The New Comorbidity Index

A Development and Validation Study

Part I - Theory

Rikke Beck Nielsen
and
Sinna Pilgaard Ulrichsen

AALBORG UNIVERSITY

Dep. of Mathematical Sciences • Master's Thesis • 1. Sep. - 30. May 2010

TITLE:

The New Comorbidity Index
A Development and Validation Study

PROJECT PERIOD:

From 1. September 2009
To 30. May 2010

PROJECT GROUP:

Rikke Beck Nielsen
Sinna Pilgaard Ulrichsen

SUPERVISOR:

Poul Svante Eriksen

COPIES: 10

NUMBER OF PAGES: 95

AALBORG UNIVERSITY

Dep. of Mathematical Sciences • Master's Thesis • 1. Sep. - 30. May 2010

Abstract

When the prognosis of a disease is studied it has been shown that comorbid diseases have an influence on the outcome. To adjust for this influence a comorbidity index can be used. The most used index is the Charlson comorbidity index, which was developed in 1987 on a small cohort of patients from the medical service at New York Hospital. The aim of this thesis is to investigate the ability of the Charlson comorbidity index to predict mortality on a cohort of pneumonia patients, and to develop and validate a new comorbidity index.

We used a cohort of hospitalized pneumonia patients from the Danish National Registry of Patients. We validated the Charlson comorbidity index by including it in a logistic regression with 30 day mortality as an outcome and assessing the performance.

Both logistic regression, naive Bayes and classifications trees were used to develop new indexes. When using the logistic regression method we updated the weights on the original Charlson diseases, included three new diseases, included first degree interaction terms and a variable for 'time since diagnosis'.
The naive Bayes method and classification trees were used as alternatives to the logistic regression model. Indexes made by these methods included the original Charlson diseases and the three new diseases.

We validated the indexes by assessing their performance for both 30 day and 1 year mortality. Their crude performance was assessed by the Pearson χ^2 test of a contingency table for the index and mortality. To assess their adjusted performance we included each index in a logistic regression adjusted for sex and age.

Our analysis showed that the Charlson comorbidity index predicted death among pneumonia patients well, and therefore it is still usable. All of our developed indexes performed well and most of them better than CCI. Our analysis showed that four of our indexes were better than the rest. For these indexes their complexity increased with performance. Choosing the best index out of these is therefore a balance between performance and simplicity and depends on the situation at hand.

Resumé

Når prognosen af en sygdom studeres, er det vist, at komorbide sygdomme har en effekt på resultatet. For at justere for denne effekt kan et komorbiditets indeks bruges. Det indeks, der oftest anvendes, er Charlsons komorbiditetsindeks, som blev udviklet i 1987 på en lille kohorte af patienter fra New York Hospital.
Formålet med dette speciale er at undersøge Charlson komorbiditetsindeksets evne til at prædiktere dødeligheden i en kohorte af lungebetændelsespatienter, og at udvikle samt validere et nyt komorbiditetsindeks.

Vi anvender en kohorte af hospitalsindlagte lungebetændelsespatienter fra Landspatient Registeret. Vi validerer Charlson komorbiditets indekset ved at inkludere det i en logistisk regression med 30 dages dødelighed som udfald og dernæst vurdere dets præstation.

Både logistisk regression, naive Bayes og klassifikations træer blev brugt til at udvikle et nyt indeks. Da den logistiske regression blev brugt, opdaterede vi vægtene på Charlsons originale sygdomme, inkluderede vi tre nye sygdomme, inkluderede vi førstegrads-interaktionsled og en variable for 'tid siden diagnose'.
Som et alternativ til den logistiske regression blev naive Bayes og klassifikationstræer brugt. Indekserne udviklet på baggrund af disse metoder inkluderede de originale Charlson sygdomme samt de tre nye sygdomme.

Vi validerede indekserne ved at undersøge deres prædiktionsevne for både 30 dages- og 1 årsdødelighed. Deres rå prædiktionsevne blev vurderet ved hjælp af en Pearson χ^2 test på en frekvenstabel indeholdende indekset og dødeligheden. For at vurdere deres justeret prædiktionsevne blev hvert indeks inkluderet i en logistisk regression justeret for køn og alder.

Vores analyser viste, at Charlson komorbiditetsindekset er udmærket til at prædiktere død blandt lundebetændelsespatienter, og derfor er det stadig brugbart.
Alle vores indekser præsterede udemærket og de fleste af dem bedre end Charlson komorbiditetsindekset. Vores analyser viste, at fire af vores

indekser var bedre end resten. For disse indekser steg kompleksiteten med prædiktionsevnen. Det at vælge det bedste indeks er derfor en balance mellem prædiktionsevne og enkelhed og afhænger af den enkelte situation.

Preface

This master thesis is written by Rikke Beck Nielsen and Sinna Pilgaard Ulrichsen in the period from September 2009 to May 2010. The thesis is made at the Department of Mathematical Sciences, Aalborg University, in cooperation with the Department of Clinical Epidemiology, Århus University Hospital.
Basic knowledge corresponding to the bachelor degree at Mathematical Sciences at Aalborg University is required.

The thesis consists of two parts. The first part comprises the basic theory of generalized linear models with focus on logistic regression models and their applications, the naive Bayes method and the theory of classification trees. The second part provides a validation of the Charlson comorbidity index, a development of new indexes using logistic regression, naive Bayes and classification trees, and a validation of these. All the analyses are based on a cohort of pneumonia patients.

We inform the reader that two kinds of references are used. When the number is in brackets, e.g. (1.2) the reference is to an equation or line with this number, while 'table 1.2' refers to an entire table with the number. The last-mentioned reference also exists for figures, sections and chapters.
Source references are given when relevant in the beginning of sections and chapters or at the end of a sentence. The source reference [Author(s), publishing year] refers to the reference list on the last page.
In the appendix tables and figures with results not given directly in the report are listed.

