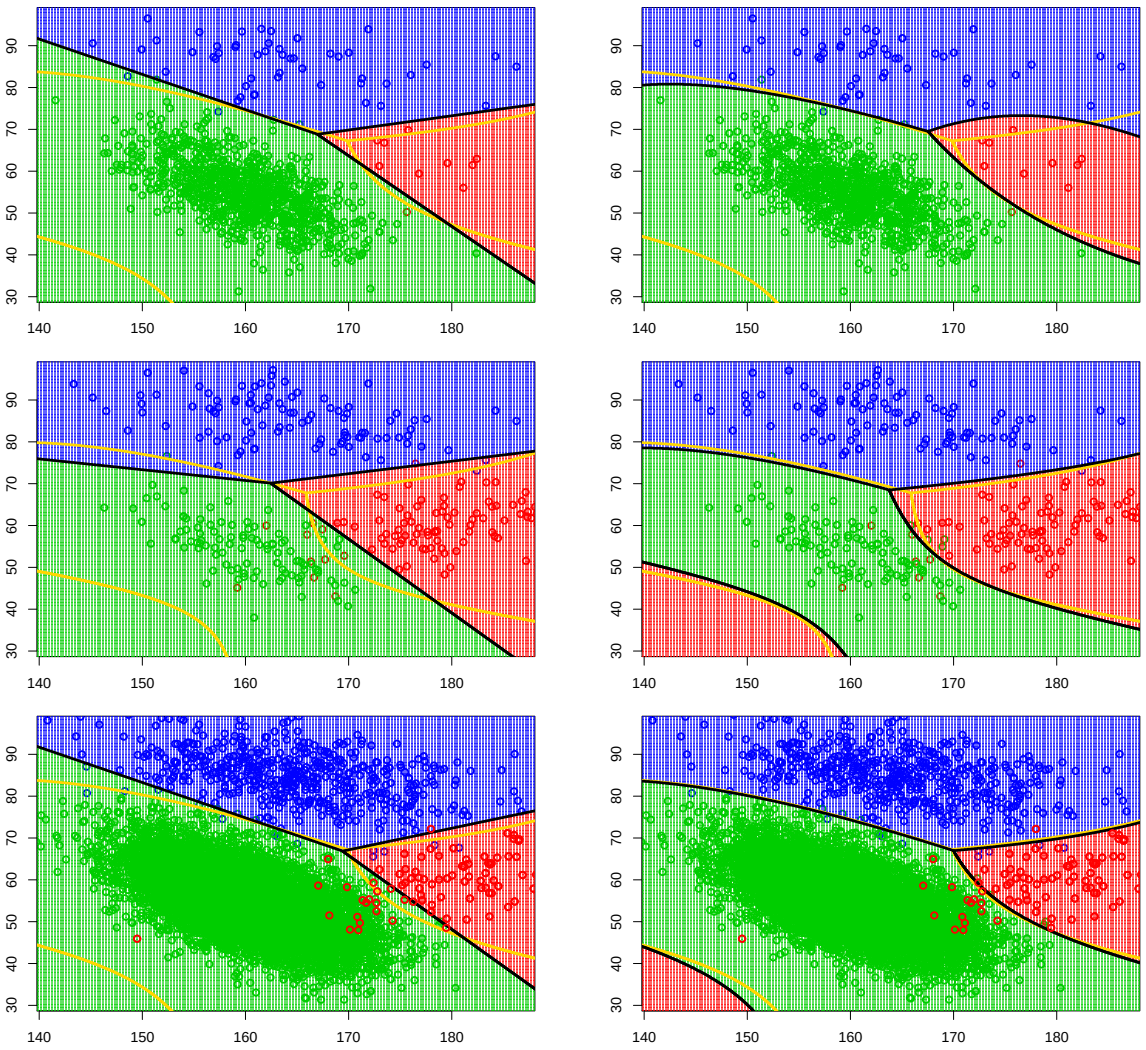


Figure 3.9: All the plots on the left show decision lines using LDA and all the plots on the right show decision lines using QDA. Bayes desision lines are depicted in gold. In the top row, the simulated data from Section 1.1 is used. In the middle row, the balanced case is seen where 100 observations for each of the classes `1percent`, `5percent` and `94percent` are drawn. The plots at the bottom are the simulated data set described in Section 1.1, where 10 times as many observations from each class are drawn.

3.6 Tree Based Methods

Classification trees are another classifier, which can classify an individual's response Y_i based on predictors $\mathbf{x}_i$. First, consider the set, R_0, called the root of the tree which consist of all observations y_i, $i = 1, \ldots, n$, from the training data. The best prediction $\hat{y}^*$ that can be made based on the tree T containing only R_0, is that it belongs to the class that is most frequent among $\mathbf{y}$. Thereby all individuals who do not belong to the most frequent class are miss classified. This can be improved by splitting the R_0 into two disjoint subsets. Split data based on a single binary split of a predictor $\mathbf{x}_{\bullet j}$. The resulting new tree consists of the root and two nodes R_1 and R_2, as there are no further splits of the sets R_1 and R_2 they are called terminal nodes or leaves. The prediction in each leave is denoted $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$, respectively, and all predictions based on each set are again the most frequent class within the set. The best split is based on a greedy algorithm, which means that the best split at a given time is made without regard to whether a better split can be performed later on. For a continuous predictor, select the sets

$$R_1(j, s) = \{i \mid x_{ij} < s\} \qquad R_2(j, s) = \{i \mid x_{ij} \geq s\}, \tag{3.167}$$

which solve the optimization problem

$$\min_{j,s} \Big(\text{Loss}\left(R_1(j, s)\right) + \text{Loss}\left(R_2(j, s)\right) \Big), \tag{3.168}$$

where a loss-function could be the classification error rate Err (4.1). In this case the Err corresponds to the proportion that does not belong to the most frequent class in that region and thus must be minimized in order to get pure response regions. The classification error rate is thus given by

$$\text{Loss}\left(R_m\right) = \text{Err}\left(R_m\right) = 1 - \max_c(\hat{p}_{mc}), \tag{3.169}$$

where $\hat{p}_{mc}$ represents the proportion of subjects belonging to class c for a given region R_m. The best predictor $\mathbf{x}_{\bullet j}$ and the best split s for this predictor are chosen. If the predictor is categorical, the classes for such predictors are split into two groups.

A larger tree is grown by splitting regions R_1 and R_2 respectively in the same manner as we did with the root R_0. This procedure is repeated until each observation has its own leaf or until a predetermined stop criterion is reached, such as if each leaf contains at least 5 observations [1]. A tree of that size will likely overfit the data, thus in order to get a better bias-variance trade-off, the tree is pruned after growing a full tree, which means omitting splits, starting from the leaves. It is done in practice by introducing a penalty based on tree size, thus penalizing large trees. Let T' denote the set of bottom regions $R_j \in T$ (only the leaves of T) and let $|T_i'|$ denote the number of leaves in the sub-tree $T_i \subseteq T$, and solve the optimization problem

$$\min_{T_i \subseteq T} \left(\sum_{R_j \in T_i'} \text{Loss}(R_j) + \alpha |T_i'| \right), \tag{3.170}$$

where the parameter $\alpha \geq 0$ indicates how hard to penalize large trees. As considering all possible subtrees can be computationally overwhelming, this is generally done step-wise by omitting splits starting from the leaves of the tree. This will not always lead to the global minimum, which is not a problem, as reaching the global minimum can lead to over-fitting.

In practice we will use the Gini index to grow a full tree instead of using classification error rate, as it more specifically prioritizes, pure leaves, meaning that each individual leaf should have a low classification error rate.

$$\text{Loss}\left(R_m\right) = \text{Gini}\left(R_m\right) = \sum_{c=1}^{K} \hat{p}_{mc}\left(1 - \hat{p}_{mc}\right). \tag{3.171}$$

This measure can thus be used in order to calculate the purity for a region R_m, since values of $\hat{p}_{mc}$ close to 0 and 1 across the K classes will result in a Gini index value close to 0.

Example

Applying a classification tree to the usual simulated data set is unsatisfactory, as the data is so nicely split that there is no meaningful difference in a "fully grown" tree and a pruned tree. We thus apply the classification tree method to a similarly drawn data set with $n = 10,000$, in which we previously have seen more mixing of `1percent` and `94percent` in eg. Figure 3.9 bottom.

Growing a "fully grown" tree is step one in the process of creating a useful classification tree, we here keep "fully grown" in quotation marks, as a fully grown tree would have n leaves. There is no need to grow the tree further than the point where all leaves are pure. Furthermore, in this example we stop growing even earlier as a "fully grown" tree in this example is a tree of dept 3, for visual reasons. In Figure 3.10 we see a "fully grown" tree on our data. At the root of the tree all observations are classified as `94percent`, resulting in 600 mis-classifications. The "fully grown" tree has the number of mis-classifications reduced to 69. Note that most of the observations end out in leaf 15, where 93.7% of all observations in the data set are classified as belonging to class `94percent`. We use the Gini index to try and minimize the effect of the dominating class `94percent` when growing the tree.

After growing a full tree we prune the tree down to the smallest tree with a total ten fold cross-validated Err within one standard deviation of the overall lowest ten fold cross-validated Err in the "fully grown" tree. The pruning results in the tree depicted in Figure 3.11. This tree mis-classifies 80 observations, but is expected to perform better on new data as the variance of the model is reduced, though introducing more bias. When the tree is pruned we use it as a classifier, and visualize the classification as we usually do with a scatter-plot and decision lines. This visualization can be seen in Figure 3.12, note how the decision lines are all straight lines as a consequence of the binary splits. □

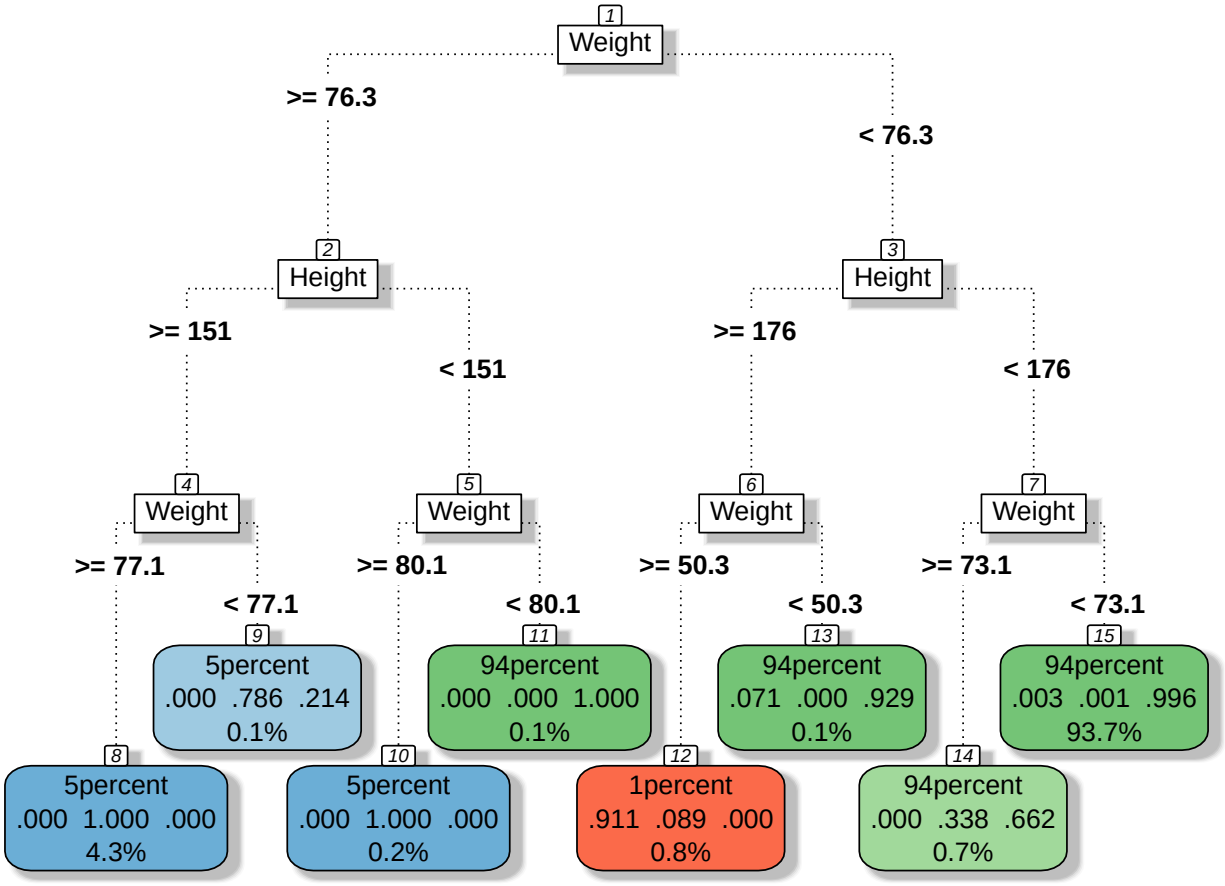


Figure 3.10: A tree grown to dept 3 on the 10,000 observations drawn as the simulated data set. The knots are splits of either `Height` or `Weight`. In the leafs we see in the first line which class is the dominant one, in the middle line it is reported how big a proportion of the observations in a given leaf belongs to the classes `1percent`, `5percent` and `94percent` respectively. The bottom line in a leaf reports the percentage of all observations that end up in a given leaf.

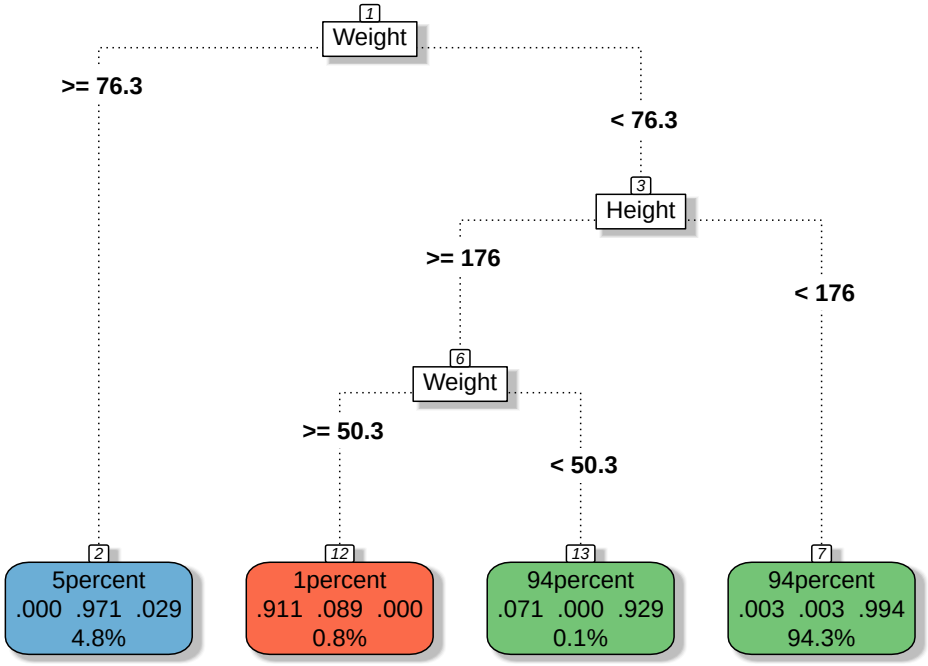


Figure 3.11: The resulting tree after pruning the tree in Figure 3.10. The knots are splits of either `Height` or `Weight`. In the leafs we see in the first line which class is the dominant one, in the middle line it is reported how big a proportion of the observations in a given leaf belongs to the classes `1percent`, `5percent` and `94percent` respectively. The bottom line in a leaf reports the percentage of all observations that end up in a given leaf.

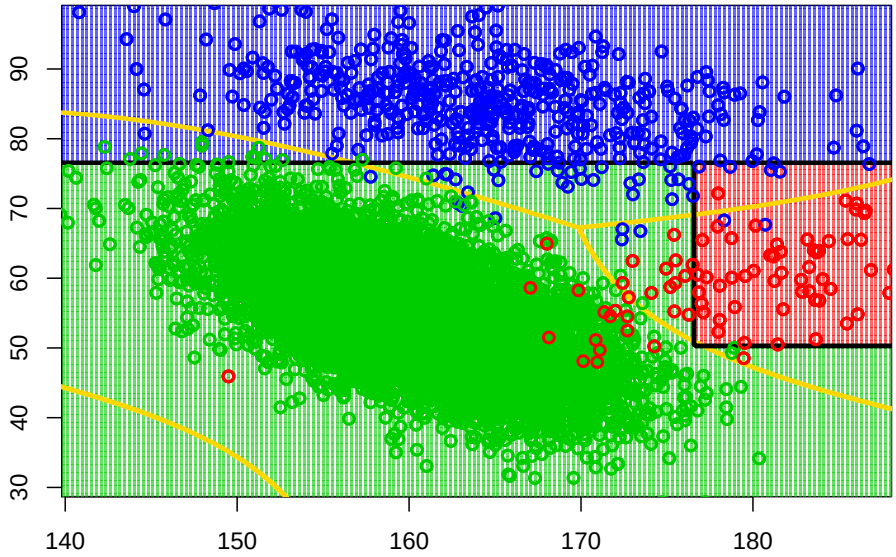


Figure 3.12: The pruned tree from Figure 3.11 applied as a classifier to the data set with $n = 10,000$. Decision lines in black and Bayes decision lines in gold.

An advantage of trees is that, at least when they are pruned, they are easy for health care professionals to use in a clinical setting as they are flowcharts with yes/no options. The real data set is most likely not as nicely separated as the simulated data, and we can thus not expect a tree as simple as the one in Figure 3.11 to be optimal. As a classifier we might expect the trees to overfit when applied to real data, but we saw no tendencies of such in our work with the simulated data.

3.7 k Nearest Neighbors[1]

For a fixed $k = 1, 2, \ldots, n$, the *k Nearest Neighbors (kNN)* classifier compares a new observation $\mathbf{x}_0^*$ to all training observations $\mathbf{x}_i$ for $i = 1, 2, \ldots, n$, chooses the k nearest observations according to some distance measure $d(\mathbf{x}_i, \mathbf{x}_0^*)$ and classifies the new observation's response $\hat{y}_0^*$ by majority vote between the K classes of $\mathbf{y}$,

$$\hat{y} = \underset{c \in \{1,\ldots,K\}}{\arg\max} \sum_{i:\mathbf{x}_i \in N_k(\mathbf{x}_0^*)} w(\mathbf{x}_i, \mathbf{x}_0^*) \mathbb{1}[y_i = c] \,, \tag{3.172}$$

where $N_k(\mathbf{x}_0^*)$ are the k nearest neighbors of $\mathbf{x}_0^*$ according to the distance measure $d(\mathbf{x}_i, \mathbf{x}_0^*)$ and $w(\mathbf{x}_i, \mathbf{x}_0^*)$ is the weight of $\mathbf{x}_i$ relative to $\mathbf{x}_0^*$. In the case of a tie between classes, the class is chosen at random. The weights $w(\mathbf{x}_i, \mathbf{x}_0^*)$ are usually set as a function of the distance, for instance

$$w(\mathbf{x}_i, \mathbf{x}_0^*) = \frac{1}{1 + d(\mathbf{x}_i, \mathbf{x}_0^*)}, \tag{3.173}$$

making training observations closer to $\mathbf{x}_0^*$ more important. Weights can be useful in our project as the occurrence of some of our classes is relatively low (ADHD 5% and ASD 1%).

[1]This section is based on [1, 2, 44].

Without weights the kNN classifier is strongly biased towards a dominant class purely through the dominant class' high occurrence. If these weights are not sufficient, a penalty can be added to equation 3.172. For example, the penalty may be that the right side is multiplied by $(1 - \alpha_c)$, where α_c is the proportion of observations belonging to class c.

The distance measure used to select the k closest neighbors is now discussed. First, note that continuous and ordinal predictors are standardized to have mean 0 and standard deviation 1 so that they can be measured on the same scale. The dissimilarity between two observations is defined by

$$D\left(\mathbf{x}_i, \mathbf{x}_{i'}\right) = \sum_{j=1}^{p} d\left(x_{ij}, x_{i'j}\right), \tag{3.174}$$

where the squared distance is often used as distance measure in case of continuous predictors

$$d\left(x_{ij}, x_{i'j}\right) = \left(x_{ij} - x_{i'j}\right)^2. \tag{3.175}$$

The categorical variables are treated by applying $d\left(x_{ij}, x_{i'j}\right) = 0$, if the j'th predictor has categorical response and the two observations have the same level and if not, then $d\left(x_{ij}, x_{i'j}\right) = 1$.
Most of the classifiers in this project are *off-line* classifiers, meaning that after a model is trained there is no need for the training data when a new observation is to be predicted. This is not the case for kNN which is an *on-line* classifier. Every time a new observation is to be predicted the n distances to all training observations must be computed. Thus, PCA should be considered to reduce the large number of predictors before the kNN is performed especially because we have a huge number of observations and therefore many distances must be calculated, and reducing the number of dimensions would ease each of these calculations. When determining the appropriate value of k for the kNN classifier, cross-validation can be used, for more on cross-validation see Section 4.2.1.

Example

In this example we use the simulated data set from Section 1.1, which involves two predictors. First we try with different choices of number of nearest points k without any kind of weights, neither $w(\mathbf{x}_i, \mathbf{x}_0^*)$ nor α_c. Thus, for a new observation the k nearest neighbors are obtained. This resulted in the plots shown in Figure 3.13. The plot at the top left shows the situation when $k = 1$ and it is seen that the decision boundaries in this case tends to overfit to the training data. In cases where the classes are not as nicely separated as our simulated data is, it will become more apparent. The plot at the top right shows the case when $k = 3$, which agrees overall with the bayes decision boundaries. The two bottom plots show the situation where k is 20 and 100, respectively. Here it is clearly seen that some kind of weight is required to control the unbalanced setup as the red decision area becomes smaller as k increases. In the case where $k = 100$, there will never be any red decision areas, as there are only 10 red observations in the simulated training data. To find the optimal k, we performed 10-fold cross-validation for $k = 1, \ldots, 100$ and based on the

evaluation measure f-score, which will be elaborated in Section 4.1, the result was $k = 3$. Cross-validation will be elaborated in Section 4.2.1. We chose to use the f-score instead of the AUC as there is no threshold in kNN. We have also tried to run a corresponding cross-validation with the weights $w(\mathbf{x}_i, \mathbf{x}_0^*)$ in (3.173) and α_c both separately and at the same time. This resulted in poorer f-scores based on cross-validation and therefore the best result was $k = 3$ without weights. Note that the f-score measure is calculated by considering a class as the event of interest and therefore we have made an f-score for each of the classes, after which the average of these three f-scores is taken to make the final f-score evaluation measure.

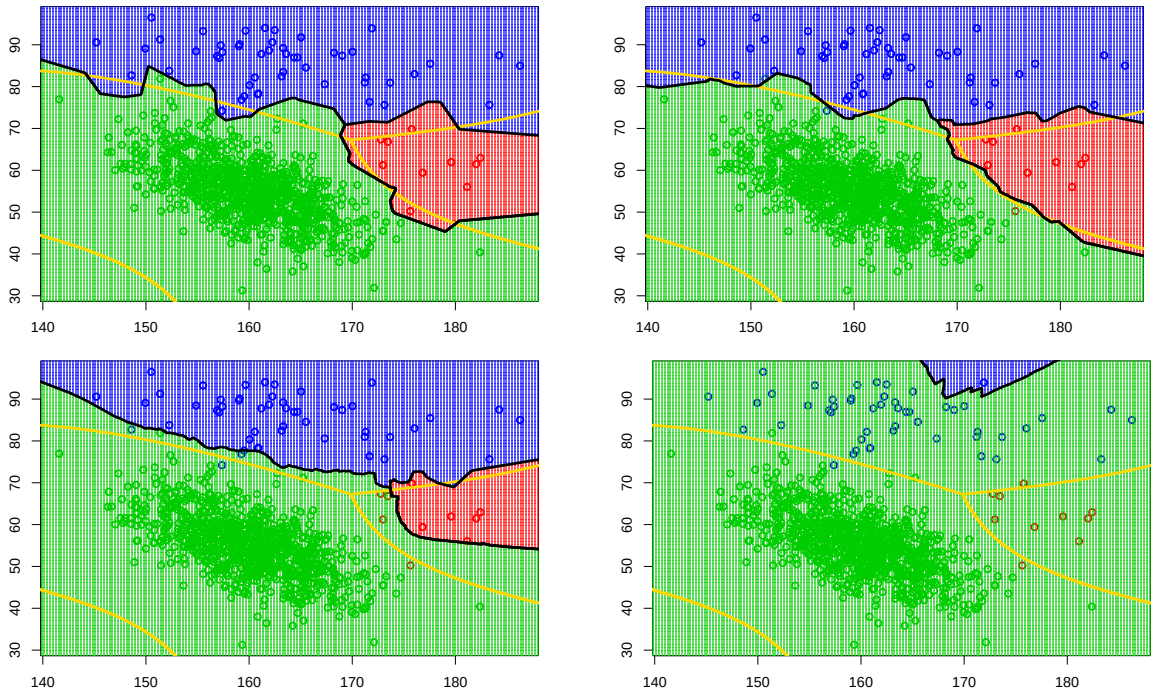


Figure 3.13: kNN applied to the simulated data set with different choice of k and no weights. The plot top left has $k = 1$, the top right plot has $k = 3$, the bottom left plot has $k = 20$ and the bottom right plot has $k = 100$. Bayes decision boundaries are depicted in gold.

Since a low k is preferred when the number of observations for the smallest class is 10 even with weights, we choose to use the simulated data set, where there are 10 times as many observations from the different classes, ie. 100 observations from the class `1percent`, 500 observations from the class `5percent` and 9400 observations from the class `94percent`. A corresponding cross-validation is performed on this data set, which resulted in the best cross-validation score for $k = 17$ when both weights $w(\mathbf{x}_i, \mathbf{x}_0^*)$ and α_c are used. Figure 3.14 shows the case where $k = 17$ and both weights are used for the large data set. It is seen that kNN makes decision boundaries close to the Bayes decision boundaries but visually, Figure 3.14 shows that the green area has generally lower priority than the other two, indicating that the green class penalty is too high but better than no penalty according to cross-validation. However, this kNN run tries to handle the overlay of the two classes `1percent` and `94percent`. This can be seen as the decision area bottom left is red and that small regions are red in the green area. It should be noted that we generally

get high f-score measures based on cross-validation because data is so nicely split. Thus a lower choice of k without any weights will also perform well for training data as not many observations are misclassified hereby.

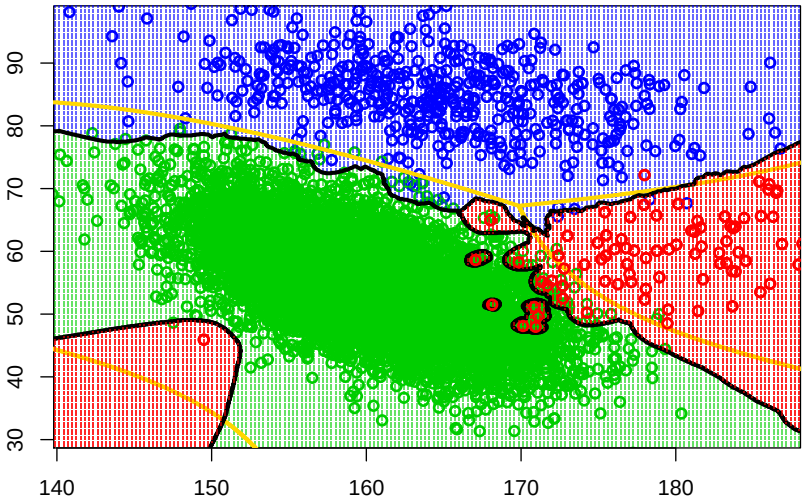


Figure 3.14: kNN applied to the simulated data set with 10 times as many observations from each class. The plot shows the case where $k = 17$ and both weights are used, as it was the best result based on 10-fold cross-validation with the evaluation measure f-score. Bayes decision boundaries are depicted in gold.

□

For our project, we could discard kNN merely on the basis of it being an online classifier, and thus not applicable in a real life setting, as the access to information on all previous patients is restricted, in such a way that a clinician would not be able to use kNN. But it is possible for us to use it in this project as we do have access to data, and we can thus use it in comparison with the other classifiers. The kNN classifier is prone to overfitting, and as there are also issues with the weights, we do not expect kNN to perform well on the real data set.

3.8 Summarizing Remarks

The aspects of logistic regression described in this chapter provides a wide range of modeling customization. The fact that logistic regression produces probabilities instead of just pure classification further adds to our appreciation for this classifier. We choose to proceed with logistic regression and seek to optimize logistic regression classification on our real data set by using splines, LASSO and Ridge regression. Should we seek to compare logistic regression to another classifier, we choose classification trees, as this method is based on a completely different approach. It also seems that classification trees actually do appear applicable to our data, at least on the simulated data set, whereas kNN faces several issues.