We wishes to thank the Department of Clinical Epidemiology for providing data as well as computer equipment and the supervisor, Poul Svante Eriksen, for his supervision and patience. Furthermore, we want to thank all personal at Department of Clinical Epidemiology, especially Mette Nørgaard and Malene Cramer Engebjerg for their guidance.

Aalborg the 30. of May 2010

Rikke Beck Nielsen Sinna Pilgaard Ulrichsen

Table of contents

Introduction

When the prognosis of a disease is studied it has been shown that comorbid diseases have an influence on the outcome. Such mixing of effects may be handled in several ways. One method is to apply a restricted criteria to exclude patients who have comorbid diseases. This method increases the certainty that any observed effect is caused by the disease of interest, and not by the confounding influence of comorbid diseases. However, studies that focus on the prognosis of a disease among patients with no comorbid diseases can not be generalized. Alternative methods such as matching, stratification and adjustment all require that the amount of comorbid diseases can be measured. Such a measure can be a comorbidity index, which can be used to e.g. adjust for the confounding effect of comorbid diseases. [Charlson, 1987],[Extermann, 2000]

There are several different indexes to choose from when wanting to adjust for comorbidity in a study. Before preceding to the problem definition a short review on the four most commonly used indexes is given.[Extermann, 2000]

Charlson Comorbidity index (CCI)

The Charlson comorbidity index was developed by Mary Charlson and colleagues in 1987. They used data from a medical service at New York hospital to analyze the 1 year mortality as a function of different comorbidities. For each disease a relative risk for death was calculated and those with a relative risk for death ≥ 1.2 were retained. This analysis resulted in a list of 19 diseases where each disease was given a weight. If the relative risk was ≥ 1.2 and < 1.5 the weight was 1, if ≥ 1.5 and < 2.5 the weight was 2, if ≥ 2.5 and < 3.5 the weight was 3 and the two diseases with a relative risk of 6 or more were given the weight 6. The index was calculated by adding the weights of those of the 19 diseases the patient had. The index was validated on breast cancer patients with 10 year mortality as endpoint. The index can then e.g. be collapsed into four categories; 0, 1-2, 3-4 and ≥ 5. [Charlson, 1987], [Extermann, 2000]

The Cumulative Illness Rating Scale (CIRS)

The CIRS was designed by Linn and colleagues in 1968. The aim was to record all the comorbid diseases of a patient. It classifies comorbidities according to the organ system affected and rates them according to their severity from 0 to 4. The CIRS has 14 organ system subdivisions. If two diseases are present in an organ system, the disease with the highest severity is used. The scale can then be summarized as a total number of categories used, total score, mean score or number of diseases with a grade of 3 or 4. [Extermann, 2000]

The index of Coexistent Disease (ICED)

The ICED was developed in 1987 by Greenfield and colleagues to address issues of intensity of care. The ICED consists of two subscales, a physical and a functional. The physical scale rates the diseases from 0 to 4 according to severity, and then regroups them in 14 disease categories. The functional scale has 12 categories of functional impairment, and each impairment is rated from 0 to 2. The scales are each summarized by the highest score and they are then lumped together according to a grouping system, to form an overall score ranging from 0 to 3. The grouping system can be seen in table A.4 in the appendix. [Extermann, 2000]

The Kaplan-Feinstein index

The Kaplan-Feinstein index was developed in 1974 by these two authors. The index consists of some diseases "that might be expected to impair a patient's long-term survival". The diseases are grouped in 12 categories and then rated from 0 to 3 according to severity. The number and severity of the diseases are evaluated and the overall comorbidity score ranging from 0 to 3 is calculated. The overall score is the grade of the disease with the highest score, but if two or more diseases have the grade 2 the overall grade is then 3. [Extermann, 2000]

Problem definition

The fact that the Charlson comorbidity index was based only on presence of diseases and not on severity of diseases or functional impairment makes it very easy to use. The index is actually the most simple and most used index [Extermann, 2000]. Because of this, it is the index we choose as a starting point.
The CCI has however some problems. Firstly, the index was developed in 1987 and since the treatment of the different diseases has become much more efficient, the weights appointed to the 19 diseases should probably be changed. Secondly, the index would probably not contain the same 19 diseases if it was made today since the improved treatments results in patients dying from other diseases. Thirdly, the dataset used by Mary Charlson only consisted of 604 patients so the dataset was relatively small. Some comorbidities had a low prevalence and occured only a few times if any at all in the dataset. This resulted in poor relative risk estimation. Since all the patients were from the New York Hospital, there could be bias so the results might not be generalizable to the rest of the USA or the world. The patients in the study were collected during 1 month, which may have resulted in bias since the seasonal variation in disease prevalence was not taken into account. In addition to these issues there might be a time effect to consider. As an example a cancer diagnosis may not have the same influence on the mortality 10 years after the diagnosis compared to 10 days after the diagnosis.

We wish to keep the simplicity of the CCI, but at the same time develop it to handle some of the above problems. To handle the first two problems we could, with a representative dataset, do as Mary Charlson and use the relative risks to find an updated list of diseases. However this may result in computational difficulties, since all the diseases recorded would be included in the model. Another approach could be to consult a clinician for a list of additional diseases and then calculate new weights for both the additional and the original diseases (some of the diseases might be given the weight zero).

To address the issue of seasonal variation and the issue of generalization subjects are found in the Danish National Registry of Patients (DNRP). The DNRP is used because it is a nation based registry

containing 99.5% of all hospitalizations [Andersen et al., 1999], which mean that the generalization to Denmark is apparent, but generalization to other countries has to be validated. The seasonal variation can be expected to even out if the study period is a number of whole years. Using the DNRP as a basis of the dataset will thereby make the dataset representative and minimize a potential bias. To deal with the time effect of the comorbidity diagnoses mentioned above, some diseases could be subdivided into different groups depending on the 'age' of the diagnosis.

Aim of the thesis

The objective of this study is to investigate the ability of the Charlson comorbidity index to predict mortality on a cohort of pneumonia patients and on the basis of this investigation to develop and validate an updated comorbidity index.

We start part I by introducing the general concepts of study design and how to choose a statical model. Next basic theory on logistic regression, naive Bayes and classification trees, with focus on application is provided.

We start part II by investigating how well the CCI predicts the mortality on a cohort of pneumonia patients. Next we modify CCI by updating the weights in the index and by including three new diseases. We gradually increase the complexity of the index by including first degree interaction terms and the 'age' of the diagnosis. Finally we validate the new indexes and see if any improvements have been made.

The choice of study population

We choose hospitalized pneumonia patients as our group of interest, since the prevalence of pneumonia is increasing and because pneumonia is the most frequent cause of death in Danish hospitals [Christensen et al., 2007]. Furthermore the well defined medical group consisting of hospitalized pneumonia patients is known to have a high mortality and a high number of comorbidities [Thomsen et al., 2006]. When the mortality and prevalence of the comorbidities are high the

statistical results have more power. Choosing the homogeneous and general internal medical group of hospitalized pneumonia patients gives the results medical credibility, and means that the admission diagnosis does not need to be taken into account. The pneumonia discharge diagnoses are validated and has a positive predictive value of 90%, which is high compared to other diagnoses [Thomsen et al., 2006].

Part I

Theory

Study design

In this chapter an introduction to the main observational study designs used in clinical epidemiologic research is given. The chapter is written on the basis of [Hulley et al., 2001].
In an observational study design the study population is only observed and values of variables are registered, where in experimental studies part of the predictor variables are modified and controlled by the investigator.
Observational studies are especially useful in situations, where modifications of variables will cause ethical problems or simply be not possible.

1.1 Cohort studies

A cohort study is an observational study, where subjects are followed over time in order to describe the incidence of a condition and to analyze predictors (risk factors) for a chosen outcome. Baseline is the time where subjects enter the study and follow-up is the time where the outcome is registered. The strength of a cohort design can be compromised by incomplete follow-up.

Prospective cohort study
In prospective studies, illustrated in figure 1.1, the investigator defines the sample of subjects, and measures the predictor variables, before any outcomes have occurred. The main strength of a prospective study design is, that it is possible to measure predictor variables thoroughly without risking, that they are biased by outcomes or by poor memories (e.g. it can be difficult to remember what you ate a month ago). This study design is however expensive and inefficient when outcomes are rare in a population.

Retrospective cohort study
In retrospective studies, illustrated in figure 1.1, the investigator defines the sample of subjects and measures the predictor variables after the outcomes have occurred. This study design requires that measurements of the predictor variables are available for a cohort of subjects. Typically these data have been gathered for other purposes. The strength is that like in prospective studies the measurements of the predictor

variables are not biased by the outcome. The retrospective study is less time consuming and expensive than the prospective study, since measurements already are made and time to outcome already has past. However the investigator has limited control over study design, study population, predictor variables, measure methods and so on, which may result in incomplete and inaccurate key variables.

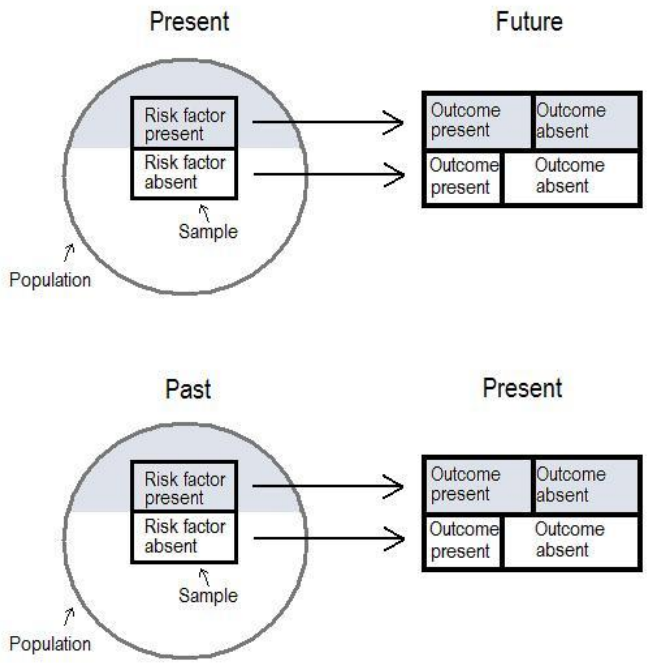

Figure 1.1 *The upper figure illustrates the prospective study design and lower figure the retrospective study design.*

1.2 Cross-sectional study

A cross-sectional study, illustrated in figure 1.2, is an observational study, where all variables are measured at the same time. On the basis of subject matter, prior knowledge and of the distributions of the variables, the investigator then chooses the outcome and predictors among the measured variables. The advantages of this study design

compared to cohort studies are time efficiency, and the fact that there is no problem with subjects being lost to follow-up. It may reveal unknown associations and can be implemented as an extra study in a cohort setting when waiting for the follow-up.