4. Model Selection

In Chapter 3 various classifiers were presented. This chapter is about how to choose between all these classifiers and variations of such. Since this master thesis aims at constructing the best predictive model, we need to use one or more evaluation measures to be able to compare such models. A presentation and discussion of different evaluation measures are therefore presented in Section 4.1. Furthermore, when selecting a model, the evaluation measure should be calculated for data that the model is not trained on. This avoids overfitting and two methods that take this into account are cross-validation and bootstrap, which are introduced in Section 4.2. Moreover, it is suspected that some predictors may cause unnecessary noise that will lead to poorer predictions based on an evaluation measure. Thus shrinkage methods were introduced in Section 3.4.3, but these are, as mentioned in the section, likelihood based and therefore best subset selection is introduced in Section 4.3, which can be used both for likelihood based and non likelihood based classifiers.

4.1 Evaluation Measures[1]

When one or more models are built, a evaluation measure is needed to determine which model predicts best and whether the best predictive model gives good prediction at all. Since our setup is a classification problem, it is desired to use an evaluation measure based on the correct classification of observations. An evaluation measure that can be used, after fitting a model f to some training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, is called the training error rate of $\hat{f}$, which is calculated by

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[y_i \neq \hat{y}_i]\,, \tag{4.1}$$

where $\mathbb{1}$ is the indicator function and $\hat{y}_i$ is the prediction from $\hat{f}(\mathbf{x}_i)$. The training error rate is not as interesting as the test error rate

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[y_i^* \neq \hat{y}_i^*]\,, \tag{4.2}$$

[1]This section is based on [1, 35, 36].

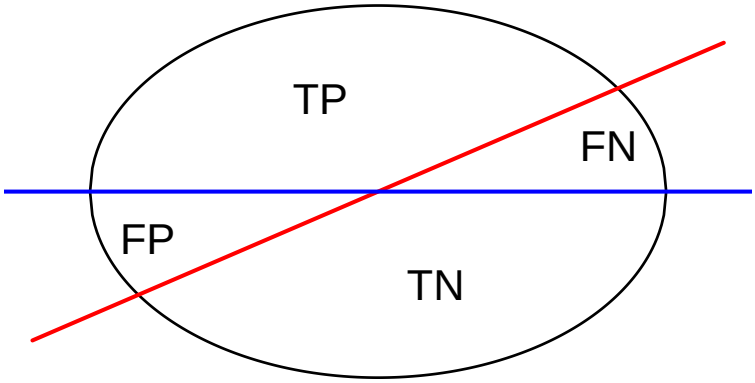


Figure 4.1: Above the blue line: The observation has the event of interest. Below the blue line: The observation does not have the event of interest. Above the red line: Positive prediction. Below the red line: Negative prediction.

where $\hat{y}_i^*$ is the prediction from $\hat{f}(\mathbf{x}_i^*)$, where $\{(\mathbf{x}_1^*, y_1^*), \ldots, (\mathbf{x}_m^*, y_m^*)\}$ are new observations and thus not used to train $\hat{f}$.

We want our models to predict future unknown responses, therefore we prefer evaluating them on other data than what we used to train it with. Such an evaluation can be considered *internal* or *external*. Within the category of internal evaluation lie all evaluation measures that can be performed on the available data set, herein of course the training error rate of (4.1), but also all other methods described later in this section. Even though they can be used for external evaluation, we only perform internal evaluation in this project. External evaluation consists of computing e.g. (4.2) for a data set not only new to $\hat{f}$, but also of different origin. Evaluating $\hat{f}$ on data collected later in time or at another location than the original data will disclose information on the generalizability of the model $\hat{f}$ and thereby evaluating the applicability of the model.

As mentioned, there are several evaluation measures. The training error rate indicates how many observations are incorrectly classified, but since the incidence of ADHD(5%) and ASD(1%) is low, a sensible classifier can be made by classifying all persons to not have any of the mentioned diagnoses. By using such a classifier, 5% and 1% errors are made respectively. Therefore, a precise classifier is made based on the training error rate, but the classifier is of course useless since it does not identify the issues of interest. Another and more useful evaluation measure for our setup should therefore be used. The binary classification problem is considered below, as it applies similarly to the multiple class case where there are several possibilities for a person not to have the event of interest.

Consider Figure 4.1, where the space within the ellipse denotes data that is either training or test data. The ellipse is divided into two parts separated by the blue line, where observations above the blue line have the event of interest, for example ADHD, and observations below do not have the event of interest. In addition, a red line is displayed, which represents a classifier who has respectively predicted a person to have the event of interest or not having it based on data. Observations above the red line are predicted to have the event of interest, for example ADHD, and the opposite is the case for observations below the red line. Thereby, the ellipse is divided into 4 subsets, each

		Actual response			
		Positive	Negative		
Classified by model	Positive	#TP	#FP	PPV	$\frac{\#\text{TP}}{\#\text{TP}+\#\text{FP}}$
	Negative	#FN	#TN	NPV	$\frac{\#\text{TN}}{\#\text{TN}+\#\text{FN}}$
		Sensitivity	Specificity		
		$\frac{\#\text{TP}}{\#\text{TP}+\#\text{FN}}$	$\frac{\#\text{TN}}{\#\text{TN}+\#\text{FP}}$		

Table 4.1: Confusion matrix.

representing one of the following situations

1. True positive: Correct positive prediction (TP)
2. False positive: Incorrect positive prediction (FP)
3. True negative: Correct negative prediction (TN)
4. False negative: Incorrect negative prediction (FN).

An observation belongs to one of the four subsets and based on this, the following so-called basic measures can be calculated, which are used in several evaluation measures within classification. The first basic measure is called sensitivity and describes the proportion of observations that are correctly classified with the event of interest out of all observations which actually have the event of interest, that is

$$Sensivity = \frac{\#TP}{\#TP + \#FN}. \tag{4.3}$$

The sign # means the number of people belonging to the subset. Sensitivity thus describes how good the model is at identifying the observations that have the event of interest. Another basic measure is specificity, which similarly describes how good the model is to identify people who do not have the event of interest

$$Specificity = \frac{\#TN}{\#TN + \#FP}. \tag{4.4}$$

The last two basic measures presented in this project are two measures that are extremely important for the doctor to know in our case. One is called positive predictive value (PPV), which is the probability that a person has the event of interest given that the person is classified as such

$$PPV = \frac{\#TP}{\#TP + \#FP}. \tag{4.5}$$

Similarly, the last basic measure is called negative predictive value (NPV), which is the probability that the person does not have the event of interest given that the model classifies the person as such

$$NPV = \frac{\#TN}{\#TN + \#FN}. \tag{4.6}$$

These basic measures are often gathered in a so-called confusion matrix, as in Table 4.1.

Note that there exists a confusion matrix for each choice of threshold for a threshold based model. When evaluating the models, it is easier to evaluate by 1 measure, thus the so-called F-score measure is introduced, which combines PPV and sensitivity

$$F_\gamma = (1+\gamma^2)\frac{PPV \cdot Sensivity}{(\gamma^2 \cdot PPV) + Sensivity}, \tag{4.7}$$

where $\gamma \geq 0$ is a parameter that determines how much PPV and sensitivity are weighted in the overall measure F_γ. If $\gamma < 1$ more focus is on PPV and if $\gamma > 1$ is the case, then there is more focus on sensitivity. It is interesting for us to try different values for γ such as $\frac{1}{2}, 1, 2$, since both PPV and sensitivity are important to us.

Another evaluation measure is the area under the curve (AUC), where curve refers to the receiver operating characteristic curve (ROC). A ROC curve is a discrete decreasing function of sensitivity and specificity and can be made for all classifiers that require a threshold. A confusion matrix such as Tabel 4.1 can be made for a classifier based on a threshold and a ROC point has the coordinates $(\text{Specificity}, \text{Sensivity})$ that appear from the confusion matrix. If several different thresholds are selected, several confusion matrices and hence more ROC points are obtained. An example of a ROC curve is seen in Figure 4.2, where it is noted that our x-axis ranges from 1 to 0, and therefore it is desired to get points at the top left corner of the plot in order to have both high specificity and sensitivity. Note that a ROC curve on this form always starts in (1,0) and ends in (0,1). The advantage of ROC curves is that two models can be compared, for example, as in Figure 4.2. The figure shows that the red model is better at predicting than the black model, because this model generally has a better trade-off between sensitivity and specificity. Which models the red and black curve represent is elaborated in the example later in this section. ROC curves can therefore both be used to determine which model is preferred, but also contribute to the choice of threshold value which provides a sensible trade-off between sensitivity and specificity. In other cases, it is not as clear which model is preferred and, moreover, a single value is desired to evaluate models instead of a plot and therefore the concept of AUC is introduced. Note that the area of the entire square in Figure 4.2 is $1 \times 1 = 1$. Thus, the AUC evaluation measure is the ratio of the area under the curve to the total area. Generally, a reasonable model is obtained if AUC ≥ 0.7. The AUC value is calculated by adding the area of the polygons under the ROC curve.

Example

This example supports the explanation of ROC curve above this example by specifically considering a logistic regression model. In this example, the `5percent` class is considered to be the event of interest, that is, the positive prediction. The logistic regression model is fitted with the predictors `Height, Height1, Height2` and `Height3`. These variables are chosen specifically because they do not produce a particularly good model, but still a model that is better than random guessing. The simulated data set is so nicely split by design, that including eg. `Height` and `Weight` produces a model with AUC almost 1, this can also be imagined by examining the top right plot in Figure 3.3. Note that

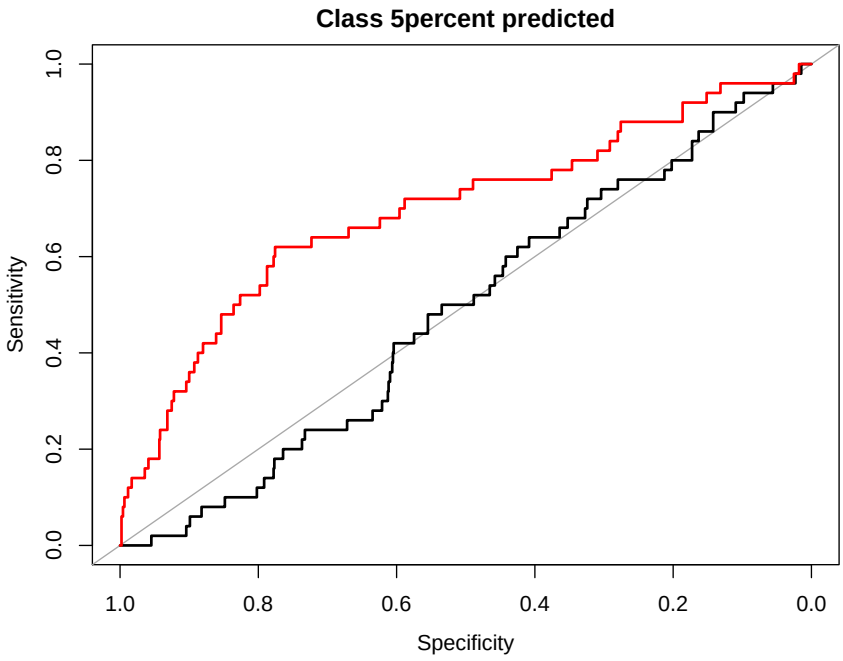


Figure 4.2: ROC curves from a random model (black) and a logistic regression model (red) `5percent`∼`Height+Height1+Height2+Height3`.
Note that the x-axis indicating sensitivity goes from 1 to 0.

for this simulated data set there are 1000 observations, where 5% have observed response `5percent`. In order to compare the model with something we know predicts worse, we compare this model with a classifier that returns probabilities uniformly between 0 and 1.

Based on the logistic regression model, the probability of a person belonging to the `5percent` class is calculated and compiled in a list as in Table 4.2, where the first mentioned person has the largest probability score, in this case it is the greatest value for $P(Y_i = \texttt{5percent} \mid \mathbf{x}_i)$.

We now select a range of threshold values between 0 and 1. Then all persons are classified to have positive or negative response based on their probability score for one specific threshold value and a similar list is made for other threshold values.

> If the probability score $\geq$ threshold, then the person is predicted as having a positive response.

Based on this, a confusion matrix is made for each threshold value, one such is seen in Table 4.3. As we in our project both want high sensitivity and specificity, we have chosen to show the confusion matrix with a threshold of 0.057, which is the threshold that gives the highest value of sensitivity + specificity.

The ROC curve seen in Figure 4.2 can now be made based on these confusion matrices, where sensitivity and specificity are included. To compare the logistic regression model, the ROC curve for the random model is also made. It is seen that the logistic regression

ID	Observed Class	Probability
457	Negative	0.61
2	Negative	0.56
53	Positive	0.54
23	Positive	0.45
60	Positive	0.42
10	Negative	0.35
62	Negative	0.32
48	Positive	0.31
173	Negative	0.26
253	Negative	0.26
⋮	⋮	⋮

Table 4.2: The list is a summary of the individuals ID number, their observed response class and the probability score calculated by logistic regression. The people are sorted so that the person with the highest probability score is at the top of the list.

		Actual response			
		Positive	Negative		
Classified by model	Positive	31	213	PPV	0.13
	Negative	19	737	NPV	0.98
		Sensitivity	Specificity		
		0.62	0.78		

Table 4.3: Confusion matrix for the logistic regression model `5percent` ∼ `Height+Height1+Height2+Height3` with a threshold of 0.057, which is the threshold that gives the largest value of sensitivity+specificity.

model predicts better based on the ROC curve than the random model since the ROC curve for the logistic regression model is always above the random model and thus has a better tradeoff between sensitivity and specificity.

Other times, the distinction is not as clear and therefore the measures, other than sensitivity and specificity, should also be taken into account when evaluating a ROC curve. Thus, the results from the logistic regression model `5percent∼Height+Height1+Height2+Height3` and the random model are gathered in a table such as Table 4.4, which is used to select models besides just the ROC curve. Table 4.4 shows that the logistic regression model generally predicts better than the random model as it has higher basic and evaluation measures. □

	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
Random	0.044	0.946	0.360	0.592	0.054	0.079	0.149	0.485
LogReg	0.127	0.975	0.620	0.776	0.151	0.211	0.349	0.696

Table 4.4: Comparison of the logistic regression model `5percent∼Height+Height1+Height2+Height3` and the random model. The comparison consists of basic measures and evaluation measures mentioned in this section. A threshold of 0.057 is chosen for the logistic regression model, and a threshold of 0.58 for the random model.

For classification the F-score is always applicable, which is why we use it when comparing across classifiers where some are not probability based, or if a threshold is set for the probability based models. When we compare probability based models (all of our logistic regressions), we choose to use AUC, as this gives a measure of the models overall predictive performance regardless of threshold.

All of the measures can be calculated for training data, whereby good results are often achieved due to overfitting the model. Instead, it should be calculated on data that the fitted model has not seen. Cross-validation and bootstrap are approaches that can be used to try and address this dilemma.

4.2 Cross-Validation and Bootstrap

This section is about cross-validation and bootstrap, which also can be used when evaluating a model. Instead of using all training data to fit a model, cross-validation and bootstrap can be used such that all (almost all, for bootstrap) training data is still used, but data used to train the model is not used to evaluate it.

4.2.1 Cross-Validation

Cross-validation is an internal validation tool used to estimate the external validity of a given procedure. If one had, as we do in this project, a fair amount of data, a way to evaluate a model on "new" data, could be to randomly take out a part of the training data for validation use. This approach is computationally inexpensive and simple. Some considerations about the size of such a holdout data set should be done. One would usually use as much data as possible to train the model, but the holdout data set should also be large enough for the evaluation to make sense.

The main concern with the holdout approach is the consideration that we would not be sure whether the evaluation results obtained were due to chance and only applicable to this particular split of data.

A way to overcome this is to do the split more than once. We split the data into k subsets. For each subset we fit the model on the remaining $k-1$ subsets and predict the responses not used in training. This is done k times and the mean of the evaluation measure is obtained. This is called *k-fold cross-validation*

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} \text{Eval}_i. \tag{4.8}$$

With the vast amount of computational power available today, it could be possible to perform a k-fold cross-validation where the amount of data used for training is maximized to the extend where $k = n$ and only one observation is predicted based on a model trained on all remaining data, this is called *leave-one-out cross-validation*. Choosing this approach makes the estimate of the evaluation measure almost unbiased. But as the training sets in leave-one-out cross-validation are almost identical, the models fitted are highly correlated, and as the mean of highly correlated quantities have a high variance,

leave-one-out cross-validation results in an unsatisfactory bias variance trade-off in the estimate of the evaluation measure. We use 10 fold cross-validation, as [1] states, that this produces a good balance of variance and bias.

4.2.2 Bootstrap

Another approach to selecting parts of training data is by using bootstrap. Assume training data $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ is given. Bootstrap is about creating a new data set by randomly drawing observations from the original training data with replacement. This is done n times, such that a bootstrapped data set has the same size as the original training data set, where an observation in the bootstrapped data set can occur several times. The advantage is that B bootstrapped datasets can be used to build a model where there is still test data that the individual models have never seen.

Example

In Table 4.4, various basic and evaluation measures were calculated for the logistic regression model `5percent~Height+Height1+Height2+Height3`. In this example, it is desired to compare these basic and evaluation measures with some found by 10-fold cross-validation and bootstrap. For 10-fold cross-validation, data is divided randomly into 10 parts, one of the parts being used for evaluation and the remaining parts are each time used to train a model. The final basic measures and evaluation measures are the average of the different basic and evaluation measures. For bootstrap it was chosen to make $B = 1000$ bootstrap data sets, where some observations can occur several times. These bootstrapped training sets are used to train 1000 different models, after which they are evaluated using the observations not included in the certain bootstrapped data sets. Again, the average of all these basic and evaluation measures is finally taken to arrive at final basic and evaluation measures.

Table 4.5 shows basic and evaluation measures for the logistic regression, for the logistic regression where cross-validation is used, and the bootstrapped logistic regression model respectively. It is seen that the values are generally higher for the logistic regression, which is suspected to be due to overfitting as the same data set has been used for training and testing. To investigate this we drew $100,000$ new observations from the same distributions as our simulated data set. We fitted the model `5percent~Height+Height1+Height2+Height3` to $90,000$ of these new observations and evaluated on the remaining $10,000$, and reported the basic and evaluation measures as the *pseudo truth*.

We see in Table 4.5, that all the basic and evaluation measures are fairly close, this might mostly be due to the ill construction of the example. The simulated origin of the data set might interfere with the results from cross-validation and bootstrap, and we further do not trust the "pseudo truth" as it might just emulate bootstrapping, at least if n is sufficiently high for the simulated data set from which the bootstrap samples are drawn.

	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LogReg	0.127	0.975	0.620	0.776	0.151	0.211	0.349	0.696
CV 10 fold	0.151	0.972	0.650	0.703	0.174	0.229	0.354	0.667
BS 1000	0.116	0.971	0.590	0.715	0.136	0.185	0.301	0.647
Pseudo truth	0.123	0.969	0.489	0.818	0.144	0.196	0.306	0.682

Table 4.5: Comparison of the logistic regression model 5percent ∼ Height+Height1+Height2+Height3. Various basic and evaluation measures are gathered here based on respectively training data, 10-fold cross-validation and 1000 bootstrap samples. Pseudo truth is the evaluation measures of a fit to a new data set with 90,000 observations drawn from the original distributions that our simulated data set was drawn from evaluated on 10,000 observations again drawn from the same distribution.

□

It seems that testing cross-validation and bootstrap on simulated data, at least data simulated the way we have done, produces inconclusive results. We still do trust cross validated and bootstrapped evaluation measures to be better estimates of the external evaluation measures than the training evaluation measure, based on among others [2].

4.3 Best Subset Selection

For non-likelihood based classifiers such as kNN and LDA we can not use LASSO to choose which predictors to include in order to get the best predictive model. Another method called best subset selection can be used, which is applicable on all the classifiers we have presented, also the likelihood based classifiers.

The idea in best subset selection is that p models are fitted with only one predictor in each model, then the best model is selected based on which one has the best evaluation measure. Similarly, a number of models are fitted that include exactly two predictors ie. in total

$$\binom{p}{2} = \frac{p!}{2!(p-2)!} = \frac{p(p+1)}{2} \tag{4.9}$$

models. Again, choose the best model based on an evaluation measure. The same procedure is also performed for $3, 4, \ldots, p$ predictors. The best model among the $p+1$ models (including the null model) is chosen based on the evaluation measure, preferably cross validated or bootstrapped. The best model is among the 2^p models and therefore p should not be much larger than 40, as it is beyond the limit of our accessible computational power, but also due to the possibility of overfitting the model when fitting that many models.

4.3.1 Stepwise Selection

To avoid fitting 2^p models, when p is large, forward stepwise selection can be performed. First, consider the null model M_0. Then build p new models by including one predictor

and choose the model that provide the best evaluation measure as M_1 . Then, the model M_1 is considered, and the remaining $p-1$ predictors are added correspondingly in turn to this model and the predictor that causes the best evaluation measure is added to the model M_1, thus creating the model M_2. This process is repeated until the full model, where all the predictors are included, is reached. Finally, one of the $p+1$ models are chosen based on a cross-validated or bootstraped evaluation measure. Thus, one has to fit way less models by using stepwise selection as the models chosen by stepwise selection are among the 2^p models. Another method of selecting variables is by using backward stepwise selection, which is very similar to forward stepwise selection. Instead of starting with the null model, start with the full model and instead of adding predictors, they are removed one by one, again based on a evaluation measure.

Example

As an example we apply forward and backwards stepwise selection to our simulated data set, including the variables `Height`, `Height1`, `Weight`, `Weight1`, `Uniform`, `Uniform1`, `Normal` and `Normal1` after which the two variable selection algorithms are compared to LASSO.

To compare the three approaches we generate a new class called `BernLogReg` based on logistic regression, such that we know which predictors we should expect to be selected. The class `BernLogReg` is drawn from a Bernoulli distribution with each observation having the probability

$$P(Y_i \in \texttt{BernLogReg} \mid \mathbf{x}_i) = \frac{\exp(2.2 - 0.14\texttt{Height} + 0.18\texttt{Weight} + 0.5\texttt{Uniform1})}{1 + \exp(2.2 - 0.14\texttt{Height} + 0.18\texttt{Weight} + 0.5\texttt{Uniform1})}. \tag{4.10}$$

As we use ten fold cross-validation the simulated data set is too small, since some folds end up containing no observations from the class `BernLogReg`, we thus increase n to $10,000$. We use AUC to determine which variables to include/exclude, but when comparing the models with different number of predictors we use 10-fold cross-validated AUC.

In Table 4.6 the variables selected by the two algorithms and LASSO are marked by x'es. All three models include `Weight`, `Weight1` and `Uniform1`, even though `Weight1` was not part of the original formula. This is interesting as the correlation resulting from `Weight1` being a transformation of `Weight` should make the selection methods discard at least one of them, like forward and backward stepwise selection did with `Height` when including `Height1`.

It seems that the choice made by LASSO to exclude the information from the height variables resulted in a slightly smaller cross-validated AUC, but as the AUC's differ only from the third digit, we can not conclude whether any of the methods are better than the other based on this example.

We changed the coefficients in (4.10) in numerous different ways, but in most cases the three variable selection approaches chose the exact same variables, and mostly the correct ones thus making the shown example the most interesting. □

	Height	Height1	Weight	Weight1	Uniform	Uniform1	Normal	Normal1	CV AUC
LASSO			x	x		x			0.8853
Forward		x	x	x		x			0.8884
Backward		x	x	x		x			0.8884
Truth	x		x			x			1

Table 4.6: The variables selected by LASSO, forward step-wise selection and backward step-wise selection marked by an x. The truth row is marking by x, the predictors on which the class `BernLogReg` is actually based. The cross-validated AUC for each of the models is reported in the far right column.

All three methods seem useful, and in most of our simulations they do find the correct variables. We though once again end up inconclusive in our example construction, as we do not have enough reasons to prefer any of the three methods. According to [45], step-wise selection is not a good idea, but most of his arguments are founded in explanatory modeling or are against the use of p-values as decision basis for variable inclusion/exclusion. Since the three methods seem to produce the same, and [45] advises against step-wise selection we choose to go forward with only LASSO.

5. Application to Real Data

This chapter presents the data-management and analysis of a data set commissioned by the Research unit for Child and Adolescent Psychiatry and The Psychiatric Research Unit, North Denmark Region Psychiatry, Aalborg, Denmark. Section 5.1 describes the data sources and general characteristics of the data set. Section 5.2 contains the application of the theory in chapters $2-4$ to the aforementioned data set.

5.1 Data

Here, we go through the different public registers and the variables chosen for the analysis.

5.1.1 Registers & Variables

In this section we first present the 8 registers used to define variables in this master thesis. At the bottom of each register description we list the variables originating from the particular register. A detailed description of how the predictors are specifically made from the variables of the registers can be seen in Appendix B. Finally, after all the registers and variables are presented Table 5.1 collects the most important information of this section and Appendix B. The table shows all predictors, their type, a coverage period, and a brief description of what a predictor indicate.

The Danish Registers can be linked uniquely through the personal CPR-number, making the combination of data across registers possible, including linking persons to their parents.

CPR

The Central Register of Persons (CPR, det centrale personregister) holds a large amount of administrative data on the Danish population, such as where they live and who they and their families are. The register was founded in 1968 and linked all existing information, such that people born before 1968 also were included. We expect the CPR register to be almost complete and correct, as so many aspects of people's lives depend on the data to be correct. Thus, each individual in the register has a self interest in the data being

correct, as well does the Danish administration. The CPR register contributes in this thesis to the variables: `PNR`, `M_PNR`, `F_PNR`, `M_Age`, `F_Age`, `Sex` and `BirthYear`.

LPR, MiniPas, LPR-PSYK & DCPR

In Denmark the public health care system covers everybody and it is free. Even though it is free, there is still a payment system in place such that each individual treatment facility gets funded according to the amount of treatment they deliver. For the purpose of both this payment scheme and research, the national patient registers were founded in 1976, comprising of all contacts to the Danish Regional health services begun or finished after 1977. The national health registers are seen as four different registers: the "lands patientregister" (LPR), which is the main register that today actually holds all information, but until 1995 psychiatric contacts were registered in a separate register "det centrale psykiatriregister" (DCPR). As the use of a separate psychiatric register still was needed after 1995 a special table in the LPR was made called the "lands patientregister-psykiatri" (LPR-PSYK). The fourth register is, like LPR-PSYK, an integrated part of the LPR, but as private health care facilities conducting work for the national health service, report their activities to the national health service in a different manner than the public facilities, all these activities can be seen as a separate register or table called the "MiniPas-LPR" (MiniPas). All four registers are seen as complete when it comes to activities paid for by the Danish health care system.

In our data set the MiniPas is nested within the LPR, whereas the DCPR and LPR-PSYK registers are delivered separately. Thus we go through three tables each time we seek information from the LPR. The national health register is used to create the variables: `ADHD`, `ASD` and `Jaundice`.

MFR & L_FOED

Every time a child is born in Denmark, given that the birth is conducted in an official setting, the midwife records several pieces of information in the "medicinske fødselsregister" (MFR). The MFR is, like LPR-PSYK, a part of the LPR, and is delivered as a separate table. Until 1995 the MFR was not a part of the LPR, and consist of data digitized from paper questionnaires filled by the midwives. This part of the register, covering births from 1973 to 1994, is called the "register over levendefødte" L_FOED, meaning the register of live-born, thus only covering the births resulting in a living child. We get the following variables from MFR/L_FOED: `M_Age`, `F_Age`, `M_BMI`, `M_Smoking`, `GestAge`, `Ext_Preterm`, `Ver_Preterm`, `Mod_Preterm`, `Visit_Mid`, `Visit_Doc`, `Visit_Spe`, `Sepsis`, `M_Spon_Abort`, `Parity`, `Cont_Stim`, `Malformations`, `Sectio`, `B_Length`, `B_Weight`, `Apgar5minOK`, `Med_Initiate`, `Epidural`, `Infections` and `In_Asfyxi`.

LSR

The Register of Medicinal Product Statistics or "Lægemiddelstatistikregisteret" (LSR) consist of information on all prescriptions filled in a Danish pharmacy. The purpose of

the register is to monitor the use of drugs. Thus the validity and quality of the register is protected by law, making it illegal not to report or to report incorrectly to the register. The register covers from 1995. We use the LSR to create the variables: `M_ADHD_Meds`, `F_ADHD_Meds`, `M_Alc_Meds`, `F_Alc_Meds`, `M_Drugs_Meds` and `F_Drugs_Meds`.

Variables	**Covering**	**Short description**
Binary:		
`Sex`	1968-	1 indicates that the subject is male and 0 indicates female
`ADHD`	1976-	The subject received a diagnosis of ADHD before the age of 18
`ASD`	1976-	The subject received a diagnosis of ASD before the age of 18
`M_ADHD_Meds`	1995-	The mother got ADHD medicine before the subject's 1st birthday
`F_ADHD_Meds`	1995-	The father got ADHD medicine before the subject's 1st birthday
`M_Alc_Meds`	1995-	The mother got anti-alcohol medicine before the subject's 1st birthday
`F_Alc_Meds`	1995-	The father got anti-alcohol medicine before the subject's 1st birthday
`M_Drugs_Meds`	1995-	The mother got anti-drug medicine before the subject's 1st birthday
`F_Drugs_Meds`	1995-	The father got anti-drug medicine before the subject's 1st birthday
`M_Smoking`	1991-	Smoking mother during pregnancy
`Ext_Preterm`	1973-	Gestational age at birth within the weeks $[22, 27]$ (rounded)
`Ver_Preterm`	1973-	Gestational age at birth within the weeks $[28, 31]$ (rounded)
`Mod_Preterm`	1973-	Gestational age at birth within the weeks $[32, 37]$ (rounded)
`Cont_Stim`	1978-96, 1999-	The mother got contraction stimulation during birth
`Epidural`	2000-	The mother got an epidural during birth
`Med_Initiate`	1991-	The birth was medically induced
`Sectio`*	1978-	The subject was born by cesarean section
`Apgar5minOK`	1978-	Acceptable APGAR score 5 minutes after birth (7,8,9 or 10)
`Malformations`	1978-86, 1991-	The subject was born with a malformation
`In_Asfyxi`	1997-	The subject experienced asfyxia in uterus during birth
`Sepsis`	1997-	The subject experienced sepsis in uterus during the pregnancy
`Infections`	1997-	The subject had infections at birth
`Jaundice`	1976-	The subject is diagnosed with jaundice within the 1st year after birth
Categorical:		
`Parity`*	1973-	Mothers births including the subject (levels: $1, 2, 3, 4, 5, \leq 6$)
Continuous:		
`M_Age`	1968-	Age of the mother when she gave birth to the subject
`F_Age`	1968-	Age of the father when the subject was born
`M_BMI`	2003-	BMI for the mother at first doctor visit during pregnancy
`M_Spon_Abort`	1997-	Number of previous abortions for the mother
`Visit_Mid`	1978-	Number of visits to midwife during pregnancy
`Visit_Doc`	1978-	Number of visits to doctor during pregnancy
`Visit_Spe`	1978-	Number of visits to special doctor during pregnancy
`BirthYear`	1968-	The year that the subject was born
`GestAge`	1973-	Gestational age measured in days
`B_Length`	1973-	The subject's length at birth in centimeter
`B_Weight`	1973-	The subject's weight at birth in gram

Table 5.1: The variables used in the thesis, constructed as described in Appendix B. For the binary variables 1 indicate yes and 0 indicate no. *Note that the variable `Sectio` can be misleading in the period $1978 - 1996$ and that `Parity` can be misleading in the period $1973 - 1996$.

5.1.2 Initial and Exploratory Data Analysis

The commissioned data set consists of individuals born between the 1st of January 1977 and the 31st of December 2012, who have lived in Denmark for at least two years between the 1st of January 1977 and the 31st of December 2018. The members of this population are followed from birth until they turn 18, or until the 31st of December 2018, emigration or death. Because of the hereditary components of ADHD and ASD we also have information on the parents of these individuals, this information is only limited by the temporal coverage of the registers.

We have data on 3,260,957 subjects and 1,262,911 of their mothers and 1,230,674 of their fathers.

The outcomes of interest are the diagnoses of ASD and ADHD registered as primary or secondary diagnosis not given in an emergency department. They are defined in detail in Appendix B. Note that for those not followed until they turn 18 (censored observations) and those not receiving a diagnosis in their follow-up period are set as "no diagnosis", although they could potentially still receive a diagnosis before turning 18. In the left plot in Figure 5.1 we have plotted histograms over the number of new ASD diagnoses in children/adolescents under the age of 18 in Denmark, this is known as the incidence. Note that a subject need not be living in Denmark at the time of first diagnosis. This incidence is merely diagnoses given to a subject who have not previously gotten a diagnosis in Denmark. Likewise the incidence of ADHD is depicted in the plot to the right in Figure 5.1. We see that the incidence of both diagnoses increases steadily in the new millennium. Though it might seem that ADHD levels out from 2010 the incidence could rise from 2013-2017 without this plot showing it, as we stop including subject in the cohort in 2012. Thus the population of Danes under 18 in these plots is reduced by a whole year's births each year, making the incidence in 2017 for the population in the age-range 5-18. It is interesting though that the incidence of ASD keeps increasing, even though the cohort gets smaller. Furthermore, we know from the psychiatric consultant on this project, Marlene Briciet Lauritsen[1], that ASD can be diagnosed even before the age of three, whereas ADHD is usually not diagnosed before the school age of five to six years. Thus the flatting out of the ADHD incidence might be correct, but the ASD incidence should be expected to rise even faster than depicted in Figure 5.1.

[1]Child and Adolescent Psychiatry, Region of Northern Jutland Psychiatry, Aalborg, Denmark

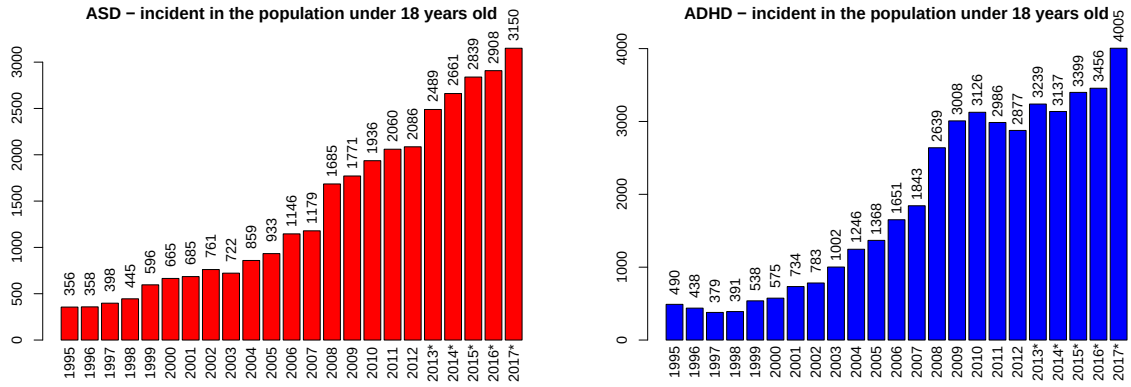


Figure 5.1: Number of incident cases of ASD and ADHD in the Danish population under the age of 18. *Note that no new subjects are included into the study from 2013 and forward.

In Figure 5.2 we see the prevalence, which we have chosen to be the proportion of the 5-18 year old Danes diagnosed with ASD or ADHD. We see that the prevalence is steadily increasing every year. Note that for a subject to contribute in a given year, the subject needs to be between the age of 5 and 18, be alive and living in Denmark at some point in that year. For the subject to belong to a diagnosed group, be it ASD or ADHD, the subject needs to have gotten a diagnose of ASD or ADHD, respectively, prior to or in a given year.

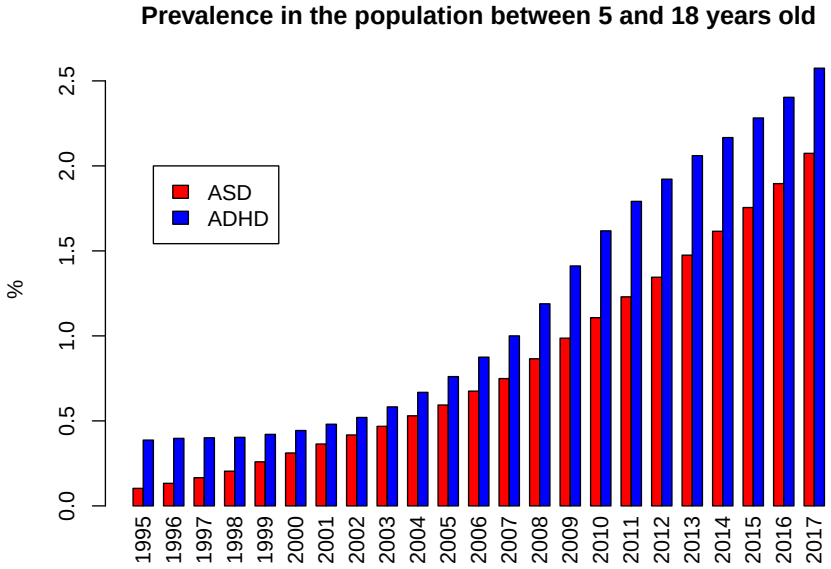


Figure 5.2: Percentage of the Danish population between the ages of 5 and 18 with a diagnosis of ASD or ADHD.

Not all the constructed variables are available for all $3,260,957$ subjects. If a child for instance is a refugee alone in Denmark, we have no knowledge about them in the MFR

as they were born outside Denmark, and we have no knowledge of such a child's parents, as they do not have a CPR number, we thus define the data set **AllObs**.

> **AllObs**: We have removed subjects with only information about `Sex` and `BirthYear` from the data set, such that the data set now consist of 2,461,082 subjects, where we have at least one of our variables not missing. For this data set 1.5% have a diagnosis of ASD and 2.0% have a diagnosis of ADHD. This data set will henceforth be referred to as **AllObs**.

In Table 5.2 and Table 5.3 we present descriptive statistic for all the variables in Table 5.1, based on the subjects in **AllObs**. All continuous variables are summarized by their mean and standard deviation and the binary and categorical variables by a count and percentage. To compare the difference between the diagnosed and undiagnosed we have calculated the odds-ratio (OR). The odds-ratio is defined by $OR = \left(\frac{p_1}{1-p_1}\right) / \left(\frac{\tilde{p}_1}{1-\tilde{p}_1}\right)$, where p_1 is the probability of a subject belonging to class 1 given one predictor value x, and $\tilde{p}_1$ is the probability of a subject belonging to class 1 given another value of the same predictor $\tilde{x}$. For logistic regression it is thus calculated by

$$OR = \frac{\left(\frac{p_1}{1-p_1}\right)}{\left(\frac{\tilde{p}_1}{1-\tilde{p}_1}\right)} = \frac{\left(\frac{P(Y_i=1|x)}{1-P(Y_i=1|x)}\right)}{\left(\frac{P(Y_i=1|\tilde{x})}{1-P(Y_i=1|\tilde{x})}\right)} = \frac{\exp(x\beta)}{\exp(\tilde{x}\beta)} = \exp(\beta)^{x-\tilde{x}}, \tag{5.1}$$

where $P(Y_i = 1 \mid x) = \frac{\exp(x\beta)}{1+\exp(x\beta)}$ from (3.104). Thus for Table 5.2 and Table 5.3, we report OR as $\exp(\beta)$ for all predictors, indicating how much a subject's odds is multiplied if the given predictor is changed by 1. Note that we calculate the OR by fitting a logistic regression for all subjects with only the one predictor.

The first thing worth noting in the two tables is the vast amount of missingness, many variables are missing over half the observations. Secondly, we note that having either ASD or ADHD severely increases the risk of having the other to a degree where a fifth of ADHD patients also have ASD and a third the other way around. This overlap may be the reason we see the same tendencies in some of the ORs of the predictors, such that the same predictors are protective for both diagnosis and the same predictors are harmful. This is not the case for the parental age `M_Age` and `F_Age` that seem to be protective for ADHD and harmful for ASD. We furthermore see that `M_ADHD_Meds` and `F_ADHD_Meds` have much higher ORs for ADHD than for ASD, which should be expected. Another odd observation is that the parents being treated with medication for alcohol or drug abuse is protective for ASD. Finally note that the only predictor that seem not significant is `Infections`, as 1 is within the confidence intervals (this also seem the case for birth weight, but this is likely due to the measuring of weight in grams).

ASD	No ASD	ASD	Missing(%)	OR	95% CI	
n	2425033	36049				
Sex-male	1234433 (50.9)	26676 (74.0)	0.0	2.745	2.681	2.811
ADHD	39234 (1.6)	10268 (28.5)	0.0	24.219	23.622	24.831
AgeAtADHD	10.63 (3.94)	9.55 (3.58)	98.0	0.93	0.924	0.935
M_Age	28.55 (5.09)	29.40 (5.15)	0.4	1.033	1.031	1.035
F_Age	31.49 (5.89)	32.29 (6.15)	2.4	1.022	1.021	1.024
M_ADHD_Meds	28427 (2.2)	1854 (6.3)	47.3	2.905	2.768	3.05
F_ADHD_Meds	24101 (1.9)	1306 (4.4)	47.4	2.38	2.248	2.519
M_Alc_Meds	49534 (3.8)	662 (2.2)	46.2	0.573	0.53	0.619
F_Alc_Meds	109527 (8.2)	1456 (4.9)	44.5	0.572	0.543	0.604
M_Drugs_Meds	18576 (1.5)	268 (0.9)	47.3	0.617	0.547	0.697
F_Drugs_Medi	22943 (1.8)	288 (1.0)	47.1	0.537	0.478	0.604
M_Smoking	263005 (20.8)	7158 (25.4)	47.6	1.294	1.259	1.329
M_BMI	24.27 (4.94)	25.05 (5.58)	78.4	1.029	1.025	1.032
M_Spon_Abort	0.18 (0.45)	0.19 (0.46)	58.0	1.037	1.009	1.066
Parity			10.9			
First Birth	975514 (45.2)	16950 (50.6)				
Second Birth	794913 (36.8)	11263 (33.6)		0.815	0.796	0.835
Third Birth	288216 (13.3)	3915 (11.7)		0.782	0.755	0.81
Fourth Birth	71655 (3.3)	993 (3.0)		0.798	0.748	0.851
Fifth Birth	18966 (0.9)	279 (0.8)		0.847	0.752	0.954
Sixth or over	10519 (0.5)	113 (0.3)		0.618	0.513	0.745
Ext_Preterm	3944 (0.2)	140 (0.4)	15.4	2.171	1.834	2.571
Ver_Preterm	12254 (0.6)	329 (1.0)	15.4	1.645	1.474	1.836
Mod_Preterm	176052 (8.6)	3480 (10.4)	15.4	1.23	1.187	1.275
Visit_Mid	4.40 (2.05)	5.13 (2.00)	14.3	1.172	1.166	1.177
Visit_Doc	1.80 (1.53)	2.59 (1.17)	18.6	1.399	1.389	1.41
Visit_Spe	2.90 (2.87)	2.05 (2.39)	16.0	0.89	0.886	0.893
Cont_Stim	452240 (22.8)	7952 (25.8)	18.0	1.181	1.151	1.212
Epidural	95944 (11.8)	2272 (11.3)	66.2	0.95	0.909	0.993
Med_Initiate	162819 (12.0)	4224 (13.7)	43.6	1.161	1.123	1.2
Sectio	258743 (12.2)	6132 (18.0)	12.6	1.574	1.531	1.619
GestAge	277.21 (13.57)	276.28 (15.58)	15.4	0.995	0.995	0.996
B_Length	51.69 (2.74)	51.78 (3.13)	11.4	1.013	1.008	1.017
B_Weight	3437.35 (594.03)	3461.49 (658.96)	10.5	1	1	1
Apgar5minOK	2066739 (98.6)	32928 (97.9)	13.4	0.698	0.647	0.753
Malformations	68303 (3.2)	2230 (6.5)	12.5	2.102	2.012	2.196
In_Asfyxi	10492 (1.0)	317 (1.2)	58.0	1.198	1.071	1.341
Sepsis	16071 (1.6)	733 (2.9)	58.0	1.829	1.696	1.971
Infections	2186 (0.2)	62 (0.2)	58.0	1.123	0.872	1.445
Jaundice	68625 (2.8)	2007 (5.6)	0.0	2.024	1.934	2.119

Table 5.2: Variable summaries for the data set **AllObs**, with 2,461,082 subjects where 1.5% has a diagnosis of ASD before the age of 18. For continuous variables the mean(SD) is reported. For binary and categorical variables the count(%) is reported. Also, the percentage of missing values and odds ratio is reported for all variables.

ADHD	No ADHD	ADHD	Missing(%)	OR	95% CI	
n	2411580	49502				
Sex-male	1225837 (50.8)	35272 (71.3)	0.0	2.398	2.351	2.445
ASD	25781 (1.1)	10268 (20.7)	0.0	24.219	23.622	24.831
AgeAtASD	10.20 (4.47)	9.81 (3.81)	98.5	0.979	0.974	0.984
M_Age	28.57 (5.09)	28.11 (5.31)	0.4	0.982	0.981	0.984
F_Age	31.51 (5.89)	31.04 (6.15)	2.4	0.986	0.985	0.988
M_ADHD_Meds	25348 (2.0)	4933 (12.7)	47.3	7.08	6.855	7.313
F_ADHD_Meds	22194 (1.8)	3213 (8.4)	47.4	5.079	4.887	5.278
M_Alc_Meds	48643 (3.8)	1553 (4.0)	46.2	1.064	1.01	1.12
F_Alc_Meds	107684 (8.1)	3299 (8.4)	44.5	1.037	1	1.075
M_Drugs_Meds	18138 (1.4)	706 (1.8)	47.3	1.289	1.195	1.39
F_Drugs_Medi	22418 (1.8)	813 (2.1)	47.1	1.202	1.12	1.29
M_Smoking	256400 (20.5)	13763 (37.0)	47.6	2.283	2.235	2.333
M_BMI	24.25 (4.93)	25.57 (5.87)	78.4	1.046	1.043	1.049
M_Spon_Abort	0.18 (0.45)	0.19 (0.46)	58.0	1.059	1.034	1.085
Parity			10.9			
First Birth	970970 (45.2)	21494 (46.5)				
Second Birth	789816 (36.8)	16360 (35.4)		0.936	0.917	0.955
Third Birth	286037 (13.3)	6094 (13.2)		0.962	0.935	0.99
Fourth Birth	71070 (3.3)	1578 (3.4)		1.003	0.952	1.056
Fifth Birth	18763 (0.9)	482 (1.0)		1.16	1.059	1.272
Sixth or over	10430 (0.5)	202 (0.4)		0.875	0.761	1.006
Ext_Preterm	3912 (0.2)	172 (0.4)	15.4	1.958	1.68	2.281
Ver_Preterm	12058 (0.6)	525 (1.1)	15.4	1.946	1.782	2.125
Mod_Preterm	174310 (8.6)	5222 (11.4)	15.4	1.374	1.335	1.415
Visit_Mid	4.40 (2.05)	5.03 (2.07)	14.3	1.15	1.145	1.154
Visit_Doc	1.80 (1.53)	2.52 (1.26)	18.6	1.358	1.349	1.366
Visit_Spe	2.91 (2.87)	2.11 (2.46)	16.0	0.898	0.894	0.901
Cont_Stim	450138 (22.8)	10054 (24.0)	18.0	1.071	1.047	1.095
Epidural	95755 (11.8)	2461 (10.0)	66.2	0.823	0.789	0.858
Med_Initiate	161724 (12.0)	5319 (13.0)	43.6	1.098	1.066	1.131
Sectio	257254 (12.2)	7621 (16.4)	12.6	1.406	1.372	1.442
GestAge	277.23 (13.55)	275.58 (15.75)	15.4	0.992	0.992	0.993
B_Length	51.69 (2.74)	51.50 (3.12)	11.4	0.976	0.973	0.979
B_Weight	3438.32 (593.59)	3409.75 (660.17)	10.5	1	1	1
Apgar5minOK	2054539 (98.6)	45128 (98.1)	13.4	0.746	0.697	0.798
Malformations	68122 (3.2)	2411 (5.2)	12.5	1.635	1.568	1.704
In_Asfyxi	10401 (1.0)	408 (1.3)	58.0	1.225	1.109	1.354
Sepsis	16010 (1.6)	794 (2.5)	58.0	1.559	1.451	1.676
Infections	2184 (0.2)	64 (0.2)	58.0	0.913	0.712	1.171
Jaundice	68069 (2.8)	2563 (5.2)	0.0	1.88	1.805	1.958

Table 5.3: Variable summaries for the data set **AllObs**, with 2,461,082 subjects where 2.0% has a diagnosis of ADHD before the age of 18. For continuous variables the mean(SD) is reported. For binary and categorical variables the count(%) is reported. Also, the percentage of missing values and odds ratio is reported for all variables.

Clusters of Predictors

In Section 2.1 we described clustering methods, we will now apply k-means and hierarchical clustering to the predictors in **AllObs**. In both approaches observations including missing values were omitted. The two approaches showed similar results, the elbow plots both look like the one for K-means clustering depicted in Figure 5.3. The number of clusters could not be decided based on the elbow plots. We tried clustering by k-means into 2-20 clusters, the best way to describe our findings would be to refer to the other approach, hierarchical clustering, as we got similar results but the dendrogram from hierarchical clustering is easier to present.

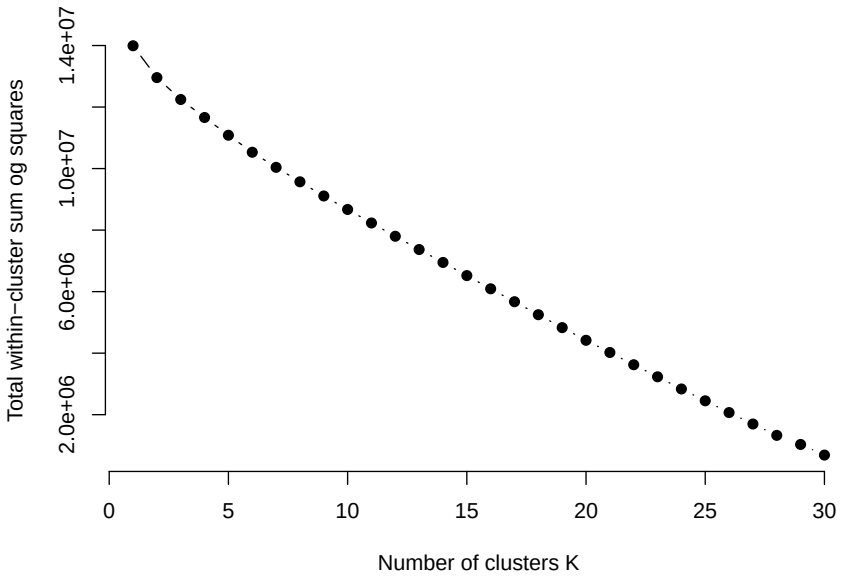


Figure 5.3: Elbow-plot of K-means clustering of 32 predictors for the data set **AllObs** with 2,461,082 observations (428,243 observations used).

For both methods, `B_Weight` and `B_length` were the closest related variables, easily seen in the dendrogram from hierarchical clustering in Figure 5.4, but also clear from k-means as they would always appear in the same cluster. It is clear directly in the dendrogram, what we saw in the elbow-plots, that a preferred number of clusters is not easily determined. Most clusters are non-surprising, like `F_Age` and `M_Age` clustering. But also the parity variables do cluster, which is expected, as they are constructed to be mutually exclusive. An interesting cluster could be `Epidural`, `Cont_Stim` and `Med_Initiate`, as one could speculate that these three all would be part of a long and hard labor.

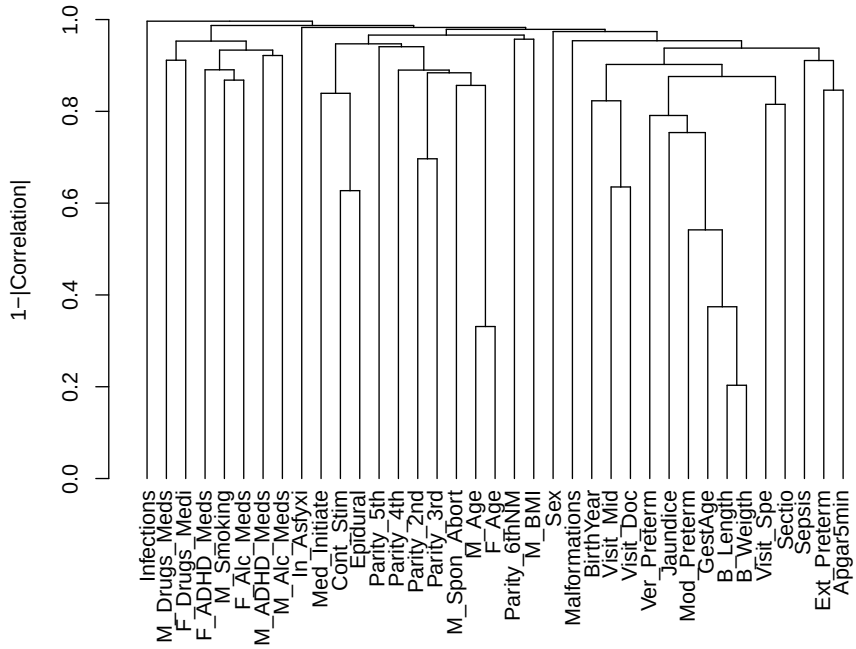


Figure 5.4: Dendrogram of hierarchical clustering of 32 predictors for the data set **AllObs** with 2,461,082 observations (428,243 observations used).

As we did not find a definitely preferred number of clusters and 32 predictors is not a particularly large amount of predictors, we choose not to go forward with any dimension reduction methods and continue to the prediction of ASD and ADHD.

5.2 Analysis

In Section 5.1.2 we defined the data set **AllObs**. In this section we analyze this data set using two classifiers: logistic regression and classification trees and at the end of this section, we summarize which models are the best for ASD and ADHD, respectively. In order to fit logistic regression models, we need a data set without missing values and we therefore describe in the next section, how we avoid this.

Imputation

As described in Section 5.1.2, we have a considerable amount of missing values even in our data set **AllObs**. How to best handle this missingness is outside the scope of this master thesis, but we cannot completely ignore it. We have thus chosen to impute missing categorical values with an extra level indicating missing and the continuous variables are imputed as the mean of the non-missing values. This imputation enables us to conduct analysis on $2,461,082$ subjects utilizing 57 variables, as the categorical predictors are converted to binary predictors.
We further expand the number of variables to 81 by replacing six of our predictors with natural splines. Our incidence and prevalence plots, Figure 5.1 and Figure 5.2, indicate that the expected risk of getting a diagnose is not linear over time, thus we spline the

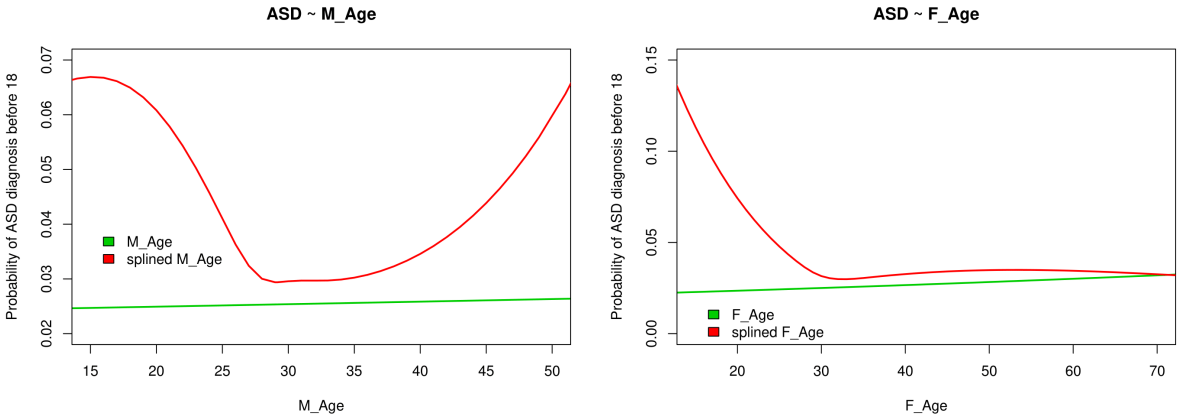
predictor `BirthYear`. The predictors `M_Age` and `F_Age` are splined because we expect the risk of diagnosis to be high for both young and old parents, but lower for median aged parents. The predictors `M_BMI`, `B_Weight` and `GestAge` are all expected to result in severely higher risk at low values than normal and high. We now define what we mean by an imputed data set with splined variables.

> **AllObsImputed:** This data set is made by performing imputation on the data set **AllObs** and furthermore spline 6 predictors, as elaborated above. The data set **AllObsImputed** includes $2,461,082$ subjects and 81 predictors, where 1.5% of the subjects have an ASD diagnose and 2.0% have an ADHD diagnose.

We now investigate whether it was wise to spline the six continuous predictors. Note that the continuous predictors are imputed before they are splined.

Investigation of the Splined Predictors

We set up a small example; we build 24 simple logistic regressions models, 12 models with ASD as response and 12 with ADHD as response. Out of the 12 models, we estimate the probability of ASD for 6 of these models using one of the continuous predictors `M_Age`, `F_Age`, `B_Weight`, `GestAge`, `BirthYear` and `M_BMI`, respectively. Furthermore we fitted 6 logistic regressions to the splined transformations of said six predictors, and again estimated the probabilities of getting a diagnosis of ASD. Similarly is done for the remaining 12 models which have ADHD as response.