An important descriptive statistic in cross-sectional designs is the prevalence. The prevalence is given by the number of exposed subjects over the number of subjects in the sample. This should not be mistaken for the incidence used in cohort studies given by the number of subjects getting an exposure over the number of subjects in the sample. A weakness of the cross-sectional study is that only the prevalence can be found and not the incidence. Prevalence is a mix of both incidence and duration of the disease. To show causation the investigator needs to show a difference in incidence for the different predictor levels, so causation can not be assessed by a cross-sectional study, only association.

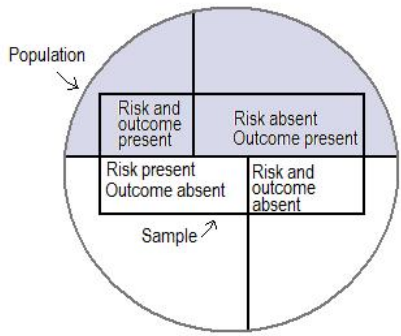

Figure 1.2 *The cross-sectional study design.*

1.3 Case-control study

A case-control study, seen in figure 1.3, is an observational study, where the investigator chooses cases from a population with presence of an outcome and controls from a population with absence of the outcome. The levels of the predictor variables are then compared for the cases and controls.

Case-control studies do not provide prevalence or incidence for the outcome, it provides odds ratios, which approximates relative risks when outcomes are rare.
The main advantage of case-control studies is that only a small number of subjects, compared to cohort studies, is required to obtain the same strength of the study. This is an advantage both when the outcome is rare and when the waiting time for the outcome is long. The main disadvantage of this study design is, that there is a large risk of bias, both from the separate sampling of the cases and controls and from the retrospective measurements of the predictors.

When sampling the cases, where the outcome is a disease, the sampling may not be representative, since misdiagnosed, undiagnosed and dead subjects are not available. To minimize this bias, only well defined and representative outcomes should be used.

Matching ensures that cases and controls are comparable with respect to characteristics, that might be related to the outcome, but of no interest to the investigator.

In general when population based registries are available these should be used to sample controls. This makes the study nested within a cohort and can minimize sampling bias. To avoid measurement bias blinding both the investigator and the patients should be done when possible.

Nested case-control study
A nested case-control study, seen in figure 1.3, is a case-control study nested within a prospective or retrospective cohort study. The cases are all the subjects in the cohort with present outcome, and the controls are sampled among the subjects in the cohort with absent outcome. If the subjects are followed for different lengths of time, it may be a god idea to match the cases and controls by the length of follow-up. In some situations matching on other characteristics such as sex and age might improve the model.

Nested case-cohort study
A nested case-cohort study is the same as a nested case-control study except that the controls are selected in the entire cohort and not only among those who did not develop the outcome. This has the advantage,

that a set of randomly selected controls can be used in several studies with the same cohort and different outcomes. The random sample also contains information on the overall prevalence of the risk factors.

In nested case-control and case-cohort studies the predictor variables are not needed for the entire cohort, but only the cases and controls. This is an advantages if some predictor variables which already are stored are difficult or expensive to assess.
In general all the cohort studies share the advantage that predictor variables are not biased by knowledge of the outcome, and the disadvantage that the observed associations can be caused by confounders.

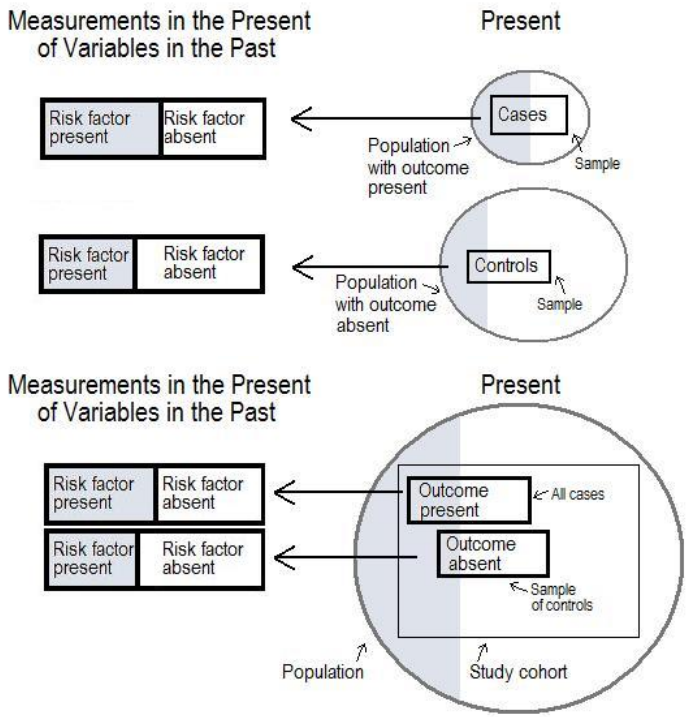

Figure 1.3 *The upper figure illustrates the case-control study design and the lower the nested case-control study design.*

1.4 Interpretation of associations

When an association between an outcome and a predictor is found, the question is if it represents a cause-effect (that the predictor caused the outcome). The main explanations of associations are cause-effect, chance (that it is due to a random error), bias (that it is due to a systematic error), effect-cause (that the outcome caused the predictor) and confounding. To minimize the possibility that the association is explained by chance, sufficient sample size and precision is needed and p-values should be assessed. Bias is minimized by choosing study design and research question carefully. Effect-cause is made less likely by carefully considering biological plausibility. Confounding happens when some third factor causes both the outcome and the predictor. It can be minimized by sampling data within a cohort with the same levels of the potential confounder, by adjusting for this third factor as described in section 2.1.5, by matching or by stratifying the subjects in the analysis according to each level of the potential confounder. One of the great advantages of adjustment is that it is possible to adjust for several potential confounders at the same time. On the other hand adjusting for too many confounders may result in problems with the statistical models used in the analysis phase.
So when evaluating whether a found association in fact represents cause-effect, the four alternative explanations must be considered and ruled out.