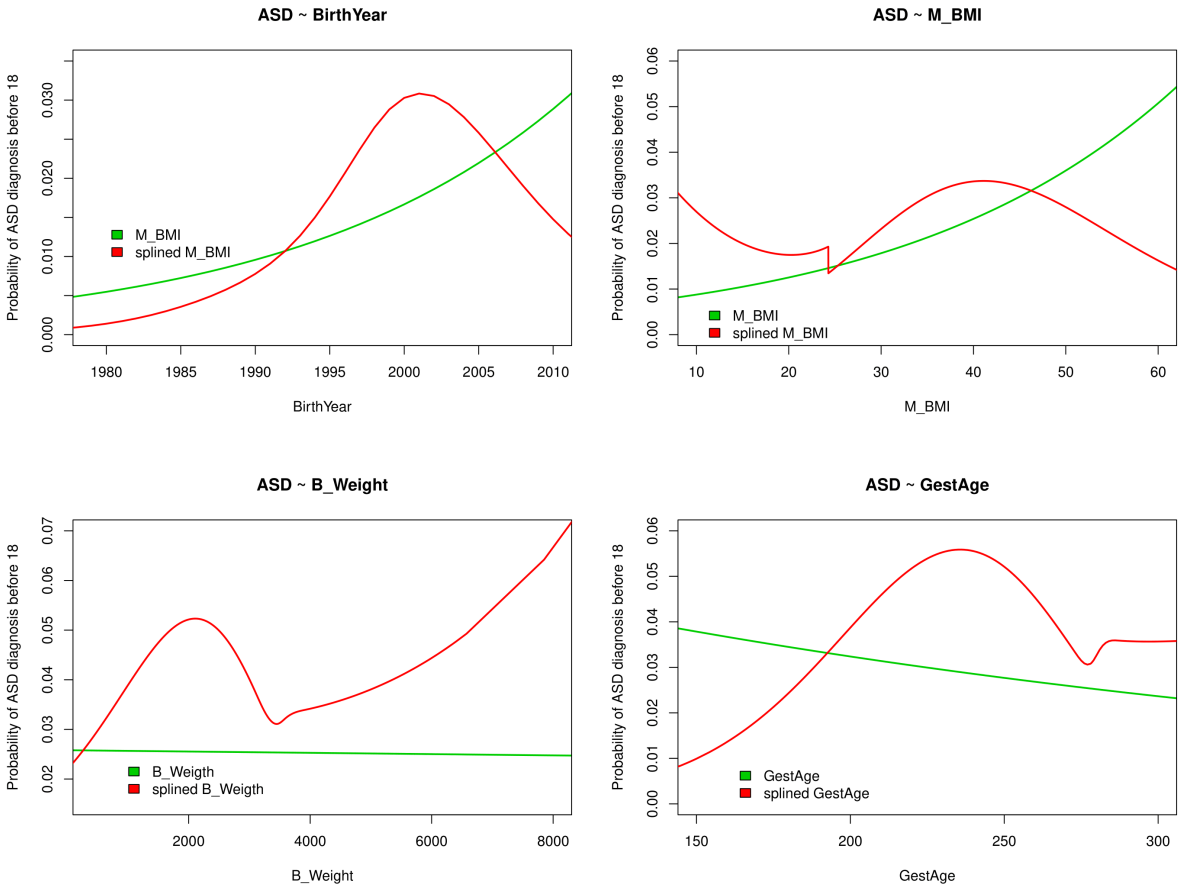


Figure 5.7: Plots of the probability of getting diagnosed with ASD based on logistic regression of only a continuous predictor or the same predictor as a natural spline. The data set used is **AllObsImputed** with 2,461,082 subjects, where 1.5% have an ASD diagnose.

In Figure 5.7 we see that modeling the probability of ASD with splined continuous variables reveals what we suspected in Section 5.2, that e.g. maternal age increases the risk of ASD both for low ages and high ones. Interestingly, only low paternal age deviates from the unsplined model. We also see that if the mother has extremely low BMI or relatively high BMI then this increases the risk of the subject having ASD. One of the most interesting findings in these splined variables are that splined birth year decreases for the late years and therefore captures that many of the late-born children do not have full follow-up and therefore do not get any diagnose even though they may get it later in their youth. It is also interesting that higher birth weight increases the probability of diagnosis, we only expected this to be the case for low birth weight. The low probabilities for the extremely low birth weight and gestational age might be due to the difficulty of extremely early born babies to live long enough to even get a diagnosis of ASD. Furthermore, note the sharp change points for especially the splined predictors `M_BMI`, `B_Weight` and `GestAge`. This occurs because the mean value of the original predictors is imputed for missing values before the predictors are splined, assigning many subjects the mean values, thus placing the spline knots (that are placed at the 20tn, 40tn, 60tn and 80tn percentiles) closer to the mean.

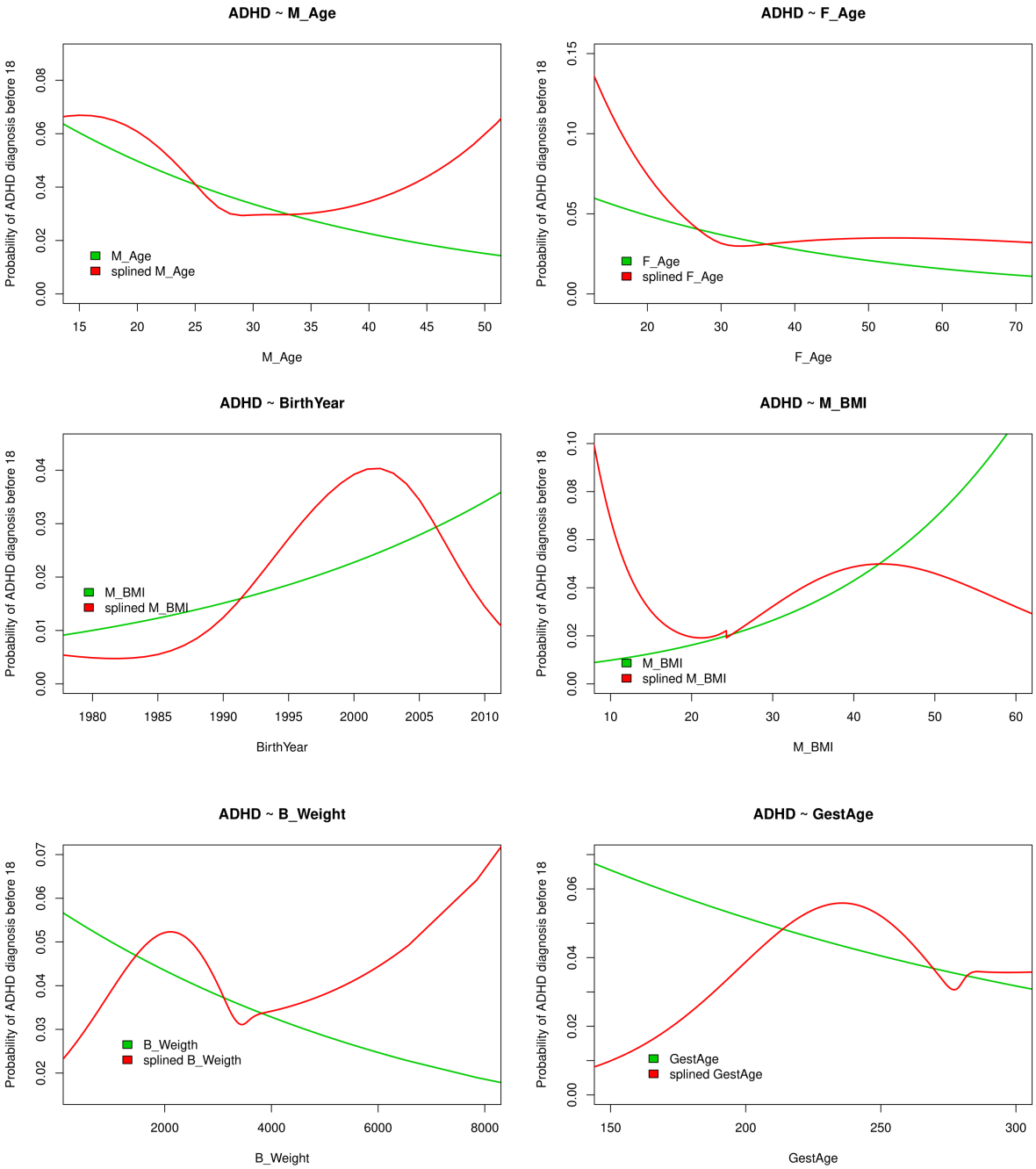


Figure 5.10: Plots of the probability of getting diagnosed with ADHD based on logistic regression of only a continuous predictor or the same predictor as a natural spline. The data set used is **AllObsImputed** with 2,461,082 subjects, where 2.0% have an ADHD diagnose.

In Figure 5.10 we see for ADHD similar results as we did for ASD. But we also see to an even bigger extent the necessity to spline these predictors. In Figure 5.7 we saw that for ASD it was convenient to spline the predictors such that the probability could deviate from the flatter nature of the unsplined predictors. But in Figure 5.10 we see the unsplined predictors actually aiming at lower probabilities for higher values of maternal age and birth weight, whereas the splined predictors again predict higher probabilities for higher values. These plots justifies the use of splines for our data.

We now fit five logistic regression models to the data set **AllObsImputed** including splined predictors.

Specification of Logistic Regression Models

We now define five different logistic regression models that will be used several times throughout the analysis, but for different subjects.

- **LASSO:** A LASSO regression model including all 81 predictors, where λ is chosen based on 10-fold cross-validated AUC.
- **Ridge:** A Ridge regression model including all 81 predictors, where λ is chosen based on 10-fold cross-validated AUC.
- **GLM Full:** A logistic regression model including all 81 predictors.
- **GLM 10:** A logistic regression model including 10 selected predictors, 5 of them replaced by splines. The predictors included were: `Sex`, `M_Smoking`, `GestAge`, `B_Weight`, `Jaundice`, `BirthYear`, `Sectio`, `Malformations`, `M_Age` and `F_Age`.
- **GLM 5:** A logistic regression model including 5 selected predictors, 2 of them replaced by splines. The predictors included were: `Sex`, `M_Smoking`, `GestAge`, `B_Weight` and `Jaundice`.

The predictors selected for **GLM 10** and **GLM 5** were chosen by the authors, we aimed to get both binary and continuous splined predictors in both models. We also included predictors that clustered in the EDA like `GestAge` and `B_Weight` as well as predictors expected to be very different like `Sex` and `M_Smoking`. The two models merely represent smaller models, and the predictors included in them could most likely have been chosen more wisely.

How We Evaluate our Models

We compare the models using the measures presented in Section 4.1, which are 10-fold cross-validated, such that e.g. the PPV is the mean of the 10 PPV's from the ten folds. The 10-fold cross-validated AUC can be computed in two ways, it can be the mean of the 10 AUC's, which is what we report in the tables of performance measures. The other way to calculate the AUC is to use the probabilities produced by each fold of the cross-validation and calculate an overall AUC, this is what we report in the figures in Appendix C. Both calculations produce almost the same AUC estimate, which is due to the large number of observations we have available.

5.2.1 The First Analysis Using Logistic Regression

We fit the five logistic regression models previously presented for the data set **AllObsImputed**, which results in the evaluation measures presented in Table 5.4 and Table 5.5. Table 5.4 is with ASD as response and Table 5.5 is with ADHD as response. In general, we want to compare the logistic regression models based on their AUC scores, but as the other evaluation measures can provide further insight about the models' qualities, these are also reported. Furthermore, for this first analysis, ROC curves are presented in Figure 5.11 for the 5 models of both ASD and ADHD, but since these curves do not

explain more than the tables, besides having an 95% confidence interval shown, we have chosen to move all other ROC curves to Appendix C. Each table caption refers to the corresponding ROC curve plot in the appendix. In Table 5.4 and Table 5.5 we see that it is generally equally difficult to predict ASD as ADHD with these five models as the AUC scores are generally very close to each other. However, the F-score measure is generally better for ASD, indicating that there is a better balance between sensitivity and PPV, even though the PPV will always be low due to the low prevalence of ASD and ADHD. Note that the model with the best AUC score for both ASD and ADHD is the logistic regression model with all predictors `GLM Full`. This indicates that we have such a large amount of subjects, that we cannot overfit the models to the 81 predictors. However, the `GLM Full` models are very close to the `LASSO` and `Ridge` model's AUC scores and therefore the models are almost equally good at predicting the response.

ASD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.030	0.994	0.759	0.639	4.746	3.038	4.746	0.764
Ridge	0.030	0.994	0.760	0.633	4.750	3.040	4.750	0.762
GLM Full	0.030	0.995	0.767	0.632	4.795	3.069	4.795	0.765
GLM 10	0.029	0.995	0.782	0.605	4.885	3.127	4.885	0.754
GLM 5	0.030	0.992	0.616	0.703	3.851	2.464	3.851	0.701

Table 5.4: 10-fold cross-validated evaluation measures for five different models predicting ASD in the data set **AllObsImputed** with 2,461,082 subjects, where 1.5% have an ASD diagnose. For ROC curves see Figure 5.11.

ADHD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.046	0.992	0.711	0.693	4.445	2.845	4.445	0.773
Ridge	0.046	0.991	0.702	0.701	4.384	2.806	4.384	0.773
GLM Full	0.045	0.992	0.724	0.682	4.528	2.898	4.528	0.775
GLM 10	0.041	0.991	0.717	0.656	4.482	2.868	4.482	0.749
GLM 5	0.038	0.989	0.647	0.659	4.043	2.588	4.043	0.695

Table 5.5: 10-fold cross-validated evaluation measures for five different models predicting ADHD in the data set **AllObsImputed** with 2,461,082 subjects, where 2.0% have an ADHD diagnose. For ROC curves see Figure 5.11.

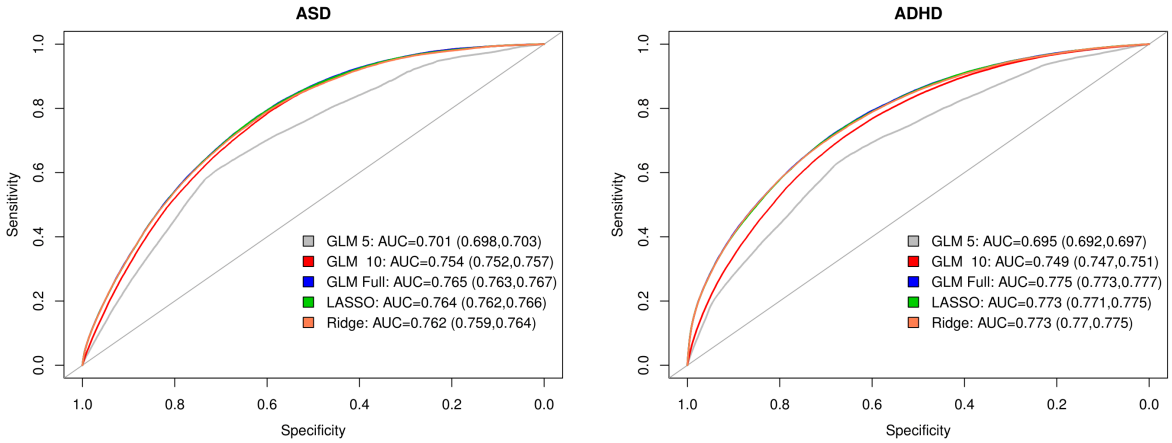


Figure 5.11: 10-fold cross-validated ROC curves for five different models predicting ASD and ADHD, respectively, in the data set **AllObsImputed** with 2,461,082 subjects, where 1.5% have an ASD diagnose and 2.0% have an ADHD diagnose.

Logistic Regression Applied to Fewer Subjects

As it is computional heavy to compute the models for the data set **AllObsImputed** (about 5 days computation time), we chose to run all models from now on only for $\frac{1}{200}$ part of any data set, randomly chosen, before using it for all subjects. Furthermore, we did it in order to investigate whether more data would always produce better predictive models as $\frac{1}{200}$ is still a considerable amount of subjects, that is $\frac{2{,}461{,}082}{200} \approxeq 12,305$ subjects. Note that this random sampling does not necessarily keep the ratio between diagnosed and non-diagnosed. Table 5.6 and Table 5.7 show evaluation measures for the five logistic regression models on such a smaller data set. This is generally the result we see in all our analysis with less data, that these models predict poorer than models made on the full data set **AllObsImputed**. However, of course we also experience some $\frac{1}{200}$ parts, which happens to be really good splits. We have also tried to run the analyses with $\frac{1}{20}$ of the data set **AllObsImputed** and draw the same conclusion. More data gives better AUC scores and thus better predictive models. Table 5.6 and Table 5.7 are analyses on the same $\frac{1}{200}$ and here the tables indicate that it is easier to predict ADHD with `LASSO` than to predict ASD with any of these logistic regression models. The random cut is thus of great importance and we should clearly use the full data set **AllObsImputed** in our analysis instead of a fraction of it.

ASD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.030	0.997	0.853	0.598	5.329	3.410	5.329	0.742
Ridge	0.030	0.997	0.852	0.612	5.323	3.407	5.323	0.749
GLM Full	0.034	0.996	0.788	0.665	4.923	3.151	4.923	0.730
GLM 10	0.030	0.997	0.840	0.615	5.250	3.360	5.250	0.742
GLM 5	0.032	0.995	0.787	0.636	4.918	3.148	4.918	0.712

Table 5.6: 10-fold cross-validated evaluation measures for five different models predicting ASD in $\frac{1}{200}$ of the data set **AllObsImputed**. For this cut we have 12,305 subjects, where 1.4% have an ASD diagnose. For ROC curves see Figure C.1.

ADHD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.054	0.993	0.756	0.688	4.726	3.025	4.726	0.760
Ridge	0.068	0.991	0.647	0.773	4.046	2.589	4.046	0.741
GLM Full	0.052	0.992	0.746	0.658	4.665	2.986	4.665	0.725
GLM 10	0.046	0.991	0.698	0.665	4.361	2.791	4.361	0.702
GLM 5	0.040	0.991	0.749	0.568	4.678	2.994	4.678	0.661

Table 5.7: 10-fold cross-validated evaluation measures for five different models predicting ADHD in $\frac{1}{200}$ of the data set **AllObsImputed**. For this cut we have 12,305 subjects, where 2.1% have an ADHD diagnose. For ROC curves see Figure C.1.

5.2.2 Observations with No Missing Values

In this section we use only a part of the data set **AllObs** to avoid imputing.

> **ObsNoMissing:** This data set is made by extracting all subjects that have no missing values for the data set **AllObs** and imputation is therefore not necessary. The data set **ObsNoMissing** has 428,943 subjects, where 2.0% have an ASD diagnose and 2.4% have an ADHD diagnose.

We avoid imputing, but on the other hand, we do not have as many observations, which we previously showed give worse predictive models. Since the variable with the latest start of coverage, `M_BMI`, starts in 2003, all observations from 2002 and before have missing values and these subjects are thus not included in this data set. As we only have observations from 2003 onward, we do not have full follow-up on these subjects, as they do not reach the age of 18, because we only have data until 2017. Table 5.8 and Table 5.9 show the evaluation measures for the five logistic regression models using the data set **ObsNoMissing**. Note that these AUC values are not higher than the AUC values found for the data set **AllObsImputed**, that is Table 5.4 and Table 5.5. We therefore do not get better predictive models by settling for observations that do not have missing values. In the next section, we try to make better predictive models by using classification trees.

ASD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.033	0.990	0.730	0.569	4.564	2.921	4.564	0.697
Ridge	0.033	0.991	0.734	0.566	4.587	2.936	4.587	0.698
GLM Full	0.035	0.990	0.704	0.599	4.398	2.815	4.398	0.700
GLM 10	0.030	0.990	0.754	0.511	4.713	3.017	4.713	0.661
GLM 5	0.030	0.991	0.765	0.499	4.779	3.058	4.779	0.655

Table 5.8: 10-fold cross-validated evaluation measures for five different models predicting ASD in the data set **ObsNoMissing** with 428,943 subjects, where 2.0% have an ASD diagnose. For ROC curves see Figure C.2.

ADHD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.052	0.989	0.690	0.690	4.311	2.759	4.311	0.754
Ridge	0.049	0.990	0.714	0.662	4.460	2.855	4.460	0.755
GLM Full	0.050	0.990	0.712	0.672	4.448	2.847	4.448	0.757
GLM 10	0.041	0.987	0.669	0.616	4.181	2.676	4.181	0.694
GLM 5	0.036	0.988	0.757	0.499	4.731	3.028	4.731	0.673

Table 5.9: 10-fold cross-validated evaluation measures for five different models predicting ADHD in the data set **ObsNoMissing** with 428,943 subjects, where 2.4% have an ADHD diagnose. For ROC curves see Figure C.2.

5.2.3 Analysis Using Classification Trees

This section includes models made using classification trees on the basis of the data set **AllObsImputed**. As we shall see later in this section, the trees helped us to identify a major problem of our logistic regression models, and that problem is calendar year. Table 5.10 and Table 5.11 show the evaluation measures for full trees grown to several different depths for ASD and ADHD respectively.

ASD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
Depth 2	0.032	0.992	0.624	0.721	3.900	2.496	3.900	0.714
Depth 8	0.028	0.995	0.783	0.602	4.893	3.131	4.893	0.757
Depth 14	0.029	0.994	0.745	0.625	4.653	2.978	4.653	0.738
Depth 20	0.027	0.992	0.653	0.653	4.081	2.612	4.081	0.678

Table 5.10: 10-fold cross-validated evaluation measures for classification trees predicting ASD, grown to different depths. The data set used is **AllObsImputed** with 2,461,082 subjects, where 1.5% have an ASD diagnose.

ADHD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
Depth 2	0.106	0.986	0.358	0.818	2.237	1.432	2.237	0.595
Depth 8	0.043	0.991	0.710	0.678	4.435	2.839	4.435	0.761
Depth 14	0.043	0.991	0.688	0.683	4.297	2.750	4.297	0.742
Depth 20	0.042	0.988	0.572	0.733	3.575	2.288	3.575	0.672

Table 5.11: 10-fold cross-validated evaluation measures for classification trees predicting ADHD, grown to different depths. The data set used is **AllObsImputed** with 2,461,082 subjects, where 2.0% have an ADHD diagnose.

Note that the trees with depth 20 have poorer AUC scores than the models with depth 8. It makes sense, as a depth of 20 corresponds to $2^{20} = 1,048,576$ leaves and since we have $2,461,082$ subjects in the imputed data set **AllObsImputed**, many observations from the training data have their own leaf, thus overfitting the model. The tree models with depth 8 have about as good AUC values as the best logistic regression models from Table 5.4 and Table 5.5. We have chosen the depths $2, 8, 14$ and 20 arbitrarily, but later in Section 5.2.5, we actually do use the optimal depth for the specific situation. Figure 5.12 and Figure 5.13 on the other hand do not show a full tree but pruned trees grown to depth 8. We show the pruned tree since a full tree, grown to depth 8, is too large to be visually displayed in a meaningful way in this thesis.

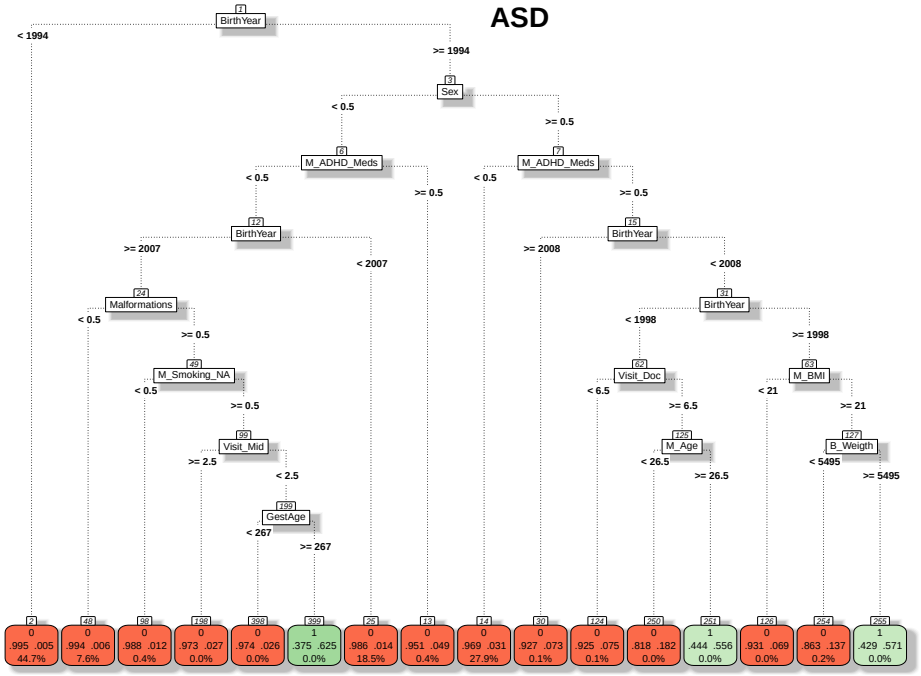


Figure 5.12: A pruned tree predicting ASD after being grown to depth 8 on the data set **AllObsImputed** with 2,461,082 subjects, where 1.5% have an ASD diagnose.

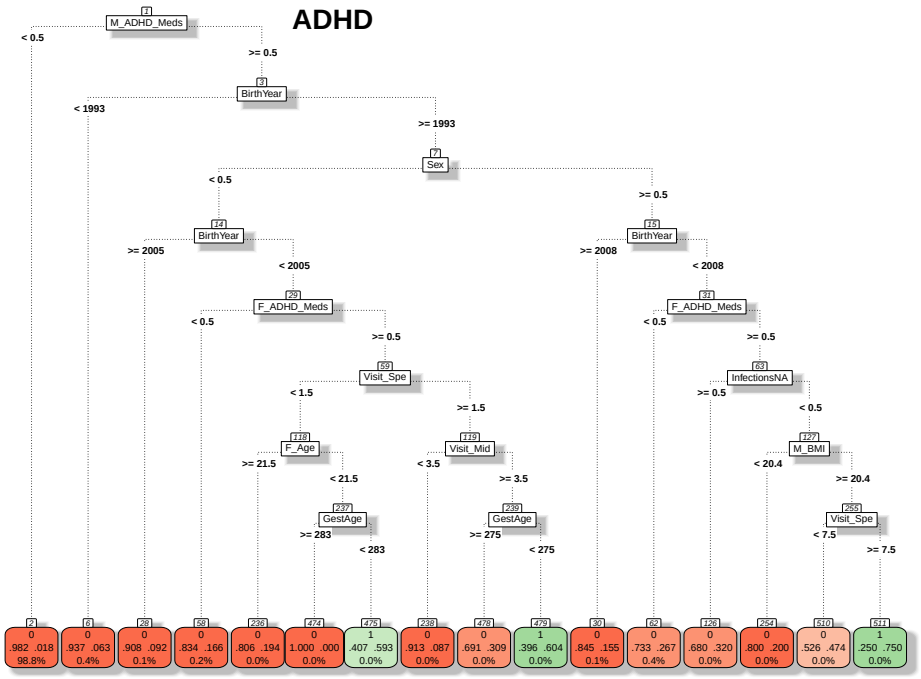


Figure 5.13: A pruned tree predicting ADHD after being grown to depth 8 on the data set **AllObsImputed** with 2,461,082 subjects, where 2.0% have an ADHD diagnose.

Note in Figure 5.12 and Figure 5.13, that one of the most important predictors is `BirthYear` since it appears relatively high in the trees and in several places for both the ASD and ADHD models. The predictor `BirthYear` is important for our models because both the number of ASD and ADHD dignoses increase over the years, which can be seen in our prevalence and incidence figures, Figure 5.1 and Figure 5.2. However, it does not make sense to include this predictor in a predictive model to predict ASD and ADHD in the future, as e.g. all boys born in or after 2008 and girls born in or after 2005 will never get a diagnosis of ADHD if we follow the flowchart in Figure 5.13

Thus, we choose to exclude `BirthYear` and fit trees on the remaining predictors. We chose not to show all trees and thus we only describe, what we saw in these plots. The plots that appear indicate that predictors such as `M_BMI` and `Epidural` are important predictors, as they appear relatively high in the trees. We suspect that these variables are chosen because they have a late covering time and therefore people before this coverage time have missings for these variables, making missingness describe time, this is thus again a calendar year problem. We therefore fit tree models based on our data set **ObsNoMissing** without the variable `BirthYear`. Table 5.12 and Table 5.13 show evaluation measures for three trees fitted on three different data sets respectively. In all trees we have chosen a depth of 8 as it seems reasonable from the first tree models. The first tree is the model based on the data set **AllObsImputed**, the second model is based on $\frac{1}{200}$ randomly selected observations of the data set **AllObsImputed** and for the last model we use the data set **ObsNoMissing** where the predictor `BirthYear` is omitted.

ASD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
AllObsImputed	0.028	0.995	0.783	0.602	4.893	3.13	1 4.893	0.757
1/200 of **AllObsImputed**	0.028	0.994	0.727	0.596	4.545	2.909	4.545	0.677
ObsNoMissing	0.031	0.990	0.744	0.525	4.649	2.976	4.649	0.669

Table 5.12: 10-fold cross-validated evaluation measures of trees predicting ASD, grown to depth 8 on three different datasets. The data set **AllObsImputed** includes 2,461,082 subjects, where 1.5% subjects are diagnosed with ASD. The data set $\frac{1}{200}$ of **AllObsImputed** includes 12,305 subjects, where 1.4% subjects are diagnosed with ASD. The data set **ObsNoMissing** includes 428,943 subjects, where 2.0% subjects are diagnosed with ASD. For ROC curves see Figure C.3.

ADHD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
AllObsImputed	0.043	0.991	0.710	0.678	4.435	2.839	4.435	0.761
1/200 of **AllObsImputed**	0.050	0.975	0.591	0.670	3.692	2.363	3.692	0.645
ObsNoMissing	0.050	0.987	0.610	0.712	3.816	2.442	3.816	0.720

Table 5.13: 10-fold cross-validated evaluation measures of trees predicting ADHD, grown to depth 8 on three different datasets. The data set **AllObsImputed** includes 2,461,082 subjects, where 2.0% subjects are diagnosed with ADHD. The data set $\frac{1}{200}$ of **AllObsImputed** includes 12,305 subjects, where 2.1% subjects are diagnosed with ADHD. The data set **ObsNoMissing** includes 428,943 subjects, where 2.4% subjects are diagnosed with ADHD. For ROC curves see Figure C.3.

Table 5.12 and Table 5.13 show that the AUC scores are best for the full imputed

data set **AllObsImputed**, but it is clear that a good classifier for the early years in our analysis is that no person has ASD and ADHD, again due to the incidence and prevalence. Thus, we will continue to study models where we try to take calendar year into account. Tree models based on the data set **ObsNoMissing** without the predictor `BirthYear` can be seen in Figure 5.14 and Figure 5.15, where the pruned trees are shown. These models indicate that predictors such as `Sex`, `M_ADHD_Meds` and `M_BMI` are of great importance, again because they appear high in the trees. The first two predictors `Sex` and `M_ADHD_Meds` are not surprising, but this is not the case for the predictor `M_BMI`, which our consulting psychiatrist Marlene Briciet Lauritsen[1] finds as having not previously been associated with ASD. In the next section we fit logistic regression models based on the fact that calendar year is a problem.

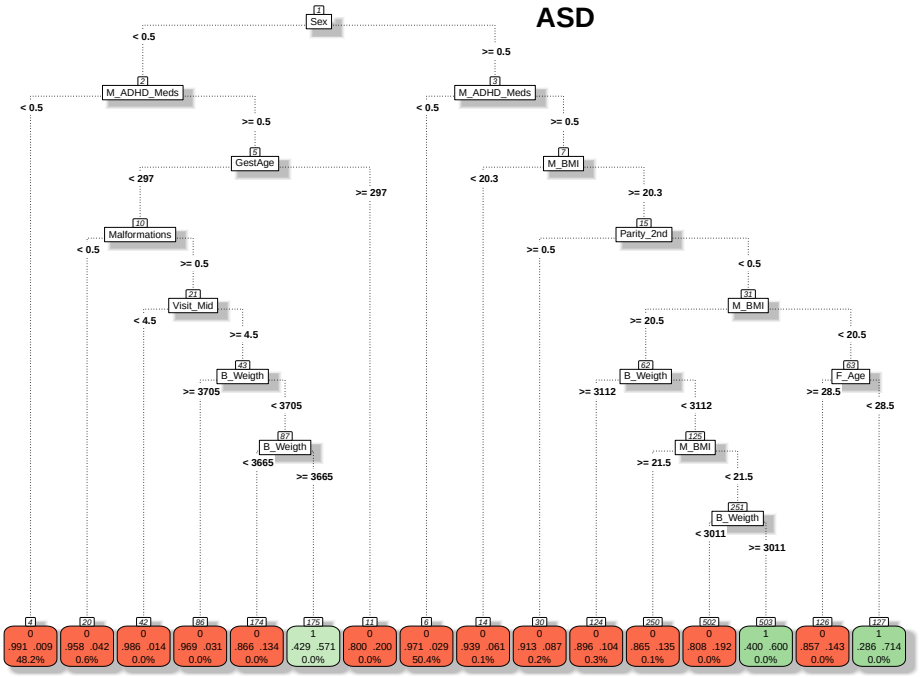


Figure 5.14: A pruned tree predicting ASD after being grown to depth 8 on the data set **ObsNoMissing** with 428,943 subjects, where 2.0% have an ASD diagnose.

[1]Child and Adolescent Psychiatry, Region of Northern Jutland Psychiatry, Aalborg, Denmark

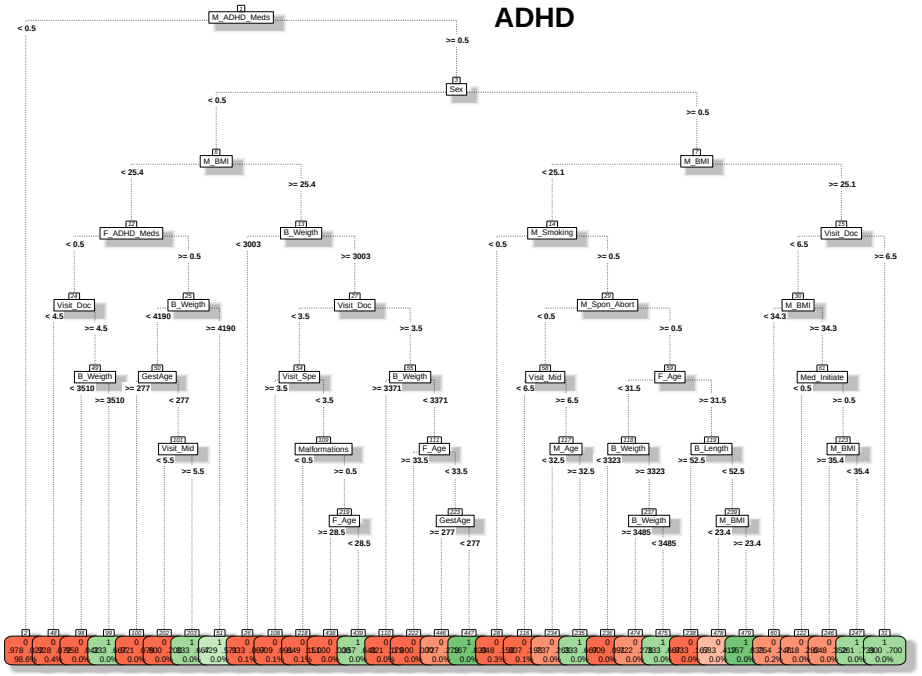


Figure 5.15: A pruned tree predicting ADHD after being grown to depth 8 on the data set **ObsNoMissing** with 428,943 subjects, where 2.4% have an ADHD diagnose.

5.2.4 Models That Take Calendar Year into Account

In order to investigate how important calendar year is for the logistic regression models, we first make three of our five logistic regression models by including only 1 predictor, `BirthYear`, for the data set **AllObsImputed**. These logistic regression models are shown in Table 5.14 and Table 5.15.

ASD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.024	0.994	0.793	0.518	4.953	3.170	4.953	0.703
Ridge	0.024	0.994	0.793	0.518	4.953	3.170	4.953	0.701
GLM	0.024	0.994	0.793	0.518	4.953	3.170	4.953	0.704

Table 5.14: 10-fold cross-validated evaluation measures for three different models predicting ASD with only splined `BirthYear` as a predictor. The data set used is **AllObsImputed** with 2,461,082 subjects, where 1.5% have an ASD diagnose. For 10-fold cross-validated ROC curves see Figure C.4.

ADHD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.032	0.991	0.786	0.512	4.912	3.143	4.912	0.683
Ridge	0.032	0.991	0.786	0.512	4.912	3.143	4.912	0.683
GLM	0.032	0.991	0.786	0.512	4.912	3.143	4.912	0.684

Table 5.15: 10-fold cross-validated evaluation measures for three different models predicting ADHD with only splined `BirthYear` as a predictor. The data set used is **AllObsImputed** with 2,461,082 subjects, where 2.0% have an ADHD diagnose. For 10-fold cross-validated ROC curves see Figure C.4.

We only fit three models because `GLM Full`, `GLM 10` and `GLM 5` are the same in this case. These models with only this one predictor predict a little worse than the models based on all our predictors shown in Table 5.4 and Table 5.5, but are not particularly worse. It is therefore clear that calendar year is a problem in our models. Based on this, we chose to fit logistic regression models with data set **ObsNoMissing** without the predictor `BirthYear`. The models that emerge from this fit have evaluation measures shown in Table 5.16 and Table 5.17.

ASD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.032	0.990	0.721	0.561	4.508	2.885	4.508	0.682
Ridge	0.032	0.990	0.727	0.556	4.543	2.908	4.543	0.683
GLM Full	0.032	0.991	0.746	0.539	4.666	2.986	4.666	0.685
GLM 10	0.031	0.990	0.745	0.520	4.658	2.981	4.658	0.661
GLM 5	0.030	0.991	0.765	0.498	4.780	3.059	4.780	0.655

Table 5.16: 10-fold cross-validated evaluation measures for five different models predicting ASD in the data set **ObsNoMissing** without `BirthYear` included as predictor. The data set consists of 428,943 subjects, where 2.0% have an ASD diagnose. For ROC curves see Figure C.5.

ADHD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.051	0.988	0.631	0.711	3.947	2.526	3.947	0.733
Ridge	0.049	0.988	0.656	0.686	4.099	2.623	4.099	0.734
GLM Full	0.051	0.988	0.647	0.698	4.043	2.587	4.043	0.736
GLM 10	0.041	0.987	0.664	0.619	4.152	2.657	4.152	0.694
GLM 5	0.036	0.989	0.770	0.487	4.812	3.080	4.812	0.673

Table 5.17: 10-fold cross-validated evaluation measures for five different models predicting ADHD in the data set **ObsNoMissing** without `BirthYear` included as predictor. The data set consists of 428,943 subjects, where 2.4% have an ASD diagnose. For ROC curves see Figure C.5.

We see that the models naturally predict worse than the full imputed data set shown in Table 5.4 and Table 5.5. On the other hand, these models make much more sense to apply for new data in 2019. Note, however, that these AUC values are about $0.68 - 0.73$, thus these models are not particularly good at predicting.

5.2.5 Full Follow-Up

To circumvent the problem of administrative censoring (uncomplete follow-up) while having information from as many variables as possible, we choose to define a data set containing subjects born between 1997 and 1999. Since there are not many observations for the years 1997-1999 without a single missing value, we choose to make an imputed data set over these years in order to get a reasonable data set with full follow-up (though subjects can still be lost to follow-up due to death or emigration).

FullFollowUp9799Imputed: This data set is made based on the data set **AllObs**, where we have excluded the predictors `Cont_Stim`, `Epidural` and `M_BMI`, as these variables are not available until a later date. We also exclude

`BirthYear` to avoid predicting by calendar year. Furthermore, we have extracted all subjects having `BirthYear` in $1997-1999$ and imputed the missing values. The data set **FullFollowUp9799Imputed** contains $220,458$ subjects, where 2.5% have an ASD diagnose and 3.5% have an ADHD diagnose.

We thereby have full follow-up on three years at a small expense of excluding three predictors besides `BirthYear`. Table 5.18 and Table 5.19 show evaluation measures for models fitted on the data set **FullFollowUp9799Imputed**. Note that we get worse AUC scores than we get for the data set **ObsNoMissing** without the predictor `BirthYear` with evaluation measures shown in Table 5.16 and Table 5.17. We think this is due to us only including subjects for three years making the data set almost half size compared to the data set **ObsNoMissing**. Furthermore, as the prevalence has risen and more actual diagnoses are added both to the predicted group and the un-predicted group, the PPV rises with the more true positives, but the sensitivity lowers with the false negatives increasing.

ASD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.039	0.986	0.683	0.562	4.266	2.730	4.266	0.653
Ridge	0.039	0.985	0.677	0.567	4.233	2.709	4.233	0.653
GLM Full	0.039	0.986	0.691	0.556	4.319	2.764	4.319	0.656
GLM 10	0.038	0.985	0.692	0.538	4.327	2.770	4.327	0.636
GLM 5	0.037	0.985	0.676	0.542	4.227	2.705	4.227	0.628
Tree Depth 7	0.037	0.986	0.718	0.513	4.485	2.871	4.485	0.642

Table 5.18: 10-fold cross-validated evaluation measures for six different models predicting ASD in the data set **FullFollowUp9799Imputed** with 220,458 observations (ASD 2.5%). Tree depth chosen between 1 and 20. For ROC curves see Figure C.6.

ADHD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.072	0.978	0.546	0.737	3.410	2.182	3.410	0.693
Ridge	0.071	0.979	0.572	0.719	3.572	2.286	3.572	0.694
GLM Full	0.069	0.979	0.575	0.712	3.595	2.301	3.595	0.696
GLM 10	0.055	0.978	0.630	0.596	3.936	2.519	3.936	0.651
GLM 5	0.051	0.978	0.674	0.531	4.211	2.695	4.211	0.628
Tree Depth 7	0.074	0.976	0.482	0.777	3.013	1.928	3.013	0.678

Table 5.19: 10-fold cross-validated evaluation measures for six different models predicting ADHD in the data set **FullFollowUp9799Imputed** 220,458 observations (ADHD 3.5%). Tree depth chosen between 1 and 20. For ROC curves see Figure C.6.

We have chosen the depth 7 for the tree classifier because it is the depth giving the highest AUC value. For the classification trees we clearly saw that we can easily overfit a model by making it too complex and therefore we choose to elaborate this it in the following example.

Overfitting in Practice

Theoretically, we dealt with the bias-variance trade-off throughout the thesis, but when we got to the actual analysis, we were not able to overfit the logistic regression models

to the data. This was possible with the classification trees, as we saw in Table 5.10 and Table 5.11, the models got worse as the trees were grown to greater depths. This should be due to overfitting and we now try to illustrate this in an example. For the data set **FullFollowUp9799Imputed** we grew trees to depths 1:20, estimating the generalizability using 10-fold cross-validated AUC as usually by predicting the test observations in the last tenth. We also predicted the $\frac{9}{10}$'s in the training sets on which each of the ten models were trained, enabling us to get a 10-fold cross-validated AUC measure of how well the tree fit the data which it was trained on, by the mean of ten internal AUCs.

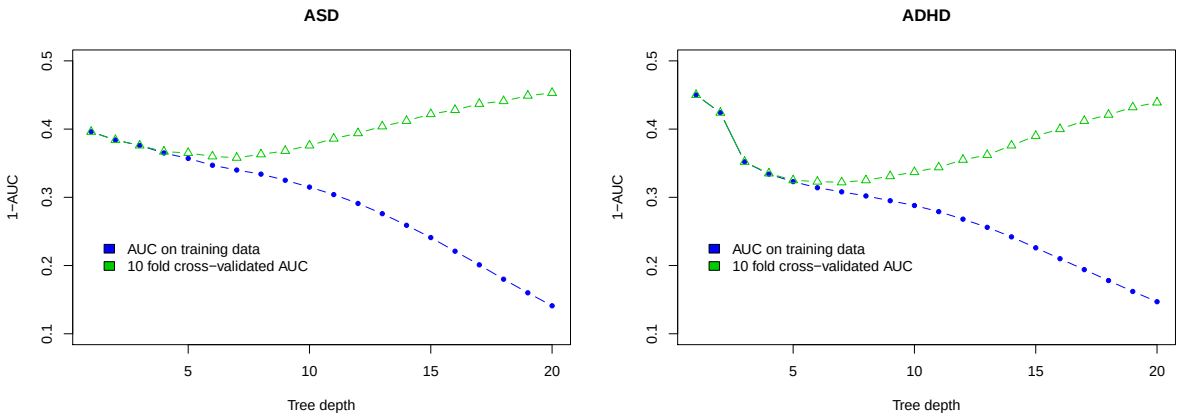


Figure 5.16: Threes grown to several depths predicting ASD and ADHD, respectively. 10-fold cross validated AUCs of trees predicting new observations (green). AUCs of trees predicting the data they were trained on (blue).

In Figure 5.16 we see that when fitting and testing on the same data set, the AUC value becomes better the deeper the tree is grown. Furthermore, we see how more complex models lead to overfitting, as the internal AUC gets better (blue line), but the more complex models lead to worse predictions (green line). For the models fitted with training data and tested by a hold-out test data, we see that the best AUC value is reached at a depth of 7. Our tree models with depth 1 are very simple, as there is only one split and therefore there is low variance, but high bias. Conversely, when we are fitting complicating trees with depth 20, the models contain high variance, but low bias. This is an example of the bias-variance trade-off in practice.

This was a little sidetrack about overfitting and we now turn our attention back to the search for better prediction models. Another way to get better predictive models is to include more or other predictors that are known to be associated with the response, but we are unable to do it in this master thesis due to time constraints. Instead, we add more variables by including interaction terms, which are a combination of our current predictors.

5.2.6 Models with Interaction Terms

In this section we try to fit models where we include interaction terms. We first chose to fit a LASSO model to the full follow-up data set **FullFollowUp9799Imputed**, where we use all the variables and all two-way interaction terms for these variables. Note that four of the original variables are splined.

The ASD and ADHD models from this LASSO fit resulted in slightly better AUC scores than those found in Table 5.18 and Table 5.19 in the previous section. The tables are omitted as we instead report other models in this section. Like for all other models, these AUC scores are 10-fold cross-validated and therefore the AUC scores are an average of 10 different splits respectively, of which different variables are selected by LASSO. For the 10 splits we counted which interaction terms were included $8, 9$ or 10 times in the 10 splits for ASD and ADHD respectively. Based on how many times an original variable is included in these interaction terms and on the basis of important predictors identified from our classification trees, we believe that the following variables are important for our models:

ASD:
`Sex`, `F_Age`, `M_ADHD_Meds`, `F_ADHD_Meds`, `M_Smoking`, `GestAge`, `Malformations`.

ADHD:
`Sex`, `F_Age`, `M_ADHD_Meds`, `F_ADHD_Meds`, `F_Alc_Meds`, `M_Smoking`, `B_Weight`.

We therefore choose to fit a LASSO, a Ridge and a full logistic regression model based on the data set **FullFollowUp9799Imputed**, where all main effects and two-way interaction terms between the 7 selected variables for ASD and ADHD, respectively, are included. The resulting cross-validated evaluation measures are shown in Table 5.20 and 5.21. The AUC scores for both ASD and ADHD are slightly lower than those found in Table 5.18 and Table 5.19 indicating that this selection of variables and the choice of including interaction terms did not give us better prediction models. As we do not seem to be close to overfitting our logistic regression models by including too many predictors, the equally good or worse models are not due to overfitting, so maybe they are just bad predictors. Instead, we made models based on the data set **FullFollowUp9799Imputed**, but now including both all the predictors originally in this data set but also the interaction terms between the seven selected variables described above. We did not get better or worse prediction models than before, and thus including interaction terms seem not to be the way to go.

ASD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.038	0.986	0.698	0.542	4.361	2.791	4.361	0.646
Ridge	0.038	0.986	0.698	0.538	4.362	2.792	4.362	0.646
GLM Full	0.038	0.986	0.695	0.546	4.341	2.778	4.341	0.647

Table 5.20: 10-fold cross-validated evaluation measures for three different models predicting ASD in the data set **FullFollowUp9799Imputed** 220458(ASD 2.5%), with only 7 selected predictors and their interactions. For ROC curves see Figure C.7.

ADHD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.070	0.977	0.541	0.725	3.380	2.163	3.380	0.685
Ridge	0.070	0.977	0.534	0.737	3.336	2.135	3.336	0.685
GLM Full	0.073	0.977	0.526	0.747	3.286	2.103	3.286	0.687

Table 5.21: 10-fold cross-validated evaluation measures for three different models predicting ADHD in the data set **FullFollowUp9799Imputed** 220458(ADHD 3.5%), with only 7 selected predictors and their interactions. For ROC curves see Figure C.7.

5.2.7 Final Models

In this section we review in details the models fit to the data set **FullFollowUp9799-Imputed** with evaluation measures found in Table 5.18 and Table 5.19. We find that these models are currently the best predictive models as they take our calendar year problem into account. Furthermore, we have full follow-up for these subjects. We do not choose the models including interaction terms as they do not show a difference in the evaluation measures and they are considerably more difficult to interpret.

We now review the three logistic regression models `LASSO`, `Ridge` and `GLM Full` fitted to the data set **FullFollowUp9799Imputed**, but this time we do not use 10-fold cross-validation to estimate the coefficients for the logistic regression models, but use the entire data set in order to include the remaining 10% data into the fit. In Table 5.22 and Table 5.23 we see evaluation measures for the three logistic regression models fitted to the entire data set **FullFollowUp9799Imputed** and tested on the same data set. We see that we are not close to overfitting our models as the evaluation measures are only slightly better than those found by 10-fold cross-validation in Table 5.18 and Table 5.19. For both ASD and ADHD, the best models based on AUC scores are the full models `GLM Full`, which include all variables from the data set **FullFollowUp9799Imputed**. This again indicates that the number of predictors included in our logistic regression models do not lead to overfitting, as there seem to be no use for shrinkage. We choose not to show results for the two logistic regression models `GLM 5` and `GLM 10`, as these are models containing predictors that we ourselves have chosen.

ASD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.039	0.985	0.656	0.582	4.100	2.624	4.100	0.656
Ridge	0.039	0.985	0.660	0.577	4.124	2.639	4.124	0.655
GLM Full	0.038	0.986	0.698	0.546	4.364	2.793	4.364	0.660

Table 5.22: Evaluation measures for three different models predicting ASD in the data set **FullFollowUp9799Imputed** for 220,458 subjects (ASD 2.5%).

ADHD	PPV	NPV	Sens	Spec	$F_{0.5}$	F_1	F_2	AUC
LASSO	0.064	0.979	0.599	0.679	3.743	2.395	3.742	0.695
Ridge	0.070	0.978	0.552	0.730	3.453	2.210	3.453	0.695
GLM Full	0.072	0.978	0.540	0.745	3.375	2.160	3.375	0.699

Table 5.23: Evaluation measures for three different models predicting ADHD in the data set **FullFollowUp9799Imputed** with 220,458 subjects (ADHD 3.5%).

Coefficients for the three logistic regression models `LASSO`, `Ridge` and `GLM Full` for ASD and ADHD can be seen in Table 5.24 and Table 5.25, fitted to the whole data set **FullFollowUp9799Imputed**. Table 5.24 includes all non splined variables in the models and Table 5.25 includes all the splined variables included in the three logistic regression models. Note that coefficients below 0 are protective and coefficients higher than 0 are risk factors for having the diagnose. Based on coefficients higher than 0.2 and the selection done by LASSO in Table 5.24, the most important non splined predictors seem to be:

ASD:
`Sex`, `M_ADHD_Meds`, `F_ADHD_Meds`, `M_Smoking`, `Malformations`, `Sepsis`.

ADHD:
`Sex`, `M_ADHD_Meds`, `F_ADHD_Meds`, `M_Alc_Meds`, `F_Alc_Meds`, `F_Drugs_Meds`, `M_Smoking`.

We do not want to comment too much on the coefficients for the splined variables as they are not straightforward to interpret. Note, however, that LASSO chooses at least one splined predictor for both ASD and ADHD for the variables `M_Age`, `F_Age`, `B_Weigth`, but not the splined `GestAge` variables. In addition, note that the supreme most important predictors we have in our study to predict both ASD and ADHD seem to be `Sex`, `F_ADHD_Meds` and `M_ADHD_Meds`.

Note how some predictors have NA as coefficient for the GLM Full. This is because these predictors are completely explained by a linear combination of other predictors, and they are thus excluded from the coefficient estimation. Furthermore, the predictors for which the coefficient is blank for the LASSO models, indicate that the coefficient is set to 0 exactly.

In this section we also show how Ridge regression shrink the coefficient estimates that belong to each predictor towards 0 for ASD and ADHD in Figures 5.17 and Figure 5.18. In addition we show how LASSO sets the coefficient estimates to 0 exactly in Figure 5.19 and 5.20. The black dashed line in all four plots indicates the estimated penalty parameter that gives the best AUC score for the model. For example, in Table 5.24, we see that the model fitted with LASSO for ADHD has an estimate of 2.007 for the coefficient belonging to the predictor `M_ADHD_Meds`, which can also be found in Figure 5.20 (It is the point where the leftmost curve hits the black dashed line).

As an example we have tried to predict the probability of two fictive subjects, born between 1997 and 1999, getting a diagnosis of ASD or ADHD. We have constructed a girl, who is the second child that her mother has born, a mother who has redeemed

a prescription of anti drug medication prior to the girls first birthday. We have also constructed a boy, who was born three weeks (21 days) prior to the mean gestational age of all other children. His parents have both redeemed a prescription of ADHD medicine before his first birthday. Besides these characteristics, the two subjects have the mean in all continuous predictors and 0 in all binary and categorical predictors.

Predicting their probabilities of ASD from the GLM Full model with coefficients as in Table 5.24 and Table 5.25, we get that the girl has a 0.9% risk of getting a diagnosis of ASD before her 18th birthday and the boy has a 23.4% risk of getting a diagnosis of ASD before his 18th birthday. With the recommended threshold (based on sensitivity plus specificity) being 2.5% the girl would be predicted as not getting a diagnosis, whereas the boy would be predicted as getting a diagnosis.

When we predict the probabilities of getting a diagnosis of ADHD before the 18th birthday, with the GLM Full model, the girl's risk is 1.6% and the boy's risk is 48.8%. With a recommended threshold of 3.8%, the girl would be predicted as not getting a diagnosis of ADHD and the boy would be predicted as getting a diagnosis of ADHD.

ASD	LASSO	Ridge	GLM Full
(Intercept)	-4.190	-4.059	-6.371
Sex	0.802	0.339	0.878
M_ADHD_Meds	1.134	0.809	1.210
F_ADHD_Meds	0.911	0.658	1.017
M_Alc_Meds		0.086	0.121
F_Alc_Meds		0.035	0.023
M_Drugs_Meds		-0.061	-0.227
F_Drugs_Medi		0.012	-0.058
M_Smoking	0.156	0.117	0.218
M_Smoking_NA	-0.056	-0.050	-0.094
M_Spon_Abort		0.020	0.051
Parity_2nd	-0.038	-0.049	-0.222
Parity_3rd	-0.095	-0.091	-0.374
Parity_4th	-0.069	-0.113	-0.481
Parity_5th		0.022	-0.180
Parity_6thNM	-0.167	-0.275	-1.110
Parity_NA		-0.033	-0.073
Ext_Preterm		0.052	0.461
Ext_Preterm_NA		-0.017	0.049
Ver_Preterm		0.012	0.133
Ver_Preterm_NA		-0.017	NA
Mod_Preterm		-0.006	-0.082
Mod_Preterm_NA		-0.017	NA
Visit_Mid		0.002	0.003
Visit_Doc		0.007	0.012
Visit_Spe		0.004	0.004
Med_Initiate		0.075	0.154
Med_Initiate_NA	-0.445	-0.068	-0.624
Sectio	0.024	0.048	0.049
Sectio_NA	0.000	-0.069	NA
B_Length		0.003	0.002
Apgar5minOK		0.033	0.082
Apgar5minOK_NA		-0.038	0.031
Malformations	0.185	0.158	0.285
Malformations_NA	0.000	-0.069	NA
In_Asfyxi		-0.196	-0.599
In_Asfyxi_NA	0.000	-0.069	NA
Sepsis	0.324	0.275	0.453
Sepsis_NA	0.000	-0.068	NA
Infections		0.126	0.257
Jaundice		0.022	-0.013

ADHD	LASSO	Ridge	GLM Full
(Intercept)	-4.158	-3.775	-5.580
Sex	0.651	0.383	0.753
M_ADHD_Meds	2.007	1.735	2.013
F_ADHD_Meds	1.549	1.321	1.573
M_Alc_Meds	0.077	0.214	0.218
F_Alc_Meds	0.213	0.238	0.299
M_Drugs_Meds		0.185	0.156
F_Drugs_Medi	0.087	0.257	0.252
M_Smoking	0.490	0.340	0.519
M_Smoking_NA		-0.028	0.006
M_Spon_Abort	0.008	0.054	0.079
Parity_2nd		0.033	0.116
Parity_3rd		0.055	0.169
Parity_4th		0.099	0.231
Parity_5th		0.233	0.431
Parity_6thNM		-0.018	0.021
Parity_NA		-0.018	0.076
Ext_Preterm		-0.073	0.695
Ext_Preterm_NA		-0.013	0.052
Ver_Preterm		-0.090	0.208
Ver_Preterm_NA		-0.013	NA
Mod_Preterm	0.050	0.074	0.130
Mod_Preterm_NA		-0.013	NA
Visit_Mid		-0.002	0.003
Visit_Doc		0.020	0.046
Visit_Spe	0.035	0.031	0.048
Med_Initiate		0.035	0.019
Med_Initiate_NA	-0.280	-0.061	-0.548
Sectio		0.013	0.018
Sectio_NA	0.000	-0.061	NA
B_Length		-0.003	-0.011
Apgar5minOK		0.015	0.083
Apgar5minOK_NA		-0.008	0.153
Malformations	0.027	0.112	0.170
Malformations_NA	0.000	-0.061	NA
In_Asfyxi		-0.057	-0.137
In_Asfyxi_NA	0.000	-0.061	NA
Sepsis		0.119	0.198
Sepsis_NA		-0.061	NA
Infections		-0.112	-0.263
Jaundice		-0.009	-0.039

Table 5.24: Coefficients from logistic regression models fitted to the data set **FullFollowUp9799Imputed** with 220,458 subjects (ASD 2.5% and ADHD 3.5%) with ASD and ADHD as response, respectively.