1.5 Secondary data analysis

Secondary data is data gathered for a different purpose than the purpose of the analysis. Using secondary data analysis reduces time and cost in research. This leaves the investigator with limited control over data, meaning that the investigator must settle with the variables at hand, their quality and the way they are recorded. Investigators can find a dataset or a database that is useful for an existing research question or the other way around. Administrative and clinical databases and registries are very useful, especially when studying rare outcomes and assessing use and effectiveness of e.g. a medical drug or a treatment, that has been shown to work in an experimental setting.

Logistic regression

In this chapter generalized linear regression with an application to logistic regression is introduced. Furthermore the concept of hypothesis testing, goodness-of-fit statistics and validation is also given.

In statistical data analysis, regression models are the main approach, when describing the relationship between a response variable and explanatory variables. The most commonly used model is the linear regression model, but when the response variable is discrete that model is inadequate. With the principles from the linear model, a family of models called generalized linear models are employed. What distinguishes the generalized linear model and the ordinary linear model is the choice of the parametric model and the model assumptions. When dealing with a binary response logistic regression, which belongs to the generalized linear models, is the standard method of analysis. [David W. Hosmer, 2000]
Often when studying dichotomous outcomes contingency tables are used. In a case with a continuous explanatory variable, this method requires that the values of the continuous variable are divided in intervals, compromising the information contained in the variable. Using a statistical model like the logistic regression model this problem is solved, since the association between the dichotomous outcome and the continuous variable can be modeled and evaluated directly. [Frank E. Harrell, 2001]

2.1 Generalized linear models

This section is written on the basis of [Azzalini, 2002] and [Dobson, 1990] and contains a generalization of the ordinary linear regression model. The insufficiencies of the linear model are summarized in the following three items:

- The relationship between the response variable and the explanatory variables is not necessarily linear.
- The variance of the response variable is not necessarily constant.
- The response variable may not be normally distributed.

Before making the generalization we introduce a family of distributions called the exponential family, which have a number of properties in common with the normal distribution. It is defined by the following.
The distribution of the stochastic variable Y is said to belong to the exponential family if the probability density function can be written in the form

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

where θ is a vector of distribution parameters, $a(\cdot)$, $b(\cdot)$ are known vector functions and $c(\cdot)$, $d(\cdot)$ are known functions. When using the exponential family later, θ, $a(\cdot)$ and $b(\cdot)$ are one dimensional, so they are from this point regarded as such.
If $a(y) = y$ the distribution is said to be in canonical form. The term $b(\theta)$ is called the natural parameter of the distribution. In table 2.1 a list of examples of natural parameters can be seen.

Distribution	**Natural parameter**
Poisson (θ)	$\log(\theta)$
Normal (μ,σ)	$\frac{\mu}{\sigma^2}$
Binomial (π)	$\log\left(\frac{\pi}{1-\pi}\right)$

Table 2.1 *Natural parameters for the poisson, normal and binomial distributions. Note that σ is a known parameter. [Dobson, 1990]*

The mean and variance of $a(Y)$ are now found for later use. Note that $1 = \int f(y|\theta)dy$, where $f(\cdot|\theta)$ is the probability density function. Differentiating with respect to θ and reversing the integration and differentiation gives

$$\begin{aligned}
0 =& \frac{d}{d\theta}\int f(y|\theta)dy \\
=& \int \frac{d}{d\theta} f(y|\theta)dy \\
=& \int [a(y)b'(\theta) + c'(\theta)]f(y|\theta)dy \\
=& b'(\theta)\mathbf{E}[a(Y)] + c'(\theta).
\end{aligned}$$

The last equality follows by $\int a(y)f(y|\theta)dy = \mathbf{E}[a(Y)]$ from the definition of the expected value. Rearranging the above gives

$$\mathbf{E}[a(Y)] = \frac{-c'(\theta)}{b'(\theta)}. \tag{2.1}$$

Differentiating twice with respect to θ and reversing the integration and differentiation gives

$$\begin{aligned}
0 =& \frac{d^2}{d\theta^2}\int f(y|\theta)dy \\
=& \int \frac{d^2}{d\theta^2} f(y|\theta)dy \\
=& \int \frac{d}{d\theta}[a(y)b'(\theta) + c'(\theta)]f(y|\theta)dy \\
=& \int [a(y)b''(\theta) + c''(\theta)]f(y|\theta) + [a(y)b'(\theta) + c'(\theta)]^2 f(y|\theta)dy \\
=& \int [a(y)b''(\theta) + c''(\theta)]f(y|\theta) + [b'(\theta)]^2\{a(y) - \mathbf{E}[a(Y)]\}^2 f(y|\theta)dy \\
=& b''(\theta)\mathbf{E}[a(Y)] + c''(\theta) + [b'(\theta)]^2 \mathrm{Var}[a(Y)]
\end{aligned}$$

Since $\int \{a(y) - \mathbf{E}[a(Y)]\}^2 f(y|\theta)dy = \mathrm{Var}[a(Y)]$ by definition. Using equation (2.1) and rearranging gives

$$\mathrm{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{b'(\theta)^3}. \tag{2.2}$$

The scoring vector is defined by $U(\theta|y) = \frac{d\ell(\theta|y)}{d\theta}$, where $\ell(\theta|y)$ is the log likelihood function. The likelihood function is proportional to the probability density function, so for an exponential family this becomes

$$U(\theta|y) = a(y)b'(\theta) + c'(\theta).$$

Given y, it can be seen as a random variable. The Information function defined by $\mathfrak{I} = -\mathbf{E}[dU/d\theta]$ is

$$\begin{aligned}
\mathfrak{I} =& -\mathbf{E}[a(Y)b''(\theta) + c''(\theta)] \\
=& \frac{c'(\theta)}{b'(\theta)} b''(\theta) - c''(\theta).
\end{aligned}$$