ASD	LASSO	Ridge	GLM Full
M_Age1		-0.036	0.888
M_Age2		-0.031	0.958
M_Age3		0.070	0.651
M_Age4	0.188	0.160	2.617
M_Age5		-0.126	0.777
F_Age1		-0.055	0.054
F_Age2		0.005	0.082
F_Age3	0.308	0.148	0.574
F_Age4		0.048	0.190
F_Age5		0.040	0.016
GestAge1		0.024	0.922
GestAge2		-0.042	0.826
GestAge3		-0.024	0.453
GestAge4		0.037	1.727
GestAge5		-0.048	0.315
B_Weigth1		-0.048	0.122
B_Weigth2		0.014	0.255
B_Weigth3		0.052	0.101
B_Weigth4	0.164	0.124	1.055
B_Weigth5		-0.201	0.482

ADHD	LASSO	Ridge	GLM Full
M_Age1		-0.034	0.006
M_Age2		-0.076	0.078
M_Age3	-0.028	-0.133	-0.578
M_Age4	0.734	0.363	1.440
M_Age5	-0.015	-0.516	1.171
F_Age1		-0.052	-0.255
F_Age2		-0.048	-0.259
F_Age3		-0.024	-0.041
F_Age4	0.359	0.225	0.017
F_Age5		-0.304	0.005
GestAge1		-0.015	1.384
GestAge2		0.014	1.499
GestAge3		-0.006	0.823
GestAge4		0.085	2.898
GestAge5		-0.171	0.518
B_Weigth1		-0.012	0.280
B_Weigth2		-0.013	0.328
B_Weigth3		-0.064	-0.089
B_Weigth4	0.202	0.170	1.357
B_Weigth5	-0.351	-0.303	0.800

Table 5.25: Coefficients from logistic regression models fitted to the data set **FullFollowUp9799Imputed** with 220,458 subjects (ASD 2.5% and ADHD 3.5%) with ASD and ADHD as response, respectively. The splined predictors.

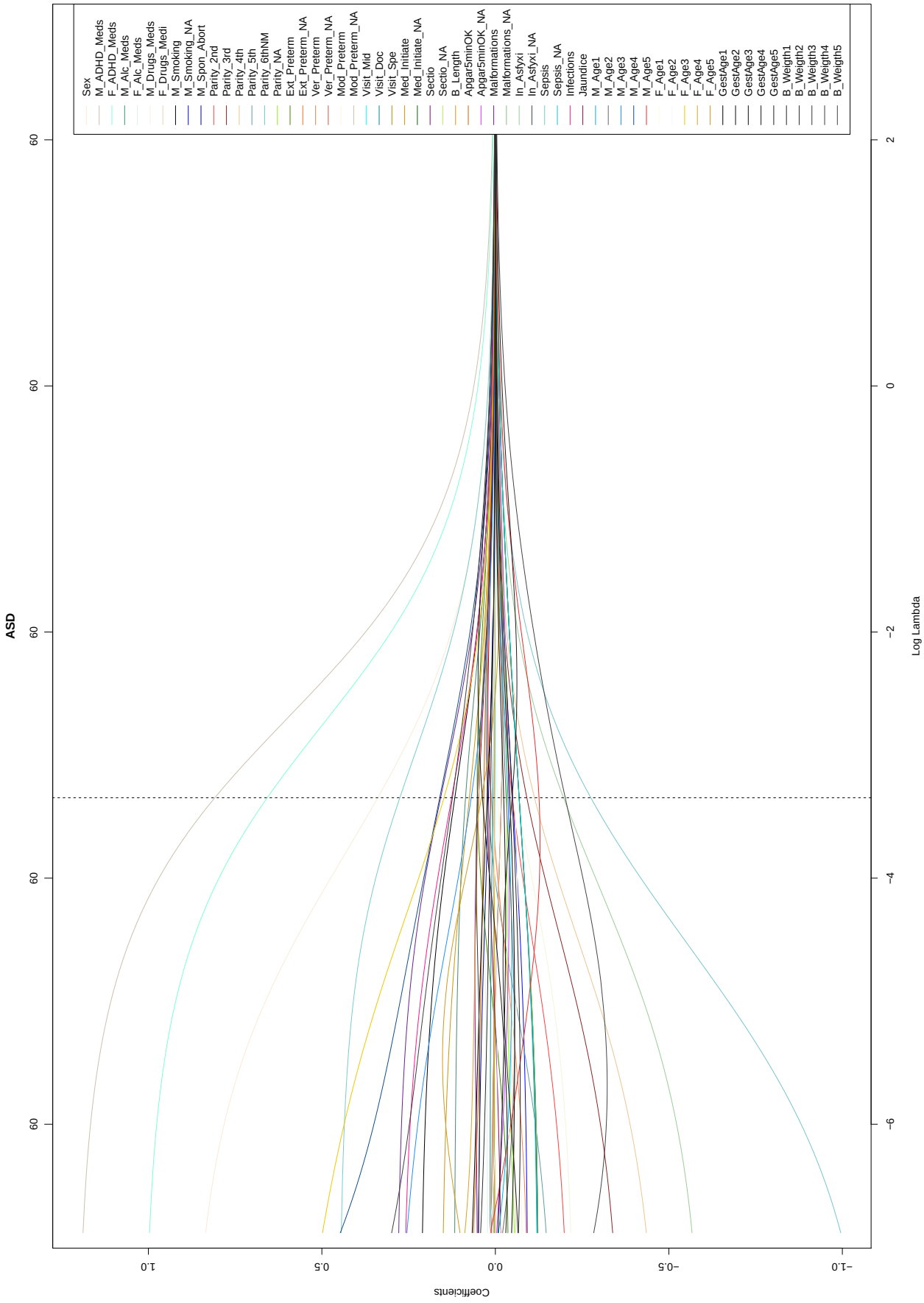


Figure 5.17: The Ridge penalized coefficients of logistic regression for different values of λ. The regression is fitted on the data set **FullFollowUp9799Imputed** with 220,458 subjects (ASD 2.5%) with ASD as response. The dotted line representing the optimal λ, chosen by 10-fold cross-validated AUC.

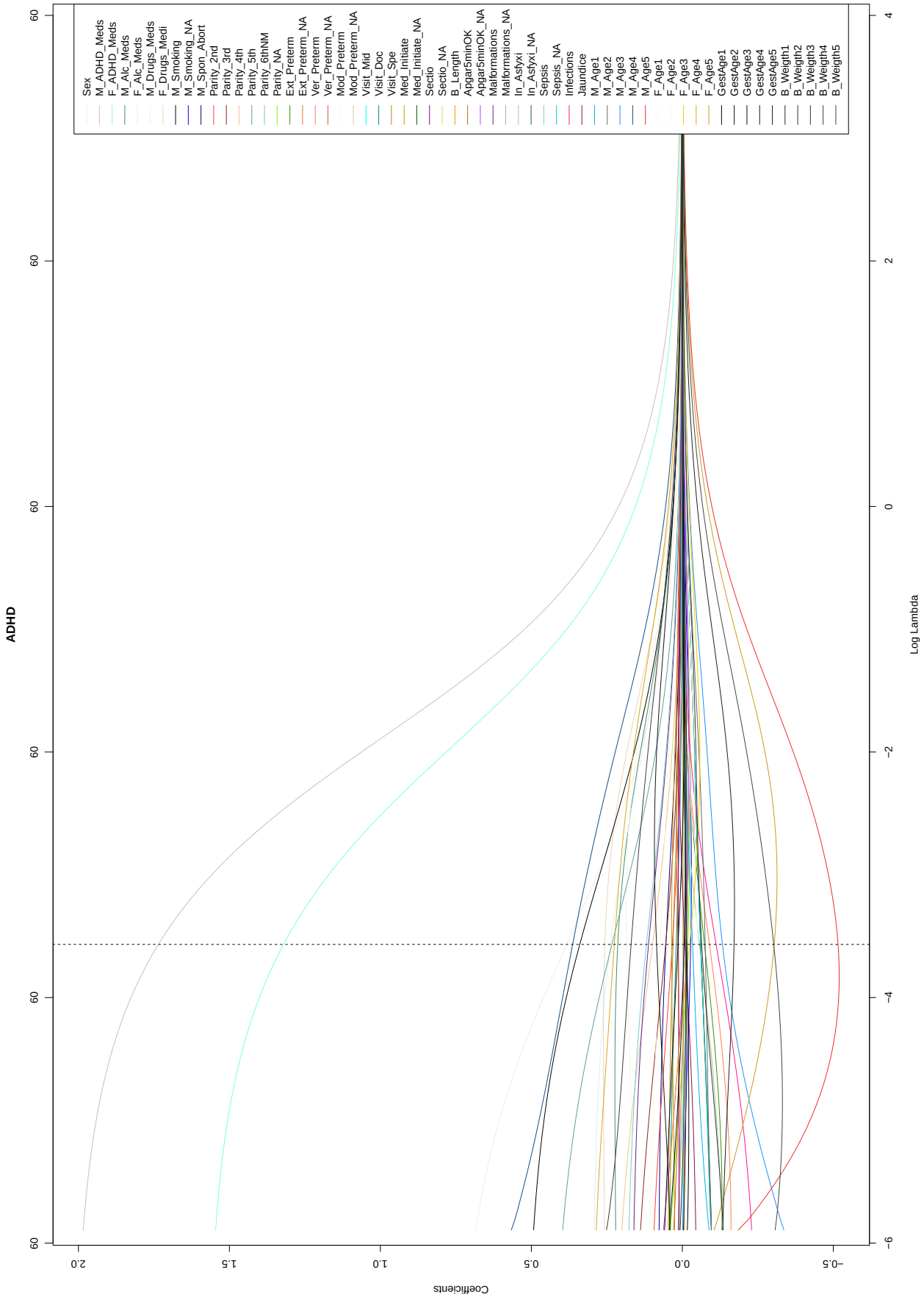


Figure 5.18: The Ridge penalized coefficients of logistic regression for different values of λ. The regression is fitted on the data set **FullFollowUp9799Imputed** with 220,458 subjects (ADHD 3.5%) with ADHD as response. The dotted line representing the optimal λ, chosen by 10-fold cross-validated AUC.

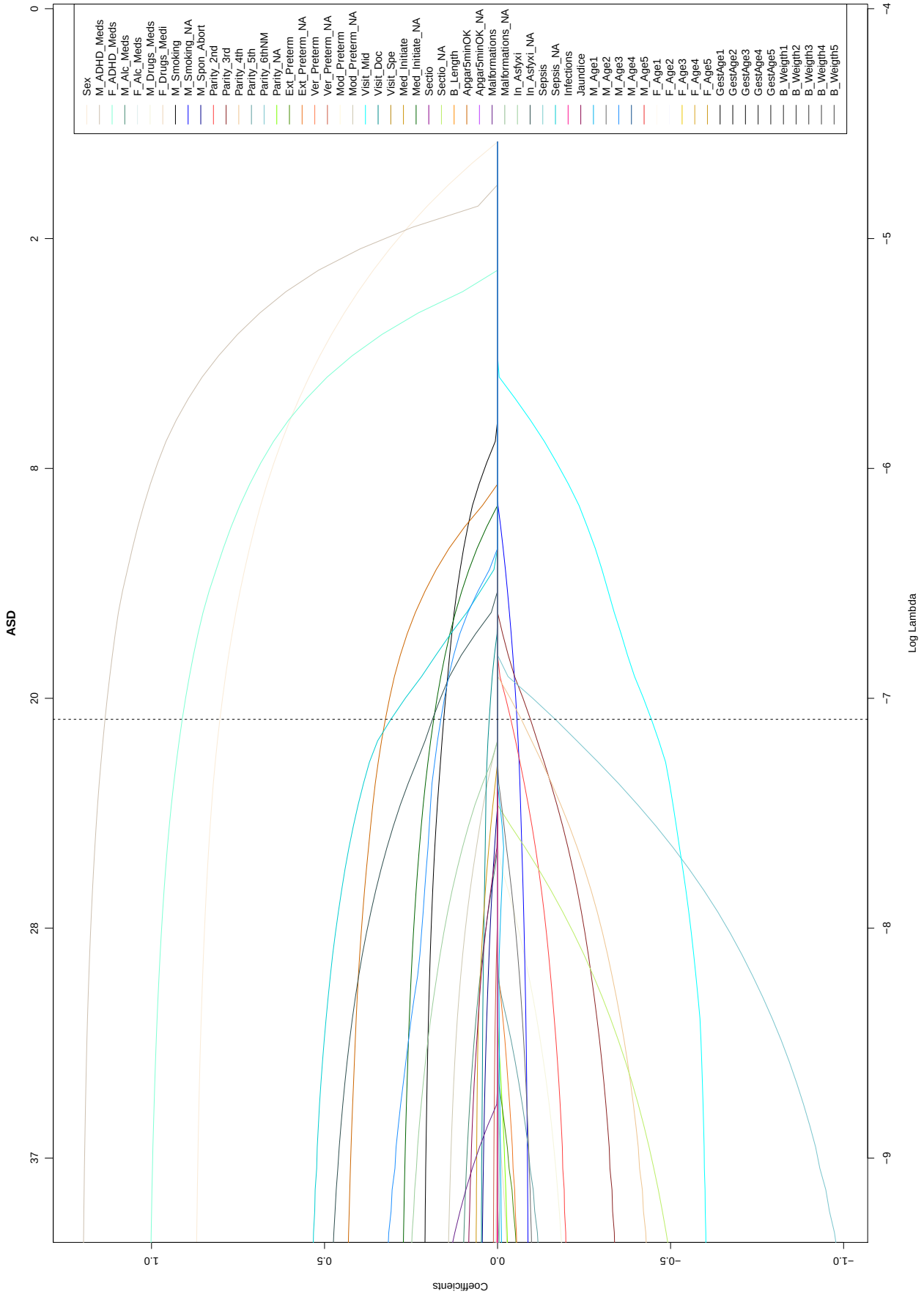


Figure 5.19: The LASSO penalized coefficients of logistic regression for different values of λ. The regression is fitted on the data set **FullFollowUp9799Imputed** with 220,458 subjects (ASD 2.5%) with ASD as response. The dotted line representing the optimal λ, chosen by 10-fold cross-validated AUC.

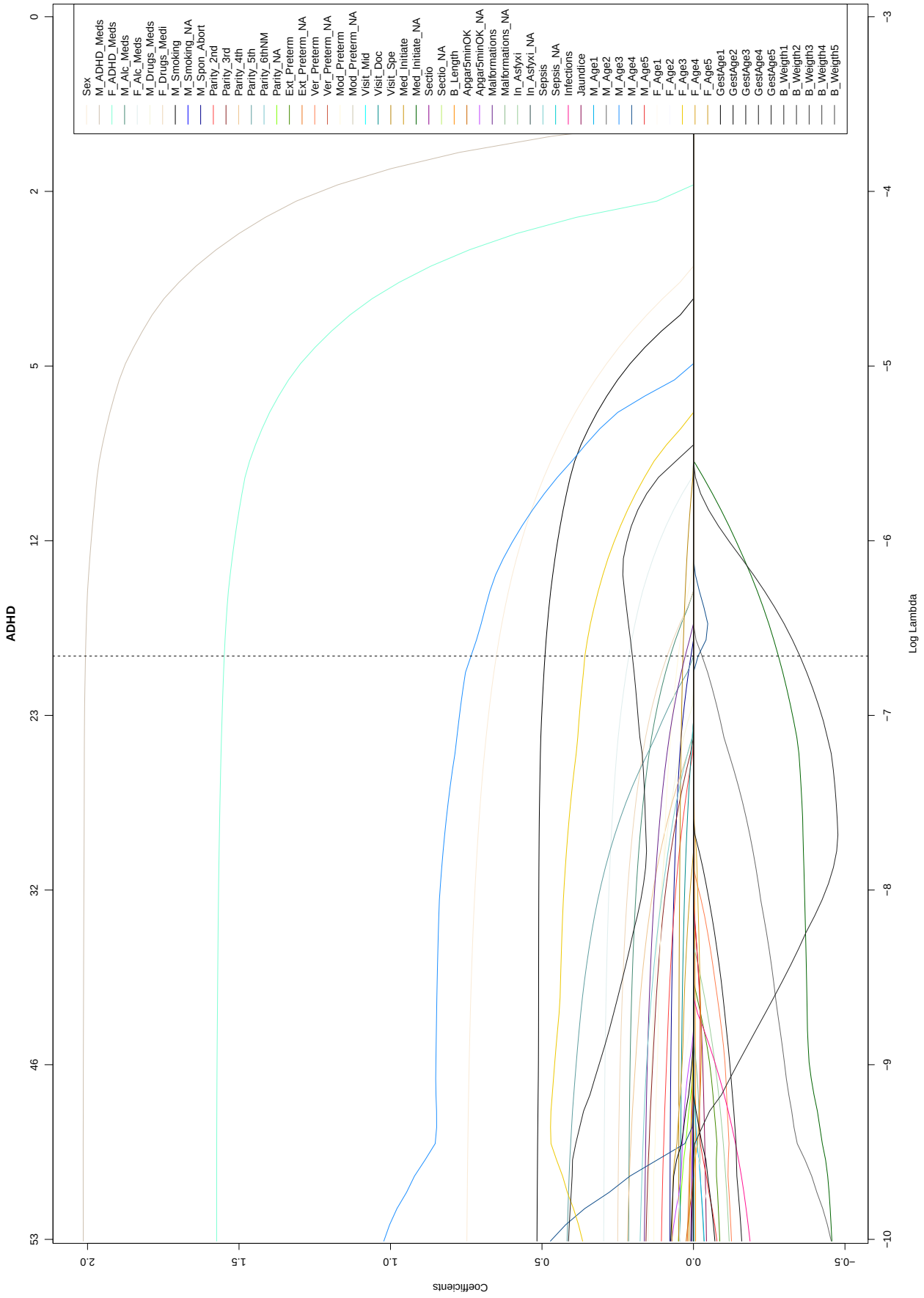


Figure 5.20: The LASSO penalized coefficients of logistic regression for different values of λ. The regression is fitted on the data set **FullFollowUp9799Imputed** with 220,458 subjects (ADHD 3.5%) with ADHD as response. The dotted line representing the optimal λ, chosen by 10-fold cross-validated AUC.

6. Concluding Remarks

We have investigated several classification methods theoretically and applied the ones, that we found most relevant to data from the Danish registers. We have found that logistic regression is a powerful classifier, but the diagnoses of ASD and ADHD are not straightforward to predict. In this chapter we discuss the main problems faced in this master thesis, how we handle them and how the work of the master thesis can be expanded. The chapter is ended by our conclusions.

Designing a Study and Commissioning a Data Set

The master thesis preparation started with us writing a study protocol in cooperation with the Research unit for Child and Adolescent Psychiatry and The Psychiatric Research Unit, North Denmark Region Psychiatry, Aalborg, Denmark. Properly designing and describing the project was necessary to obtain the approval from the relevant agencies needed to order the data set. It was soon clear too, that properly designing a study was not straightforward and it turned out to be quite time consuming. We ended up spending the first month of our master thesis on designing a study and deciding which variables to order.

In the process we acquired knowledge about how to design a predictive study, where our main source was "To explain or to Predict" by Galit Shmueli [34]. To be able to order the correct variables, we consulted Marlene Briciet Lauritsen[1], an experienced child and adolescent psychiatrist and expert on ASD. Based on her recommended variables, we had to plan how to create these variables and in which registers that information could be found. When selecting which variables to order, we focused on quality and accessibility as we had learned from Shmueli.

Furthermore, managing the acquired data and actually creating the variables, we had already designed before we ordered the data, turned out to be a time consuming task. The amount of time put into the design and creation of the actual variables used in the analysis have given us the insight into the data needed to actually conduct the analysis, and was therefore well spent.

[1] Child and Adolescent Psychiatry, Region of Northern Jutland Psychiatry, Aalborg, Denmark

Simulated Data

While waiting for the real data set, we simulated a data set, such that we could familiarize ourselves with the methods we sought to use in our study. This proved to be of great value to multiple aspects of the thesis. We got a better understanding of the theory behind our methods. By applying the methods to constructed data, we could investigate how they behave and familiarize ourselves with the different kinds of results, we could expect.

The value of getting to know how the R implementations of the models functioned and how data should be processed for a model to be applied to it, cannot be understated. We saved many hours of debugging when the real data set arrived, because of the familiarity with the R implementations earned on the simulated data set.

Being able to evaluate the different classification methods on the simulated data also helped us decide, which methods to use on the real data set, and which methods not to use. The application of the methods to the simulated data had great impact on how this thesis is formed and helped to decide where to place our focus and time. We think that one should always start out on simulated data when familiarizing oneself with new methods.

Exploratory Data Analysis

We might have laid too heavy focus on exploratory data analysis in the planning and theoretical phase. After the process of constructing the real variables from the raw data and our consultations with psychiatrist Marlene Briciet Lauritsen[1], we were quite familiar with our data set already before the actual exploratory data analysis. Thus the clustering methods only confirmed what we already knew or what was unsurprising, like birth length and birth weight clustering.

When we evaluated the number of clusters preferred (made the elbow plot) for hierarchical clustering, we used total within cluster sum of squares, as we did for k-means clustering. This might not have been the best way to evaluate the clusters, as we had already earlier decided that an Euclidean based distance measure should not be used for hierarchical clustering in our setting. We should instead have found a correlation based measure for within cluster dissimilarity.

The number of predictors we chose to construct did not necessitate dimension reduction, and we thus skipped PCA in the analysis.

Even though we did not see great use of the clusterings for our real data, one could use the clusters when evaluating the results of the LASSO variable selections. As a predictor might not be unimportant just because it is excluded by the LASSO, it might just be that the predictor is closely related to another predictor that explains almost the same.

Logistic Regression

As we never got to overfit the logistic regression models, we did not get to explore the benefits of the shrinkage provided by LASSO and Ridge on the real data. We still expect

[1] Child and Adolescent Psychiatry, Region of Northern Jutland Psychiatry, Aalborg, Denmark

there to be a benefit in shrinkage, but this might not be before more predictors are added.

We did not find time to conduct multi class logistic regression. Exploring the multi class setting might provide great insight, as we see already in Table 5.2 and Table 5.3 that there is a great overlap between the two diagnoses. When identifying predictors suited for predicting ASD, the predictors for parental ADHD medication, keeps showing up. Had we worked in a multi class setting, we expect the double diagnosed to be predictable by these predictors and could thus focus more on the pure ASD class.

Other Classifiers

Not using LDA in the analysis seemed obvious based on [1] stating that results should almost be similar to logistic regression. Had we conducted the multi class analysis, LDA, and especially QDA, could have been interesting to use.

The classifier kNN was not used on real data either. This was due to the fact that, as an online classifier, the kNN method would not be applicable to a real life setting, as data in the Danish registers is access restricted.

Classification trees on the other hand proved to be of great value to the project. As a classifier, it was not particularly worse than the logistic regression, which it was expected to be severely outperformed by. But the advantageous uses of the visual interpretations of the pruned trees justify the classification trees' role in this thesis. The pruned trees quickly visualized which predictors could be of importance to a prediction model and played a big role in us identifying `BirthYear` as an issue. This was backed by the incidence and prevalence plots. Identifying important predictors by another method than the thesis' main method, logistic regression, helped us to be more confident in the findings from our logistic regression analysis. We could have laid a heavier focus on classification trees and evaluated pruned trees as predictors. This might have led to us recommending the marginally worse predictor, but with an easy to follow flow-chart as the actual prediction model, as opposed to presenting the coefficients in Table 5.24 and Table 5.25.

Evaluation Measures

We could have split up our data in such a way that external validation could have been done on data from a held-out hospital or some held-out years. We chose to only perform internal validation, as any splitting of the original data set, to create an external one, would reduce the data on which we can perfect our models. We would rather create the best possible model than prove how generalizable it is. Furthermore, we believe that cross-validation is an acceptable approximation of generalizability. We have thus 10-fold cross-validated all of our evaluation measures. We could have used bootstrap samples instead, but did not, due to time constraints and because we find cross-validation sufficient. The choice of AUC as the main evaluation measure seemed straightforward, as it measures the overall performance of a type of model, instead of just a specific model given by a certain threshold. In the end we choose a model/threshold based on maximum sensitivity and specificity. We could instead have chosen models based on the F-scores, taking the balance between sensitivity and positive predictive value into account. It could

have been interesting to see, whether the final model had been chosen as the same, had we optimized the F-score all the way trough instead of the AUC.

Variable Selection

All variable selection ended out being done "manually", as we decided to exclude variables based mainly on temporal coverage. But in the reduced model with interaction terms, we did base our decisions on what predictors to include on LASSO results, clustering indications and the classification trees. The way we did it is in fact in violation with the cross-validation principle. We looked at LASSO performed on all ten folds, before we decided which predictors to include. The clustering methods being trained on and classification tree grown on the whole data set, and not folds of it, was also in violation with the cross-validation principle. To implement variable selection into a cross-validated setup, the variable selection needs to be done by an algorithm, which we did not find the time nor the need to do. This is also why we never got to apply forward/backwards variable selection to the real data.

The Analysis

Instead of focusing on design and evaluation of prediction models we ended out spending most of the analysis selecting a data set on which to perform the analysis. In this process we identified time (calendar year) as the main obstacle in the way of us being able to design and evaluate credible prediction models. Mainly time in the sense, that as the incidence of ASD and ADHD is rising every calendar year, this increase is not straightforward to model. The time issue extended into our handling of missing values, as the missingness often was informative about calendar year. How to handle missingness in general, and informative missingness in particular, could be the aim of a whole thesis in itself. We suspect that our handling of missing values could have been done in a more proper way, e.g. by following [46].

Conducting the evaluation of our prediction models on the data set **FullFollow-Up9799Imputed**, seems to be valid, as we have full follow-up on the subjects and the missingness should not be informative, at least not about calendar year. The models that we evaluated should be applicable for the cohorts a few years ahead, but at some point the incidence has risen to an extend where the models should be refitted. Even though the models might be valid to use, the AUC scores we obtained are not in a range where any of the models presented can be categorized as good prediction models. In a clinical setting the models presented seem to be of little value. For the purpose of investigating the correspondence between the theory behind ASD and ADHD and real life, this thesis does provide some insight. We did find that the hereditary component of ADHD is present, as the medication of parents with ADHD medication turned out to be a strong predictor. The link between malformations and ASD [47] also seem to be confirmed, as the malformation predictor contributed to the ASD prediction models.

In the sense of finding new and not yet investigated links, the link between maternal

BMI and both ADHD and ASD is interesting. Even though we could not include it in the models for the data set **FullFollowUp9799Imputed**, we still believe it is worth looking into in future studies.

According to [48], high explanatory power, e.g. a large odds-ratio does not necessarily lead to high predictive power. In our setting the high-scorers on the odd-ratio, maternal and paternal ADHD medicinal use, also seem to be central prediction variables. This is also the case for malformations for ASD. On the other hand, maternal smoking with an OR of 2.283 for ADHD has played a lesser role in our prediction models. We have not investigated the individual predictive power of our predictors, and are thus not able to conclude anything definitely.

In [34] it is stated that, how much the results of corresponding predictive and explanatory studies differ is an indication of how far the theory within a given field is from reality. We think that the theory on both ASD and ADHD is not yet complete, as we were not able to construct good prediction models, based on several predictors suggested both by the literature and a psychiatric expert. We do still believe that the prediction models can be improved by adding more predictors, which is the plan to do in the future.

Perspectives

Conducting a statistical prediction study will contribute to the debate on which of the already known, and possibly new risk factors, associate with ASD and ADHD. This might even lead to hypotheses on new predictors and their association with ASD and ADHD [34]. This is why we believe that the predictive study should be conducted to its full extent, including many more predictors.

As for the investigation into prediction models, we would like to investigate the ASD diagnoses given before a subject's third birthday. This would enable us to have full follow-up on our entire data set. Furthermore, children diagnosed with ASD so early are probably different from those diagnosed in their teenage years and it could thus be easier for us to predict such a diagnosis.

In our search for evaluation measures we have come across calibration plots several times, e.g. in [49]. These might provide even further insights into our prediction models and could thus be considered in the future.

To better handle the calendar year issues, models more suited for data collected over time, like survival analysis using Cox proportional hazards [50], could be considered. Even though the time varying prevalence is not easily handled here either, to the extend of our limited knowledge.

Another type of classifier to consider would be neural networks [41]. These are not better at predicting in general, but are well suited for large data sets. We do not know if the size of this data set could be considered "large", but as long as it is possible to fit a logistic regression, we do not see a need for neural networks.

Conclusion

We conclude that logistic regression is the best classification method for the data handled in this thesis. Classification trees are useful as a supplement and shrinkage might become important when more predictors are added.

A time dependent outcome with varying prevalence and informative censoring of variables, are the main issues for the data handled in this thesis.

To further investigate the models in this thesis, we recommend investigating the diagnosis of ASD before the age of three. To expand the work of the thesis, we recommend investigating a multi class setting. To go another way, we recommend survival analysis or neural networks. To improve the work, we recommend creating more predictors and learning how to handle the missingness.

Bibliography

[1] Gareth James et al. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York, 2013. ISBN: 978-1-4614-7137-0. DOI: 10.1007/978-1-4614-7138-7. arXiv: arXiv:1011.1669v3. URL: http://link.springer.com/10.1007/978-1-4614-7138-7.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York, 2009, p. 745. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7. arXiv: 1010.3003. URL: http://link.springer.com/10.1007/978-0-387-84858-7.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.r-project.org/.

[4] Hjördis O. Atladottir et al. "The increasing prevalence of reported diagnoses of childhood psychiatric disorders: a descriptive multinational comparison". In: *European Child and Adolescent Psychiatry* 24.2 (2014), pp. 173–183. ISSN: 1435165X. DOI: 10.1007/s00787-014-0553-8. URL: http://link.springer.com/10.1007/s00787-014-0553-8.

[5] Stefan N Hansen, Diana E Schendel, and Erik T Parner. "Explaining the increase in the prevalence of autism spectrum disorders: The proportion attributable to changes in reporting practices". In: *JAMA Pediatrics* 169.1 (2015), pp. 56–62. ISSN: 21686203. DOI: 10.1001/jamapediatrics.2014.1893. URL: http://jamanetwork.com/pdfaccess.ashx?url=/data/journals/peds/931882/.

[6] Erik T. Parner et al. "A comparison of autism prevalence trends in Denmark and Western Australia". In: *Journal of Autism and Developmental Disorders* 41.12 (2011), pp. 1601–1608. ISSN: 01623257. DOI: 10.1007/s10803-011-1186-0. URL: http://link.springer.com/10.1007/s10803-011-1186-0.

[7] Guilherme Polanczyk et al. "The worldwide prevalence of ADHD: A systematic review and metaregression analysis". In: *American Journal of Psychiatry* 164.6 (2007), pp. 942–948. ISSN: 0002953X. DOI: 10.1176/ajp.2007.164.6.942. arXiv: arXiv:1011.1669v3. URL: http://psychiatryonline.org/doi/abs/10.1176/ajp.2007.164.6.942.

[8] Anita Thapar et al. “Practitioner review: What have we learnt about the causes of ADHD?” In: *Journal of Child Psychology and Psychiatry and Allied Disciplines* 54.1 (2013), pp. 3–16. ISSN: 00219630. DOI: 10.1111/j.1469-7610.2012.02611.x.

[9] Anita Thapar. “Discoveries on the Genetics of ADHD in the 21st Century: New Findings and Their Implications”. In: *American Journal of Psychiatry* 175.10 (2018), appi.ajp.2018.1. ISSN: 0002-953X. DOI: 10.1176/appi.ajp.2018.18040383. URL: http://ajp.psychiatryonline.org/doi/10.1176/appi.ajp.2018.18040383.

[10] Elizabeth Hisle-Gorman et al. “Prenatal, perinatal, and neonatal risk factors of autism spectrum disorder”. In: *Pediatric Research* 84.2 (2018), pp. 190–198. ISSN: 0031-3998. DOI: 10.1038/pr.2018.23. URL: http://www.nature.com/articles/pr201823.

[11] Bernie Devlin and Stephen W. Scherer. “Genetic architecture in autism spectrum disorder”. In: *Current Opinion in Genetics and Development* 22.3 (2012), pp. 229–237. ISSN: 0959437X. DOI: 10.1016/j.gde.2012.03.002. arXiv: NIHMS150003. URL: http://dx.doi.org/10.1016/j.gde.2012.03.002http://www.ncbi.nlm.nih.gov/pubmed/22463983http://linkinghub.elsevier.com/retrieve/pii/S0959437X12000366.

[12] Stephen V. Faraone et al. *Molecular genetics of attention-deficit/hyperactivity disorder*. 2005. DOI: 10.1016/j.biopsych.2004.11.024. URL: https://www.sciencedirect.com/science/article/pii/S0006322304012260?via{\%}3Dihub.

[13] Padideh Karimi et al. “Environmental factors influencing the risk of autism”. In: *Journal of Research in Medical Sciences* 22.1 (2017). ISSN: 17357136. DOI: 10.4103/1735-1995.200272.

[14] Emma Sciberras et al. “Prenatal Risk Factors and the Etiology of ADHD—Review of Existing Evidence”. In: *Current Psychiatry Reports* 19.1 (2017), p. 1. ISSN: 15351645. DOI: 10.1007/s11920-017-0753-2. URL: http://www.ncbi.nlm.nih.gov/pubmed/28091799http://link.springer.com/10.1007/s11920-017-0753-2.

[15] Paul Bryde Axelsson et al. “Investigating the effects of cesarean delivery and antibiotic use in early childhood on risk of later attention deficit hyperactivity disorder”. In: *Journal of Child Psychology and Psychiatry* (2018). ISSN: 00219630. DOI: 10.1111/jcpp.12961. URL: http://doi.wiley.com/10.1111/jcpp.12961.

[16] William Thompson et al. “Maternal thyroid hormone insufficiency during pregnancy and risk of neurodevelopmental disorders in offspring: A systematic review and meta-analysis”. In: *Clinical Endocrinology* 88.4 (2018), pp. 575–584. ISSN: 13652265. DOI: 10.1111/cen.13550. arXiv: 0608246v3 [arXiv:physics].

[17] William Mandy and Meng-Chuan Lai. “Annual Research Review: The role of the environment in the developmental psychopathology of autism spectrum condition”. In: *Journal of Child Psychology and Psychiatry* 57.3 (2016), pp. 271–292. ISSN: 00219630. DOI: 10.1111/jcpp.12501. URL: http://doi.wiley.com/10.1111/jcpp.12501.

[18] Rosalind J. Neuman et al. “Prenatal Smoking Exposure and Dopaminergic Genotypes Interact to Cause a Severe ADHD Subtype”. In: *Biological Psychiatry* 61.12 (2007), pp. 1320–1328. ISSN: 00063223. DOI: 10.1016/j.biopsych.2006.08.049. URL: https://www.sciencedirect.com/science/article/pii/S0006322306011401.

[19] Malhar Kumar, Andrei Baklanov, and Daniel Chopin. “Correlation between sagittal plane changes and adjacent segment degeneration following lumbar spine fusion”. In: *European Spine Journal* 10.4 (2001), pp. 314–319. ISSN: 09406719. DOI: 10.1007/s005860000239. URL: http://link.springer.com/10.1007/s005860000239.

[20] Elaine Tierney et al. “Abnormalities of cholesterol metabolism in autism spectrum disorders”. In: *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics* 141.6 (2006), pp. 666–668. ISSN: 15524841. DOI: 10.1002/ajmg.b.30368. URL: http://doi.wiley.com/10.1002/ajmg.b.30368.

[21] C M Freitag. *The genetics of autistic disorders and its clinical relevance: A review of the literature*. 2007. DOI: 10.1038/sj.mp.4001896. URL: http://www.nature.com/articles/4001896.

[22] Ruth M. Pfeiffer et al. “Risk Prediction for Breast, Endometrial, and Ovarian Cancer in White Women Aged 50 y or Older: Derivation and Validation from Population-Based Cohort Studies”. In: *PLoS Medicine* 10.7 (2013). Ed. by Eduardo L. Franco, e1001492. ISSN: 15491277. DOI: 10.1371/journal.pmed.1001492. URL: http://dx.plos.org/10.1371/journal.pmed.1001492.

[23] Duminda N. Wijeysundera et al. “Derivation and Validation of a Simplified Predictive Index for Renal Replacement Therapy After Cardiac Surgery”. In: *JAMA* 297.16 (2007), p. 1801. ISSN: 0098-7484. DOI: 10.1001/jama.297.16.1801. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.297.16.1801.

[24] Keith A A Fox et al. “Prediction of risk of death and myocardial infarction in the six months after presentation with acute coronary syndrome: prospective multinational observational study (GRACE)”. In: *BMJ* 333.7578 (2006), pp. 1091–1091. ISSN: 0959-8138. DOI: 10.1136/bmj.38985.646481.55. URL: http://www.ncbi.nlm.nih.gov/pubmed/17032691http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1661748http://www.bmj.com/cgi/doi/10.1136/bmj.38985.646481.55.

[25] Margaret A. Shipp et al. “Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning”. In: *Nature Medicine* 8.1 (2002), pp. 68–74. ISSN: 10788956. DOI: 10.1038/nm0102-68. URL: http://www.nature.com/articles/nm0102-68.

[26] T. M. Reynolds and M. D. Penney. “The Mathematical Basis of Multivariate Risk Screening: With Special Reference to Screening for Down’s Syndrome Associated Pregnancy”. In: *Annals of Clinical Biochemistry: An international journal of biochemistry and laboratory medicine* 27.5 (1990), pp. 452–458. ISSN: 0004-5632. DOI: 10.1177/000456329002700506. URL: http://acb.sagepub.com/lookup/doi/10.1177/000456329002700506.

[27] Adam Mourad Chekroud et al. “Cross-trial prediction of treatment outcome in depression: A machine learning approach”. In: *The Lancet Psychiatry* 3.3 (2016), pp. 243–250. ISSN: 22150374. DOI: 10.1016/S2215-0366(15)00471-X. URL: https://linkinghub.elsevier.com/retrieve/pii/S221503661500471X.

[28] Søren D. Østergaard et al. “Predicting ADHD by assessment of Rutter’s indicators of adversity in infancy”. In: *PLoS ONE* 11.6 (2016). Ed. by Phillipa J. Hay, e0157352. ISSN: 19326203. DOI: 10.1371/journal.pone.0157352. URL: http://dx.plos.org/10.1371/journal.pone.0157352.

[29] D. Van Der Meer et al. “Predicting attention-deficit/hyperactivity disorder severity from psychosocial stress and stress-response genes: A random forest regression approach”. In: *Translational Psychiatry* 7.6 (2017), e1145. ISSN: 21583188. DOI: 10.1038/tp.2017.114. URL: http://www.nature.com/doifinder/10.1038/tp.2017.114.

[30] Tzlil Einziger et al. “Predicting ADHD Symptoms in Adolescence from Early Childhood Temperament Traits”. In: *Journal of Abnormal Child Psychology* 46.2 (2017), pp. 1–12. ISSN: 00910627. DOI: 10.1007/s10802-017-0287-4. URL: http://link.springer.com/10.1007/s10802-017-0287-4.

[31] E Skafidas et al. “Predicting the diagnosis of autism spectrum disorder using gene pathway analysis”. In: *Molecular Psychiatry* 19.4 (2014), pp. 504–510. ISSN: 14765578. DOI: 10.1038/mp.2012.126. arXiv: 0208024 [gr-qc]. URL: http://www.nature.com/articles/mp2012126.

[32] William J. Bosl, Helen Tager-Flusberg, and Charles A. Nelson. “EEG Analytics for Early Detection of Autism Spectrum Disorder: A data-driven approach”. In: *Scientific Reports* 8.1 (2018), p. 6828. ISSN: 20452322. DOI: 10.1038/s41598-018-24318-x. URL: http://www.nature.com/articles/s41598-018-24318-x.

[33] Hans Collins, G., Reitsma and Doug Altman. “Reporting Guideline for Prediction Model Studies : TRIPOD Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis”. In: *Annals of Internal Medicine* 162.1 (2015), pp. 55–63. ISSN: 0003-4819. DOI: 10.7326/M14-0697. URL: http://annals.org/article.aspx?doi=10.7326/M14-0697http://annals.org/article.aspx?doi=10.7326/M14-0697{\%}0Aavailableatwww.annals.org.

[34] Galit Shmueli. “To explain or to predict”. In: *Statistical Science* 25.3 (2010), pp. 289–310. DOI: 10.1214/10-STS330. arXiv: arXiv:1101.0891v1. URL: https://www.stat.berkeley.edu/{~}aldous/157/Papers/shmueli.pdfhttps://www.svds.com/machine-learning-vs-statistics/.

[35] Ewout W. Steyerberg and Yvonne Vergouwe. *Towards better clinical prediction models: Seven steps for development and an ABCD for validation*. 2014. DOI: 10.1093/eurheartj/ehu207. arXiv: arXiv:1011.1669v3. URL: http://www.ncbi.nlm.nih.gov/pubmed/24898551http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4155437https://academic.oup.com/eurheartj/article-lookup/doi/10.1093/eurheartj/ehu207.

[36] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation". In: 2006, pp. 1015–1021. ISBN: 978-3-540-49787-5. DOI: 10.1007/11941439_114. URL: http://link.springer.com/10.1007/11941439{_}114.

[37] Helle Wallach Kildemoes, Henrik Toft Sørensen, and Jesper Hallas. "The Danish national prescription registry". In: *Scandinavian Journal of Public Health* 39.7 (2011), pp. 38–41. ISSN: 14034948. DOI: 10.1177/1403494810394717. URL: http://journals.sagepub.com/doi/10.1177/1403494810394717.

[38] Morten Schmidt et al. *The Danish National patient registry: A review of content, data quality, and research potential*. 2015. DOI: 10.2147/CLEP.S91125. URL: http://www.ncbi.nlm.nih.gov/pubmed/26604824http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4655913.

[39] *Lagrange multipliers, introduction | Khan Academy*. URL: https://www.khanacademy.org/math/multivariable-calculus/applications-of-multivariable-derivatives/constrained-optimization/a/lagrange-multipliers-single-constraint (visited on 03/28/2019).

[40] I T Jolliffe. *Principal Component Analysis. Second Edition*. Vol. 98. 2002, p. 487. ISBN: 0-387-95442-2. DOI: 10.1007/b98835. arXiv: arXiv:1011.1669v3. URL: http://link.springer.com/10.1007/b98835.

[41] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006, p. 738. ISBN: 9780387310732.

[42] Søren Højsgaard. "Topics in Statistical Science TOSTA – 2018 Generalized linear models Generalized estimating equations". 2018.

[43] Henrik Madsen and Poul Thyregod. *Introduction to general and generalized linear models*. CRC Press, 2011, p. 302. ISBN: 1420091557. URL: https://books.google.dk/books/about/Introduction{_}to{_}General{_}and{_}Generalized.html?id=JhuDNgAACAAJ{\&}redir{_}esc=y.

[44] Xindong. Wu and Vipin Kumar. *The top ten algorithms in data mining*. CRC Press, 2009, p. 215. ISBN: 9781420089646.

[45] Frank E. Harrell. *Regression modeling strategies : with applications to linear models, logistic and ordinal regression, and survival analysis*. 2015, p. 582. ISBN: 9783319194257.

[46] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2002. ISBN: 9781119013563. DOI: 10.1002/9781119013563. URL: http://doi.wiley.com/10.1002/9781119013563.

[47] Katja M. Lampi et al. "Risk of autism spectrum disorders in low birth weight and small for gestational age infants". In: *Journal of Pediatrics* 161.5 (2012), pp. 830–836. ISSN: 00223476. DOI: 10.1016/j.jpeds.2012.04.058. URL: https://www.sciencedirect.com/science/article/pii/S0022347612004945.

[48] Margaret Sullivan Pepe et al. "Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker." In: *American journal of epidemiology* 159.9 (2004), pp. 882–90. ISSN: 0002-9262. DOI: 10.1093/aje/kwh101. URL: https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwh101http://www.ncbi.nlm.nih.gov/pubmed/15105181.

[49] Ewout W Steyerberg et al. "Assessing the performance of prediction models: a framework for traditional and novel measures." In: *Epidemiology (Cambridge, Mass.)* 21.1 (2010), pp. 128–38. ISSN: 1531-5487. DOI: 10.1097/EDE.0b013e3181c30fb2. URL: http://www.ncbi.nlm.nih.gov/pubmed/20010215http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3575184.

[50] John P. Klein and Melvin L. Moeschberger. *Survival analysis : techniques for censored and truncated data*. Springer, 2003, p. 536. ISBN: 9780387216454.

[51] Julie Anne Quinn et al. "Preterm birth: Case definition & guidelines for data collection, analysis, and presentation of immunisation safety data". In: *Vaccine* 34.49 (2016), pp. 6047–6056. ISSN: 18732518. DOI: 10.1016/j.vaccine.2016.03.045. URL: http://www.ncbi.nlm.nih.gov/pubmed/27743648http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5139808.

[52] Sundhedsdatastyrelsen. *Fødselsregisteret (MFR) - Sundhedsdatastyrelsen*. URL: https://sundhedsdatastyrelsen.dk/da/registre-og-services/om-de-nationale-sundhedsregistre/graviditet-foedsler-og-boern/foedselsregisteret (visited on 06/08/2019).

[53] Sundhedsdatastyrelsen. *Det Medicinske Fødselsregister*. URL: https://www.esundhed.dk/Registre/Det-medicinske-foedselsregister (visited on 06/08/2019).

[54] D S Y Chong and J Karlberg. "Refining the Apgar score cut-off point for newborns at risk". In: *Acta Paediatrica, International Journal of Paediatrics* 93.1 (2004), pp. 53–59. ISSN: 08035253. DOI: 10.1080/08035250310007295. URL: http://www.ncbi.nlm.nih.gov/pubmed/14989440.

A. Simulated Variables

Name	Origin
`Height`	As described in Section 1.1
`Height1`	$3\texttt{Height} + \mathcal{N}(0, 16)$
`Height2`	$\frac{1}{7}\texttt{Height} + \mathcal{N}(0, 0.25)$
`Height3`	$(\texttt{Height} + \mathcal{N}(0, 04))^8$
`Sex`	As described in Section 2.1.2
`Smoke`	As described in Section 2.1.2
`Weight`	As described in Section 1.1
`Weight1`	$-\exp\left(\frac{\texttt{Weight}}{30}\right) + \mathcal{N}(0, 0.16)$
`Weight2`	$8\texttt{Weight} + \mathcal{N}(0, 100)$
`Weight3`	$\log\left(\texttt{Weight}^4\right) + \mathcal{N}(0, 0.01)$
`Uniform`	$\text{Unif}(0, 1)$
`Uniform1`	$20\texttt{Uniform} + \mathcal{N}(0, 1.44)$
`Uniform2`	$-10\texttt{Uniform} + \mathcal{N}(0, 4)$
`Uniform3`	$-(\texttt{Uniform} + \mathcal{N}(0, 0.16))^2$
`Normal`	$\mathcal{N}(50, 30)$
`Normal1`	$30\texttt{Normal} + \mathcal{N}(0, 225)$
`Normal2`	$-15\texttt{Normal} + \mathcal{N}(0, 64)$
`Normal3`	$-(\texttt{Normal} + \mathcal{N}(0, 100))^2$

Table A.1: Simulated variables and transformations.

B. Real Variables

Other

- `PNR`: The pseudo randomized personal record number (PNR).
  The variable is taken directly from the CPR register and links all our observations, both subjects and parents across all the registers.

- `M_PNR`: The mother of the child's pseudo randomized personal record number (PNR).
  The variable is made based on the LFOED variable `PNRM` covering $1973 - 1996$ and the variable `PNRM` from MFR covering $1997-$, indicating the biological mother of the child. The merged variable `M_PNR` has a few missings, which was instead taken from the CPR register, this is the legal mother of the child the 1st of January 2019. The variable `M_PNR` is covering $1973-$.

- `F_PNR`: The father of the child's pseudo randomized personal record number (PNR).
  The variable is made based on the LFOED variable `PNRF` covering $1973 - 1996$ and the variable `PNRF` from MFR covering $1997-$, indicating the biological father of the child. The merged variable `F_PNR` has many missings, which was instead taken from the CPR register, this is the legal father of the child the 1st of January 2019. The variable `F_PNR` is covering $1973-$.

Binary variables

- `Sex`: The sex of the subject.
  The `Sex` variable is taken directly from the CPR register indicating the legal sex of the subject the 1st of January 2019. The variable covers all subjects and has no missing values.

- `ASD`: Indicating whether the child has gotten a diagnosis of Autism Spectrum Disorder before it's 18th birthday.
  The variable is based on the LPR, LPR-PSYK and DCPR. In all three registers we have excluded diagnoses given in the emergency room and referral diagnosis as these two are usually not confirmed by specialists. Between the remaining diagnosis, we have identified all diagnoses with the ICD 8 classification 299 and the ones with the IDC 10 classification F84. For all of these diagnoses we have reduced them to the

first one given to each subject and evaluated whether the admission date for the record leading to the diagnosis was prior to the subject's 18th birthday. If this was the case `ASD` is set to 1, otherwise `ASD` is set to 0. Overall `ASD` is an indicator variable indicating if the child has gotten an ASD diagnosis before it's 18th birthday. The variable covers all children in the study, thus 1977-.

- `ADHD`: Indicating whether the child has gotten a diagnosis of Attention Deficit Hyperactive Disorder before it's 18th birthday.
  The variable is based on the LPR, LPR-PSYK and DCPR. In all three registers we have excluded diagnoses given in the emergency room and referral diagnosis, as these two are usually not confirmed by specialists. Between the remaining diagnosis, we have identified all diagnoses with the ICD 8 classifications 406.39 and 308, and the ones with the IDC 10 classifications F90.0, F90.1, F90.8 and F98.8. For all of these diagnoses we have reduced them to the first one given to each subject and evaluated whether the admission date for the record leading to the diagnosis was prior to the subject's 18th birthday. If this was the case `ADHD` is set to 1, otherwise `ADHD` is set to 0. Overall `ADHD` is an indicator variable indicating if the child has gotten an ADHD diagnosis before it's 18th birthday. The variable covers all children in the study, thus 1977-.

Note: All medicine variables are made based on the register LSR and are made in the same way. Therefore we only describe in detail how the variable `M_ADHD_Meds` is made.

- `M_ADHD_Meds`: Indication of whether the mother of the child redeemed a prescription of ADHD medicine before the child's first birthday.
  In the register LSR an observation represents a redeemed prescription, for each of these transactions, among other things, we get the three variables `PNR`, the variable `ATC`, denoting which medicine is redeemed during the transaction and the variable `expd`, which refers to the dispatch date from the pharmacy. A person is only in this database if she gets any medicine. Based on these three variables, we have made the variable `M_ADHD_Meds` in the following way. First, we extracted all observations with ATC codes N06BA02, N06BA04, N06BA07, N06BA09, N06BA12, C02AC02, as these are the ATC codes for ADHD medicine. For each personal record number `M_PNR`, we then only store the first dispatch date. If the first dispatch date is before the child's first birthday, we choose to assign the mother as being an ADHD medicine user. We chose that it should be before the child's first birthday as in a predictive study, we must not use variables recorded later in time to describe a response. In this case, the child having ADHD and according to psychiatrist Marlene Briciet Lauritsen[1], children generally does not get diagnosed with ADHD before the age of 1. Overall, the variable `M_ADHD_Meds` is an indicator variable, where 1 means that the mother of the child used ADHD medicine before the child's first birthday and 0 means that the mother has not taken any ADHD medicine before the child's first

[1]Child and Adolescent Psychiatry, Region of Northern Jutland Psychiatry, Aalborg, Denmark

birthday or that the mother has not taken ADHD medicine at all. The variable covers 1995−.

- `F_ADHD_Meds`: ADHD medicine for the father.
  We make this variable by using the same ATC codes as for the variable `M_ADHD_Meds`. The variable `F_ADHD_Meds` is an indicator variable, where 1 indicates, that the father has received ADHD medicine before the child's first birthday. This variable covers the period 1995−.

- `M_Alc_Meds`: Alcoholic medicine for the mother.
  We make this variable by using the ATC codes N07BB01, N07BB02, N07BB03, N07BB04, N07BB05. The variable `M_Alc_Meds` is an indicator variable, where 1 indicates, that the mother has received alcoholic medicine before the child's first birthday. This variable covers the period 1995−.

- `F_Alc_Meds`: Alcoholic medicine for the father.
  We make this variable by using the same ATC codes as for the variable `M_Alc_Meds`. The variable `F_Alc_Meds` is an indicator variable, where 1 indicates, that the father has received alcoholic medicine before the child's first birthday. This variable covers the period 1995−.

- `M_Drugs_Meds`: Drug medicine for the mother.
  We make this variable by using the ATC codes N07BC01,N07BC02, N07BC03, N07BC04, N07BC05, N07BC06, N07BC51. The variable `M_Drugs_Meds` is an indicator variable, where 1 indicates, that the mother has received drug medicine before the child's first birthday. This variable covers the period 1995−.

- `F_Drugs_Meds`: Drug medicine for the father.
  We make this variable by using the same ATC codes as for the variable `M_Drugs_Meds`. The variable `F_Drugs_Meds` is an indicator variable, where 1 indicates, that the father has received drug medicine before the child's first birthday. This variable covers the period 1995−.

- `M_Smoking`: Maternal smoking prior to birth.
  The variable is made based on the LFOED variable `B_RYGER` covering $1991 - 1996$ and the variable `RYGERSTATUSMODER` from MFR covering 1997−. The variable `B_RYGER` has non-smokers marked with 0 and smokers marked with 1 and a level with blanks, which we set as missing. The variable `RYGERSTATUSMODER` also has a level with blanks that is set as missing. Furthermore the variable has the levels $00, 10, 11, 20, 21, 22, 23, 29, 99$, which we have interpreted as if 00 is means 0, that is non-smokers. The level 99 indicates NA and the remaining levels are different kinds of smokers, but since we are interested in merging the variable with `B_RYGER`, we simply consider all these as the level 1, that is, they are smokers.
  Overall, the variable `M_Smoking` is thus an indicator variable containing the levels $0, 1$ and NA, which indicates whether the mother has been smoking prior to birth and the variable is covering 1991−.

- `Ext_Preterm`: This variable and the next two indicate whether the child was prematurely born based on the continuous variable `GestAge` (see continuous variables). This specific variable indicates an extreme premature birth.
  According to [51], doctors should try to rescue the child if the gestational age is at least 22 weeks and the source calls the child extremely early born if the gestational age is between $22 - 27$ weeks. We therefore make an indicator variable to indicate whether the subject is born with a gestational age in between $22 - 27$ weeks by dividing all entries in the variable `GestAge` by 7 because it is measured in days. We then rounded these values to the nearest whole number of weeks. If the child is born with a gestational age of 22 or 27 weeks or a number in between, then the person belongs to the level 1, indicating extremely early born. If the gestational age for a birth is less than 22 or more than 27, the person belongs to the level 0, indicating that the birth was not extremely preterm. In addition to these two levels we also have some missing values. The variable `Ext_Preterm` is covering 1973-.

- `Ver_Preterm`: This variable indicates whether the child was prematurely born based on the continuous variable `GestAge` (see continuous variables). This specific variable indicates a very premature birth.
  This variable is made exactly like the previous one, where this is just a very early birth instead of an extremely early birth. According to [51] a very early birth is indicated by having a gestational age of $28 - 31$ weeks, including week 28 and 31. Thus for the variable `Ver_Preterm`, a person belongs to level 1 if the person has a gestational age of $28, 29, 30$ or 31 weeks and if the person has a other number reported, then the person belongs to the level 0. There is also a level indicating missing.

- `Mod_Preterm`: This variable indicates whether the child was prematurely born based on the continuous variable `GestAge` (see continuous variables). This specific variable indicates a moderately premature birth.
  This variable is made exactly like the previous one, where this is just a moderate early birth instead of a very early birth. According to [51] a moderate early birth is indicated by having a gestational age of $32 - 37$ weeks, including week 32 and 37. Thus for the variable `Mod_Preterm`, a person belongs to level 1 if the person has a gestational age of $32, 33, 34, 35, 36$ or 37 weeks. If the person has another number reported, then the person belongs to the level 0. There is also a level indicating missing.

- `Cont_Stim`: This variable indicates whether the mother of the child received contraction stimulation during the birth of the child.
  The variable is made based on the LFOED variables `B_I1` covering $1978 - 1990$ and `B_VESTIM` covering $1991 - 1996$ and furthermore based on the variable `MARKOER_VESTIMULATION` from MFR covering $1999-$. For the LFOED variables 1 indicates that the mother of the child has been given contraction stimulation during the birth and 0 indicates that the mother has not received it. There are many miss-

ings for both these two variables, which is because they both only cover part of the overall period $1973 - 1996$. Therefore, these variables are combined by taking the variable `B_VESTIM` and changing its values to 0 if there is 0 in the variable `B_I1` for one person and correspondingly done with the level 1. For MFR, the variable `MARKOER_VESTIMULATION` covers $1999-$ and we have thus not observed the variable in the period $1997 - 1998$. Furthermore, the variable `MARKOER_VESTIMULATION` only contain values for the levels 1 and blank. Many of the blank ones are of course 0, but some of these observations can also belong to level 1, because we have an uncovered period. We therefore choose to put all persons to 0 who have the year of birth $1999-$ and people who were born in $1997 - 1998$ are set as missings.
Overall, the merged variable `Cont_Stim` indicates that if a person has 1, the mother of the child has received contraction stimulation and if the person has 0, the subject's mother has not been given contraction stimulation. The variable `Cont_Stim` covers $1978 - 1996$ and $1999-$.

- `Epidural`: Indication of whether there has been an epidural blockade during childbirth.
  The variable is made based on the categorical MFR variable `EPIDURALBLOKADE` covering $2000-$. The MFR variable has the levels blank,NAAD0, NAAD00, NAAD01, NAAD02, NAAD03, NAAD0A, NAD0B. All the levels starting with NAA are the ICD 10 codes for a kind of epidural that the mother has had during the childbirth, but since we in this master thesis are not interested in which kind of epidural, the variable is made as a binary variable. Here, 1 indicates that the mother had an epidural. All blanks was dealt with such that if the child has a year of birth in 2000 or later, the child belongs to the level 0 indicating that there was no epidural. Otherwise the child belongs to the level missing.
  Overall, the variable `Epidural` is an indicator variable, where level 1 means that the mother had a epidural blockade during birth and the level 0 indicates, that the mother has no epidural blockade during birth. The variable is covering $2000-$.