This is actually the $\mathrm{Var}(U)$ since

$$\mathrm{Var}(U) = \mathrm{Var}[a(Y)b'(\theta) + c'(\theta)] = \mathrm{Var}[a(Y)]b'(\theta)^2 = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta).$$

Hence

$$\mathfrak{I} = \mathrm{Var}(U) = \mathbf{E}[U^2] - \mathbf{E}[U]^2 = \mathbf{E}[U^2], \tag{2.3}$$

since $\mathbf{E}[U] = \mathbf{E}[a(Y)]b'(\theta) + c'(\theta) = -[c'(\theta)/b'(\theta)]b'(\theta) + c'(\theta) = 0$.

The generalization of the linear model

The characteristics of the linear model is first identified and then developed in two ways.
Given the linear model $Y = X\beta + \varepsilon$, the i'th observation of Y is defined by the linear predictor $\eta_i = x_i^\top \beta$, where x_i is the i'th row in X. Assume that the observations Y_i are drawn independently from $Y_i \sim N(\mu_i, \sigma^2)$, where $\mu_i = \eta_i$. This can be summarized to:

- Y_i's are mutually independent
- $Y_i \sim N(\mu_i, \sigma^2)$
- $\mu_i = \eta_i$
- $\eta_i = x_i^\top \beta$

The generalized linear models are now achieved by allowing the following two expansions:

- The distribution of the Y_i is not restricted to the normal distribution, but can have any distribution belonging to a given exponential family.
- The relationship between η_i and μ_i is not restricted to the identity, so
$$g(\mu_i) = \eta_i,$$
where $g(\cdot)$ called the link function is a differentiable and monotonous function.

This means that a generalized linear model is characterized by

$$g(\mu_i) = \eta_i, \quad \eta_i = x_i^\top \beta \tag{2.4}$$

and the distribution of the mutually independent Y_i's, which belongs to an exponential family.

A statistical model is a generalized linear model, when it satisfies these three statements:

- The observations $y_1, \ldots, y_n$ are realizations of mutually independent stochastic variables $Y_1, \ldots, Y_n$, where the distributions of the Y_i's belong to an exponential family.
- The function $g(\cdot)$ exists, so that $g(\mu_i) = x_i^\top \beta$, where β is a vector of parameters.
- The functions $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ and $d(\cdot)$ are known, and the distributions of Y_i have the same shape (either normal, binomial etc.).

2.1.1 Choosing the model

In this section a few guidelines as how to choose a model is given. This section is based on [Frank E. Harrell, 2001].

In biostatistics, epidemiology, economics and many other fields it is seldom that prior knowledge on the subject exists so that the analyst can prespecify a model, a transformation for the response variable, and a structure for how predictors appear in the model (e.g., transformations, addition of nonlinear terms, interaction terms). The analyst is therefore often forced to develop models empirically. Fortunately, a careful and objective validation of the accuracy of model predictions against observed responses can make the model more trustworthy, if a good validation is not merely the result of overfitting.

There are a few guidelines that can help in choosing the basic form of the statistical model.

1. The model must use the data efficiently. If, for example, the prediction of the probability that a patient with a specific set of

characteristics would live five years from diagnosis is of interest, an inefficient model would be a binary logistic model. A more efficient method would be a parametric survival model. Such a model uses individual times of events in estimating coefficients, but it can easily be used to estimate the probability of surviving five years.

2. Choose a model that fits overall structures which are likely to be present in the data. In modeling survival time in chronic diseases it might be important that most of the risk factors are constant over time. In that case, a proportional hazards model such as the Cox model would be a good initial choice.

3. Choose a model that is robust to problems in the data that are difficult to check. For instance, the Cox proportional hazards model and ordinal logistic regression models are not affected by monotonic transformations of the response variable.

4. Choose a model whose mathematical form is appropriate for the response being modeled. This has to do with minimizing the need for interaction terms that are included only to address a basic lack of fit. For example when an ordinary linear regression model is used for a binary response. Such a model allow predicted probabilities outside the interval $[0, 1]$, and therefore strange interactions among the predictor variables are needed to make predictions remain in the interval.

5. Choose a model that easily can be extended. The Cox model, by its use of stratification, easily allows a few of the predictors to violate the assumption of equal regression coefficients over time, that is the proportional hazards assumption.

2.1.2 Binary responses

This section is written on the basis of [Dobson, 1990] and contains the derivation of the likelihood and link function in logistic regression.

The special case where the response is binary often appears in biostatistics. As an example one may want to investigate how of a medical treatment affects the mortality among patients. In this case, the medical treatment is the explanatory variable and the response variable is

binary with the possible outcomes 'dead' or 'alive'. A general binary response variable is defined by the following, where the interpretation of this response depends entirely on the situation analyzed.

$$Z = \begin{cases} 1 & \text{if the outcome is success,} \\ 0 & \text{if the outcome is failure,} \end{cases}$$

where $P(Z = 1) = \pi$ and $P(Z = 0) = 1 - \pi$, is the familiar Bernoulli distribution $B(\pi)$ with $\mathbf{E}[Z] = \pi$. Given n independent stochastic variables $Z_1, \ldots, Z_n$ with probability $P(Z_j = 1) = \pi_j$ the joint probability density function is

$$\begin{aligned} f(z_1, \ldots, z_n | \pi_1, \ldots, \pi_n) &= \prod_{j=1}^{n} f(z_j | \pi_j) \\ &= \prod_{j=1}^{n} \pi_j^{z_j} (1 - \pi_j)^{1 - z_j} \\ &= \exp\left[\sum_{j=1}^{n} z_j \log\left(\frac{\pi_j}{1 - \pi_j} \right) + \sum_{j=1}^{n} \log(1 - \pi_j) \right]. \end{aligned}$$

This belongs to the exponential family, where $a(z_j) = z_j$, $b(\pi_j) = \log(\pi_j/(1 - \pi_j))$, $c(\pi_j) = \log(1 - \pi_j)$ and $d(z_j) = 0$ for each of the observations.
In the case, where all $\pi_j = \pi$,

$$Y = \sum_{j=1}^{n} Z_j$$

is the number of successes, which is binomially distributed,
$Y \sim \text{Bin}(n, \pi)$.
This means, that for N stochastic variables, $Y_1, \ldots, Y_N$ given by the number of successes in N different subgroups, where $Y_i \sim \text{Bin}(n_i, \pi_i)$,

the log likelihood function is

$$\begin{aligned}\ell(\pi|y) &= \log\left(\prod_{i=1}^{N} f(y_i|\pi_i)\right) \\ &= \sum_{i=1}^{N} \log\left[\binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}\right] \\ &= \sum_{i=1}^{N}\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i\log(1-\pi_i) + \log\binom{n_i}{y_i}\right]. \qquad (2.5)\end{aligned}$$

The probability density function for the stochastic variables $Y_1, \ldots, Y_N$ belongs to the exponential family where, $a(y_i) = y_i$, $b(\pi_i) = \log(\pi_i/(1-\pi_i))$, $c(\pi_i) = n_i\log(1-\pi_i)$ and $d(y_i) = \log\binom{n_i}{y_i}$. To describe the behavior of the binary response, the probability for success, $\pi_i = Y_i/n_i$ is considered. Since $E[Y_i] = n_i\pi_i$ and thereby $E[Y_i/n_i] = \pi_i$, the probabilities, π_i are modeled by the generalized linear model

$$g(\pi_i) = x_i^\top \beta.$$

Logistic regression

As seen in line (2.5) the natural parameter for the binomial distribution is $b(\pi_i) = \text{logit}(\pi_i)$. When the link function results in the natural parameter, it is called a canonical link function. In logistic regression the canonical link function is used, so the logistic regression model becomes

$$g(\pi_i) = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^\top\beta. \qquad (2.6)$$

Written in another way this gives the regression model

$$\pi_i = \frac{\exp(x_i^\top\beta)}{1+\exp(x_i^\top\beta)}. \qquad (2.7)$$

The logistic regression has a very suitable property, that is, it models the π_i's in the range between 0 and 1.

Other statistical methods

Another statistical model, with similar properties is obtained via the probit link function. This model is based on the cumulative normal distribution. It has the same shape as the logistic model, but the link function involving the inverse of the cumulative normal distribution makes the computations quite heavy. This link function is not the natural parameter, so using the iterative weighted least squares procedure, described in the following section, does not guaranty that the right parameter estimates are found as described in section 2.1.3. [Frank E. Harrell, 2001]

Discriminant analysis is a statistical tool, which is computationally easier than logistic regression [David W. Hosmer, 2000]. Discriminant analysis actually have the same assumptions about the model as logistic regression plus the additional assumptions, that the joint distribution of the explanatory variables is multivariate normal. These additional assumptions are unlikely to be met in practice, especially if one of the explanatory variables is discrete. The assumptions come from the fact, that the model in discriminant analysis is based on the distribution of $X|Y$ and has to be inverted using Bayes' rule to derive $P(Y)$. The logistic regression model on the other hand is based on $P(Y|X)$ directly. The distribution of a binary random variable is completely defined by the true probability that $Y = 1$, so no assumptions about X is made in the logistic regression model. If the assumptions in discriminant analysis are violated the logistic regression yields a better model [Press and Wilson, 1978], [Halperin et al., 1971] and if not, the logistic regression is just as good [Frank E. Harrell and Lee, 1985].

2.1.3 Estimating model parameters in generalized linear models

This section is written on the basis of [Dobson, 1990] and contains a mathematical derivation of the procedure used when estimating the parameters in a generalized linear model with the maximum likelihood method.
When fitting the ordinary linear model the method of least squares is used. Using the method of least squares, the parameters, which minimize the distance between the observed response and the model predicted values, is chosen. When the error terms are normally distributed,

this method is equivalent to the maximum likelihood method. The method of least squares is not suitable for the generalized linear model, but the maximum likelihood is a convenient replacement. Since the assumption of normal distributed error terms does not apply for the generalized linear model, the maximum likelihood method results in a more complex equation. The idea in maximum likelihood estimation is to find the parameters, which make the observed data most probable. That is, to maximize the likelihood function with respect to the parameters or equivalently the log likelihood function. Using the maximum likelihood method to fit the generalized linear model can be done with an iterative weighted least squares procedure. This procedure is used in most software packages and is described in the following.

The method of maximum likelihood applied on generalized linear models

Except in special cases, where all the explanatory variables are discrete, the maximum likelihood problem can not be solved directly. The fastest and most applicable method for solving a function iteratively is generally the Newton-Raphson method, that approximates the given function with a linear function in a small region. [Frank E. Harrell, 2001]

We now derive the so called iterative weighted least squares procedure, to determine the maximum likelihood estimates in a generalized linear model. The derivation is for the generalized linear model with a canonical likelihood function, so $a(Y_i) = Y_i$.