- `Med_Initiate`: Indication whether the mother of a child has received medical commencement during birth.
  The variable is made based on the LFOED variable `B_F3` covering $1991 - 1996$ and the variable `MARKOER_IGANGSAETTELSE` from MFR covering $1997-$. The LFOED variable `B_F3` does not cover a large period of LFOED and therefore there are of course many blanks for this variable. All persons who have a year of birth in the year 1990 or earlier are therefore set as missing, which resulted in all blanks becoming NA. For the MFR variable `MARKOER_IGANGSAETTELSE`, there are only the levels blank and 1. Since the variable covers the entire period for MFR, all these blanks are set to 0, as it is assumed that it is a check box where no marking means no medical commencement.
  Overall, the variable `Med_Initiate` is an indicator variable, where 1 indicates that the mother of the child has received medical commencement and 0 indicates, that the mother has not received medical commencement. The variable is covering $1991-$.

- `Sectio`: Indication of whether the child is delivered through a cesarean section. The variable is made based on the LFOED variables `B_I11` covering $1978-1990$ and `B_SECTIOU` covering $1991-1996$ and furthermore based on the variable `MARKOER_KEJSERSNIT` from MFR covering $1997-$. The LFOED variables `B_I11` and `B_SECTIOU` contains the levels $0, 1$ and blank and we thus choose to make it as an indicator variable. According to the official documentation of the register [52], 0 means no cesarean section, 1 means cesarean section and blank means not informed for both variables. The variables are covering different time period and thus everything should be okay, but these two variables give us $72,593$ sections and according to Sundhedsdatastyrrelsen [53]there should have been $141,597$ in this time period. We do note that some individuals have been removed from the data set, but not that many. This is odd and thus we should be careful when including this variable before 1997. For the MFR variable `MARKOER_KEJSERSNIT`, the levels are blank and 1, where blank is treated as no cesarean section since this variable covers the entire MFR period, referred as 0. The level 1 means that the mother had cesarean section when giving birth to the child. Note that this variable was not as we had hoped, as we are not only interested in whether the mother had a cesarean section during birth, but more interested in whether the mother got an acute cesarean section during birth. There are two variables in LFOED `B_I9` and `B_I12`, where `B_I9` indicates whether the mother has received a scheduled cesarean section and `B_i12` indicates whether the cesarean section was acute. Since both of these variables contained only the levels 0 and NA and each cover the periods $1978-1990$, we contacted Sundhedsdatastyrrelsen to hear more about these cesarean section variables, but we are still waiting for an answer and must therefore omit the acute cesarean section in our master thesis.
  Overall the variable `Sectio` is an indicator variable, where 0 means that the mother of the child has not been given a cesarean section during birth and 1 means that the mother of the child had a cesarean section during birth. The variable covers 1978-, but we are aware that the variable may be odd before 1997.

- `Apgar5minOK`: The Apgar score of the child recorded 5 minutes after birth. The variable is made based on the LFOED variable `V_APGAR5` covering $1978-1996$ and the variable `APGARSCORE_EFTER5MINUTTER` from MFR covering $1997-$. An Apgar score indicates the degree to which the child is alright, in this case 5 minutes after birth. The highest score is 10 indicating that the child breathes and looks healthy. The lowest score is 0, indicating that the child is dead 5 minutes after birth. All integers in between are also possible scores. Since 93% of newborns scored 10, there is not much variation in this variable. Therefore, we made it as an indicator variable based on [54], which suggests that the levels $8, 9$ and 10 correspond to a fine Apgar score and that the levels $0-7$ indicate a troublesome Apgar score. For this indicator variable 1 means that the child has a fine score, whereas level 0 indicates that the child has a poor score. Values that are not $0-10$ are set to missing, as these are not possible scores. We treat the MFR variable `APGARSCORE_EFTER5MINUTTER` in

the same way.
Overall, the combined variable `Apgar5minOK` is an indicator variable, where 1 indicates a fine Apgar score and 0 indicates a poor score. The variable is covering 1978−.

- `Malformations`: Indication of whether the child has malformations after birth noted by the midwife.
  The variable is made based on the LFOED variable `C_MISDAN` covering $1978 - 1996$ and furthermore based on the variable `MARKOER_B_MISDANNELSE` from MFR covering 1997−. The LFOED variable `C_MISDAN` contains the levels blank,$0, 1$ and 2 and since the MFR variable is an indicator variable, we also choose to make this as an indicator. According to the official documentation [52], 0 means not informed, 1 means the child has malformations and 2 means no malformations. As there were almost just as many observations for each level, it indicated that extremely many children were born with malformations and therefore we contacted Sundhedsdatastyrrelsen to hear if it could be right. They responded by sending a detailed documentation. Here it appears that for the period $1978 - 1986$, 0 means malformation, the level 1 has no observations included and 2 means no malformation / uninformed. In the period $1987 - 1990$, all observations are missing and for the period $1991 - 1996$, 0 means uninformed, 1 means no malformation and 2 means malformation. This answers our concern that we had from the start on this variable. We chose to put all children with malformations as 1 and children without malformations are set as 0. For the MFR variable `MARKOER_B_MISDANNELSE`, the levels are blank and 1, where blank is treated as no malformation since this variable covers the entire MFR period. The level 1 means malformation.
  Overall the variable `Malformations` is an indicator variable, where 0 means no malformations for the child at birth noted by the midwife and 1 means that the child has malformations. The variable covers 1978-1986 and 1991-.

- `In_Asfyxi`: Indication of whether the child had inter-uterine asphyxia during the mother's pregnancy period.
  The variable is made based on the binary MFR variable `INTRAUTERIN_ASFYXI` covering 1997−. The MFR variable includes only the levels blank and DO363, where DO363 indicates, that the child had asphyxia in the mother's uterus and thus referred as 1 for the newly made variable `In_Asfyxi`. Since the entire MFR period is covered, it is chosen to put all blanks as 0 indicating that the child had not asphyxia
  Overall, the variable `In_Asfyxi` is an indicator variable, where level 1 means that the child had asphyxia in uterus during the mother's pregnancy period and the level 0 indicates that the child had not. The variable is covering 1997−.

- `Sepsis`: Indication of whether the child had sepsis during the mother's pregnancy period.
  The variable is made based on the categorical MFR variable `SEPSIS_BARN` covering 1997−. The MFR variable has the levels blank, DP36, DP360, DP361, DP362,

DP363, DP364, DP365, DP368, DP369, which are ICD 10 codes for different kinds of sepsis. Since we are not interested in which kind of sepsis the children had, we choose to gather all the levels such that 1 indicates, that a child had sepsis. As the variable covers the entire MFR period, the blanks are considered non-sepsis.
Overall, the variable `Sepsis` is an indicator variable, where level 1 means that the child had sepsis in uterus during the mother's pregnancy period and the level 0 indicates that the child had not. The variable is covering 1997−.

- `Infections`: Indication of whether the child had an infection shortly after birth.
  The variable is made based on the binary MFR variable `MARKOER_INFEKTIONER` covering 1997−. The MFR variable include only the levels blank and 1, where 1 indicates, that the child had infections. Since the entire MFR period is covered, it is chosen to put all blanks as 0 indicating that the child had no infections.
  Overall, the variable `Infections` is an indicator variable, where level 1 means that the child had an infection shortly after birth and the level 0 indicates, that the child had no infections. The variable is covering 1997−.

- `Jaundice`: Indicating whether the child had gotten a diagnosis of jaundice before it's first birthday.
  The variable is based on the LPR. In the LPR we have excluded diagnoses given in the emergency room and referral diagnosis, as these two are usually not confirmed by specialists. Between the remaining diagnosis we have identified all diagnoses with the ICD 8 classifications 282, 774 and 7852 and the ones with the IDC 10 classifications A270, D58, D598, D599, E031, E742, E804, E805, E848, K729, K831, P550, P551, P559, P579, P58, P59, Q441, Q443 and R17. For all of these diagnoses we have reduced them to the first one given to each subject, and evaluated whether the admission date for the record leading to the diagnosis was prior to the subject's first birthday. If this was the case `Jaundice` is set to 1, otherwise `Jaundice` is set to 0. Overall `Jaundice` is an indicator variable indicating whether the child had gotten a jaundice related diagnosis before it's first birthday. The variable covers all children in the study, thus 1977-.

Categorical variables

- `Parity`: Number of births including the newborn child.
  The variable is made based on the LFOED variable `V_TIDLLEV` covering $1973-1996$ and the variable `PARITET` from MFR covering 1997−. The variable `V_TIDLLEV` indicates how many live born children the mother has given birth to, excluding the newborn child. We choose to include the new baby as MFR does and therefore we add 1 to every levels. In addition, the variable is made as a categorical variable with the levels NA, $1, 2, 3, 4, 5$ and the level 6 or more births. The variable `PARITET` indicates how many birth the mother has including still born children. We must therefore be aware that the two variables do not indicate exactly the same when we evaluate our models. The variable has a level with blank that is set as missing.

We also make this variable as a categorical NA,1, 2, 3, 4, 5 and the level 6 or more births.
Overall, the variable `Parity` is thus a categorical variable containing the levels NA,1, 2, 3, 4, 5 and 6 $\geq$, which indicates how many birth the mother has including the new born child and the variable is covering 1973−. Note that `Parity` can be misleading as the period 1973 − 1996 does not include still born children and the period 1997− includes still born children.

Continuous variables

- `M_Age`: Age of the mother at the birth of the child.
  The variable is made based on the LFOED variable `V_MALDER` covering 1973 − 1996 and the variable `ALDER_MODER` from MFR covering 1997−. The variable `V_MALDER` is numeric and indicates the age of the mother when she gives birth to the child. The variable `ALDER_MODER` indicates the same but for another period and there are no strange inputs for these variables and no missings. This could indicate, that the variable is made based on the CPR register. For the children not registered in the MFR and the children with missing `V_MALDER` we calculate the maternal age from the mothers birthday and the child's birthday in the CPR register.
  Overall, the variable `M_Age` is thus a numeric variable indicating the age of the mother when giving birth to the child and the variable is covering 1973−.

- `F_Age`: Age of the father at the birth of the child.
  The variable is made based on the LFOED variable `V_FALDER` covering 1973 − 1996 and the variable `ALDER_FADER` from MFR covering 1997−. The variable `V_FALDER` is numeric and indicates the age of the father when the child is born. The variable `ALDER_FADER` indicates the same but for another period. There are some few levels indicating that the father's age is 6 years or younger and even negative values, which is set to NA, as the next smallest values beyond 6 are 15 years. For the children not registered in the MFR and all other children with missing `F_Age` we calculate the paternal age from the father's birthday and the child's birthday in the CPR register.
  Overall, the variable `F_Age` is thus a numeric variable indicating the age of the father when the child is born and the variable is covering 1973−.

- `M_BMI`: Mother's BMI prior to pregnancy.
  The variable is made based on the numeric MFR variable `BMI_MODER` covering 2003−. There are many different values for BMI because it is a composition of height and weight for each person. To identify strange levels, we round the BMI values, where no remarkable levels were found and thus the previous values are retained. All BMI, which are equal to or smaller than 7 and BMI values of 68 or more is set as missing as these are extremely unrealistic BMI scores.
  Overall, the variable `M_BMI` is a numeric variable indicating the BMI of the mother at first doctor visit. The variable is covering 2003−.

- `M_Spon_Abort`: Number of previous spontaneous abortions for the mother before the birth of this child.
  The variable is made based on the numeric MFR variable `TIDLIGERESPONTANEABORTER` covering 1997−. The MFR variable has the levels $1-12$ and NA. Since the variable covers the entire MFR period, we choose to consider the the level NA as no previous spontaneous abortions.
  Overall, the variable `M_Spon_Abort` is a numeric variable indicating the number of previous spontaneous abortions for the mother. The variable is covering 1997−.

- `Visit_Mid`: Number of visits to midwife.
  The variable is made based on the LFOED variable `V_U1` covering $1978-1996$ and the variable `BESOEGHOSJORDEMODER` from MFR covering 1997−. The variable `V_U1` is numeric and indicates how many visits to midwives the mother had during pregnancy. The same is the case for the variable `BESOEGHOSJORDEMODER` and therefore they can be merged without doing any transformations. There are also missings in both variables.
  Overall, the variable `Visit_Mid` is thus numeric and indicates how many visit to midwife the mother had during pregnancy covering 1978−.

- `Visit_Doc`: Number of visits to doctor.
  The variable is made based on the LFOED variable `V_U2` covering $1978-1996$ and the variable `BESOEGHOSLAEGE` from MFR covering 1997−. The variable `V_U2` is numeric and indicates how many visits to doctor the mother had during pregnancy. The same is the case for the variable `BESOEGHOSLAEGE` and therefore they can be merged without doing any transformations. There are also missings in both variables.
  Overall, the variable `Visit_Doc` is thus numeric and indicates how many visit to doctor the mother had during pregnancy covering 1978−.

- `Visit_Spe`: Number of visits to specialist doctor.
  The variable is made based on the LFOED variable `V_U3` covering $1978-1996$ and the variable `BESOEGHOSSPECIALLAEGE` from MFR covering 1997−. The variable `V_U3` is numeric and indicates how many visits to specialist doctor the mother had during pregnancy. The same is the case for the variable `BESOEGHOSSPECIALLAEGE` and therefore they can be merged without doing any transformations. There are also missings in both variables.
  Overall, the variable `Visit_Spe` is thus numeric and indicates how many visit to specialist doctor the mother had during pregnancy covering 1978−.

- `BirthYear`: The year, that the subject was born.
  The variable is extracted from the subject's birthday in the CPR register. The variable `BirthYear` has full coverage.

- `GestAge`: Days between last menstruation and birth also known as gestational age.
  The variable is made based on the LFOED variable `V_SVLANGDE` covering $1973-1996$

and the variable GESTATIONSALDER_DAGE from MFR covering 1997−. The variable V_SVLANGDE is numeric and indicates how many weeks have passed between last menstruation and birth. All levels are multiplied by 7, so it is instead indicated in days. This is done as the variable GESTATIONALDER_DAGE from MFR is indicated in days. Therefore nothing needs to be done about the latter, which is also a numeric variable.
Overall, the variable GestAge is thus a numeric variable indicating how many days have gone between last menstruation and the birth. The variable is covering 1973−.

- B_Length: The birth length of the child.
  The variable is made based on the LFOED variable V_LANGDE covering $1973 - 1996$ and the variable LAENGDE_BARN from MFR covering 1997−. The variable V_LANGDE is numeric and indicates the length of the child when born recorded in centimeters without decimals. The variable LAENGDE_BARN indicates the same but for another period. The levels 90 and 99 is considered as missings.
  Overall, the variable B_Length is thus a numeric variable indicating the length of the child when born recorded in centimeters and the variable covers 1973−.

- B_Weight: The birth weight of the child.
  The variable is made based on the LFOED variable V_VAGT covering $1973 - 1996$ and the variable VAEGT_BARN from MFR covering 1997−. The variable V_VAGT is numeric and indicates the weight of the child when born. Since the variable is recorded in grams without decimals, there are many different values (in total 3543). We investigated how many observations occurred for each level to find out if 99 could, for example, indicate missing. There were no remarkable strange levels and therefore it was only decided to put $0, 9900, 9990$ and 9999 as missings as these are unrealistic and not other values were close to them. Similarly the MFR variable VAEGT_BARN is numeric and we set the same levels to missing as for the variable V_VAGT and furthermore the level 9920 is set as missing.
  Overall, the variable B_Weight is thus a numeric variable indicating the weight of the child when born recorded in grams and is covering 1973−.

C. ROC Curves of Models on Real Data

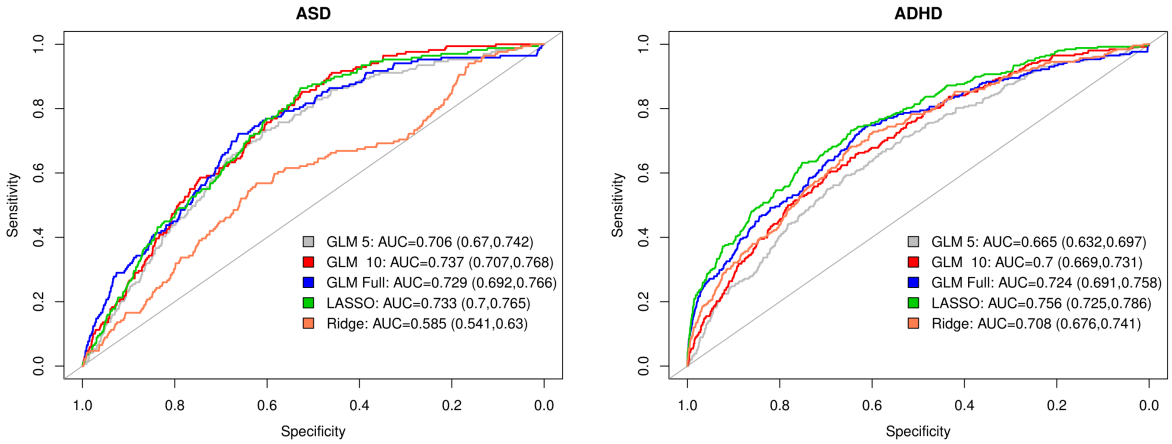


Figure C.1: 10-fold cross-validated ROC curves for five different models predicting ASD and ADHD, respectively, in $\frac{1}{200}$ of the data set **AllObsImputed** with 12,305 subjects, where 1.4% have an ASD diagnose and 2.1% have an ADHD diagnose

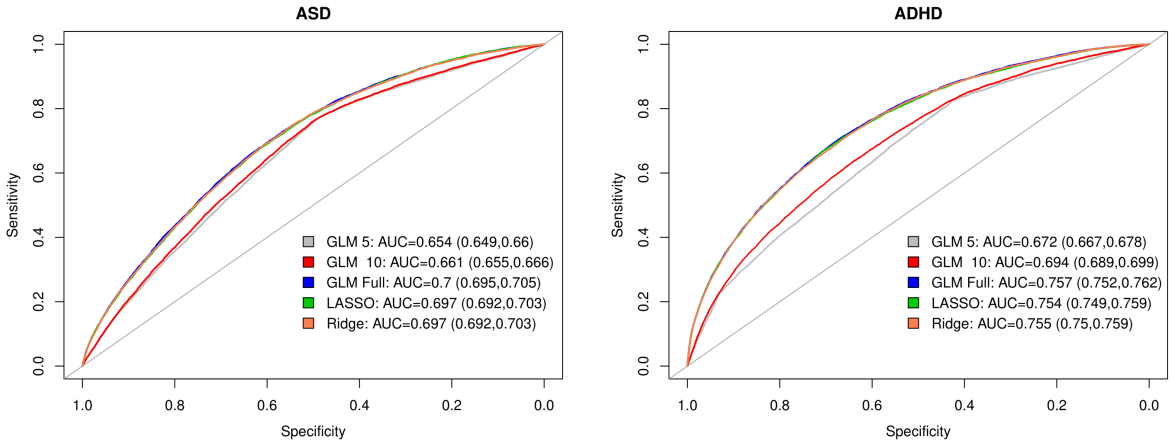


Figure C.2: 10-fold cross-validated ROC curves for five different models predicting ASD and ADHD, respectively, in the data set **ObsNoMissing** with 428,943 subjects, where 2.0% have an ASD diagnose and 2.4% have an ADHD diagnose

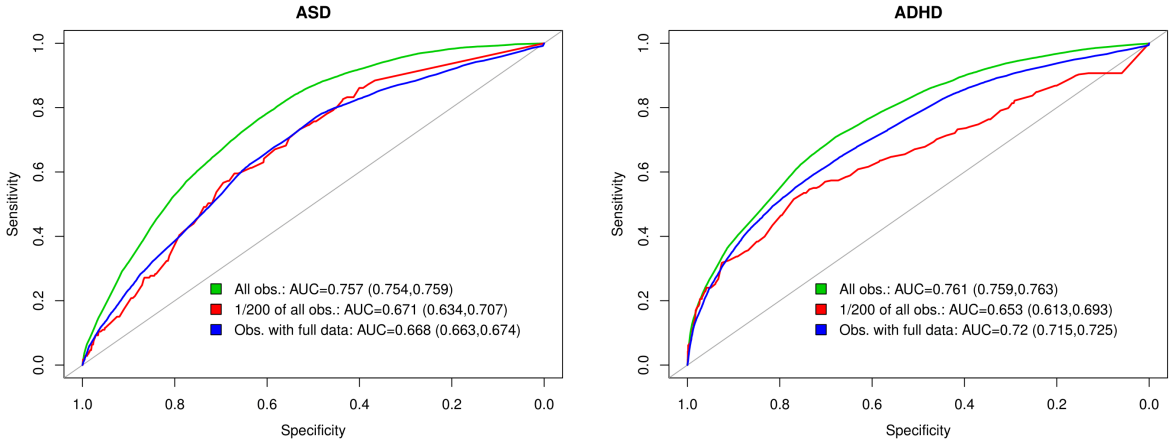


Figure C.3: 10-fold cross-validated ROC curves of trees predicting ASD and ADHD, respectively, grown to depth 8 on three different data sets. The data set **AllObsImputed** includes 2,461,082 subjects (ASD 1.5%, ADHD 2.0%). The data set $\frac{1}{200}$ of **AllObsImputed** includes 12,305 subjects (ASD 1.4%, ADHD 2.1%). The data set **ObsNoMissing** includes 428,943 subjects (ASD 2.0%, ADHD 2.4%). For ROC curves see Figure C.3

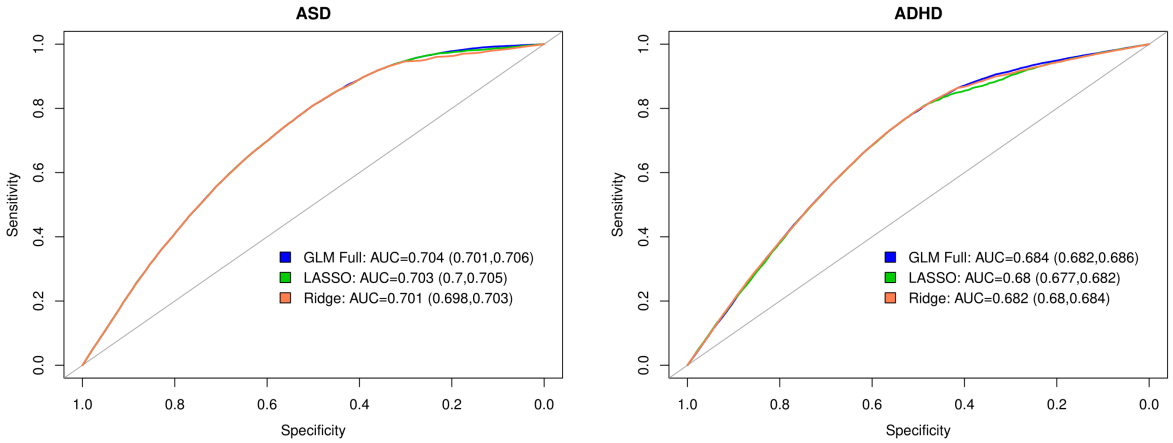


Figure C.4: 10-fold cross-validated ROC curves of three different models on the data set **AllObsImputed** with 2,461,082 subjects (ASD 1.5%, ADHD 2.0%) only using time as predictor for ASD and ADHD, respectively.

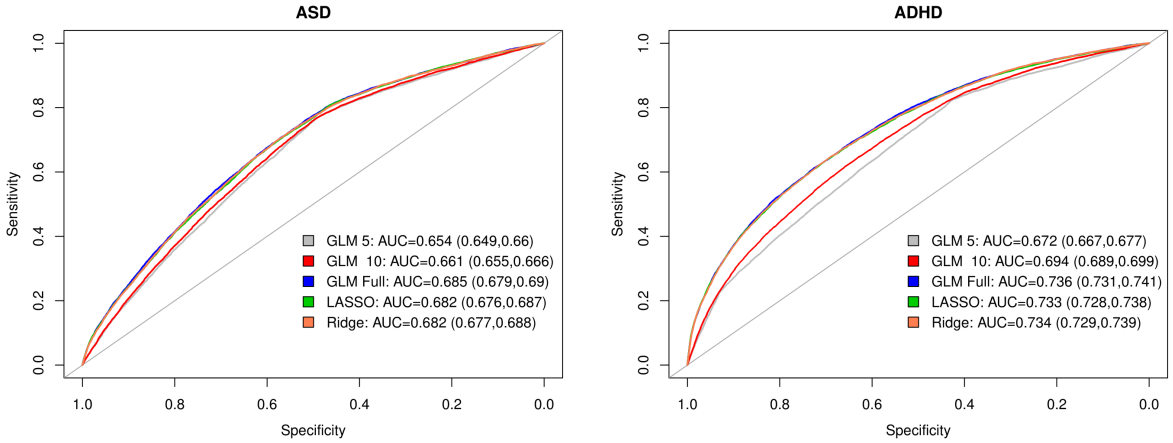


Figure C.5: 10-fold cross-validated ROC curves for five different models predicting ASD and ADHD, respectively, in the data set **ObsNoMissing** without `BirthYear` included as predictor. The data set includes 428,943 subjects (ASD 2.0%, ADHD 2.4%)

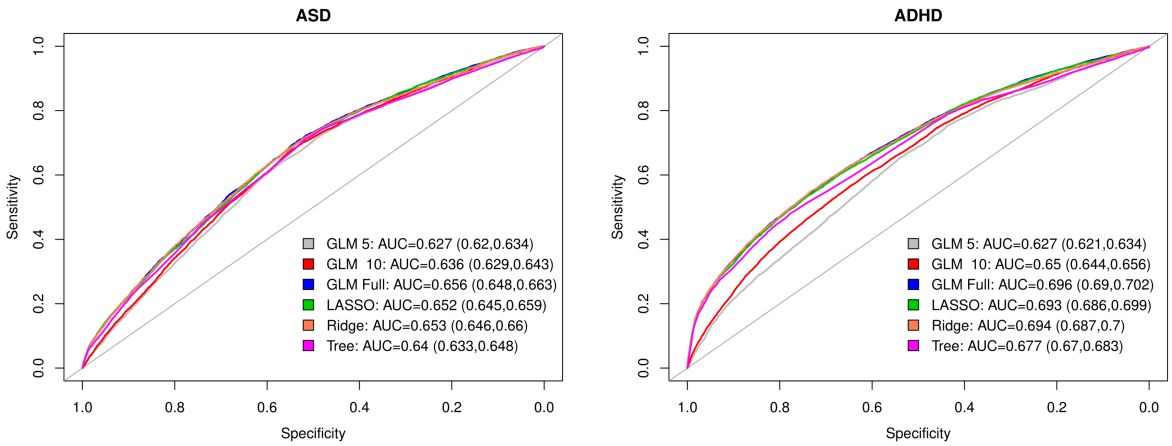


Figure C.6: 10-fold cross-validated ROC curves for six different models predicting ASD and ADHD, respectively, in the data set **FullFollowUp9799Imputed** with 220,458 subjects (ASD 2.5%, ADHD 3.5%).

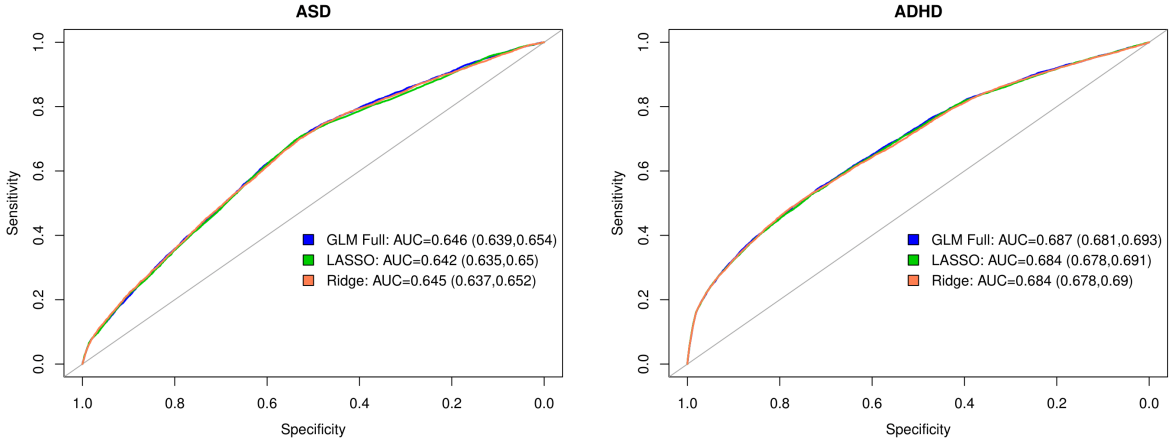


Figure C.7: 10-fold cross-validated ROC curves of three different models on the data set **FullFollowUp9799Imputed** with 220,458 subjects (ASD 2.5%, ADHD 3.5%) with imputed missing values. Only selected variables and interaction terms