Given N independent stochastic variables $Y_1, \ldots, Y_N$, which satisfies the assumptions for the generalized linear regression, the joint log likelihood function can be written in the form

$$\ell(\beta|y) = \sum_{i=1}^{N} \ell_i(\theta_i|y_i) = \sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i),$$

since it belongs to the exponential family and is on the canonical form. Remember from line (2.1), (2.2) and (2.4) that

$$\mathbf{E}[Y_i] = \mu_i = -c'(\theta_i)/b'(\theta_i) \tag{2.8}$$

$$\mathrm{Var}[Y_i] = [b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)]/[b'(\theta_i)]^3 \tag{2.9}$$

$$g(\mu_i) = \eta_i = x_i^\top \beta, \tag{2.10}$$

where x_i is the i'th row in X. These equalities also gives the connection between β and θ.
The maximum likelihood estimate can be found by solving the score equation

$$0 = \frac{\partial \ell}{\partial \beta_j} = U_j = \sum_{i=1}^{N} \left[\frac{\partial \ell_i}{\partial \beta_j} \right] = \sum_{i=1}^{N} \left[\frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \right]. \tag{2.11}$$

Each of the factors on the right side in equation (2.11) are treated separately. The first factor is by line (2.8)

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

By differentiating line (2.8) and using equation (2.9) the second factor is

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} = \left(\frac{-c''(\theta_i)b'(\theta_i) + c'(\theta_i)b''(\theta_i)}{[b'(\theta_i)]^2} \right)^{-1} = \left(b'(\theta_i)\mathrm{Var}(Y_i) \right)^{-1}.$$

By line (2.10) the last factor can be written as

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

Hence the score equation is

$$U_j = \sum_{i=1}^{N} \left[\frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right]. \tag{2.12}$$

The Information matrix from line (2.3) then becomes

$$\begin{aligned} \mathfrak{I}_{jk} &= \mathbf{E}[U_j U_k] \\ &= \mathbf{E} \left\{ \sum_{i=1}^{N} \left[\frac{(Y_i - \mu_i)}{\mathrm{Var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \sum_{l=1}^{N} \left[\frac{(Y_l - \mu_l)}{\mathrm{Var}(Y_l)} x_{lk} \left(\frac{\partial \mu_l}{\partial \eta_l} \right) \right] \right\} \\ &= \sum_{i=1}^{N} \frac{\mathbf{E}[(Y_i - \mu_i)^2]}{[\mathrm{Var}(Y_i)]^2} x_{ij} x_{ik} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2, \end{aligned}$$

as $\mathbf{E}[(Y_i - \mu_i)(Y_l - \mu_l)] = \mathbf{E}[Y_i - \mu_i]\mathbf{E}[Y_l - \mu_l] = 0$ for $i \neq l$, because of the independence of the Y_i's.

Since $\mathbf{E}[(Y_i - \mu_i)^2] = \mathrm{Var}(Y_i)$, $\mathfrak{I}_{jk}$ can be reduced to

$$\mathfrak{I}_{jk} = \sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \tag{2.13}$$

from where it can be seen, that

$$\mathfrak{I} = X^\top W X,$$

where W is an $N \times N$ diagonal matrix with the elements

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

Estimating the parameters, β is then done with a modified version of the Newton-Raphson method for maximization by

$$b^{(m)} = b^{(m-1)} + \left[\mathfrak{I}^{(m-1)}\right]^{-1} U^{(m-1)}, \tag{2.14}$$

where $b^{(m)}$ is the estimation of β at the m'th iteration, and the information matrix $\left[\mathfrak{I}^{(m-1)}\right]^{-1}$ and the scoring vector $U^{(m-1)}$ both are computed with $b^{(m-1)}$. This is actually quite similar to the equations used in linear regression, which can be seen by the following derivation. Written in a different way equation (2.14) is

$$\mathfrak{I}^{(m-1)} b^{(m)} = \mathfrak{I}^{(m-1)} b^{(m-1)} + U^{(m-1)}. \tag{2.15}$$

The expression on the right side of equation (2.15) is the vector with the elements

$$\sum_{k=0}^{p} \sum_{i=1}^{N} \left(\frac{x_{ij}x_{ik}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 b_k^{(m-1)} \right) + \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right),$$

by equation (2.13) and (2.12), where $p+1$ is the number of parameters. This means that the expression on the right side of equation (2.15) is

$$X^\top W z,$$

where z has the elements

$$z_i = \sum_{k=0}^{p} x_{ik} b_k^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i}\right), \tag{2.16}$$

with μ_i and $\partial\eta_i/\partial\mu_i$ computed with $b^{(m-1)}$.
All in all equation (2.15) is equivalent to

$$X^\top W X b^{(m)} = X^\top W z. \tag{2.17}$$

This equation is similar to the weighted normal equations $X^\top W X b = X^\top W y$ used in ordinary weighted linear regression for parameter estimation. Ordinary weighted linear regression is similar to ordinary linear regression except from that fact, that each observation has a weight. Most statistical programs use this iterative weighted least squares procedure, when estimating parameters in a generalized linear regression model.

The procedure starts with an initial value, $b^{(0)}$, used to compute z and W. With these and equation (2.17), $b^{(1)}$ is computed. Then new z and W is computed and so on. If the log likelihood gets worse at $b^{(i+1)}$ than at $b^{(i)}$, $b^{(i+1)}$ is replaced by $(b^{(i)} + b^{(i+1)})/2$. If this does not help, $b^{(i+1)}$ is replaced by $(3b^{(i)} + b^{(i+1)})/4$, and the normal procedure is resumed. [Frank E. Harrell, 2001]

When the -2 log likelihood function changes with less than some threshold, k, the procedure is stopped and the maximum likelihood estimate is found. Often $k = 0.05$ is chosen because a change in the parameter values of this size, does not affect the statistical conclusions since the -2 log likelihood function is χ^2 distributed. [Frank E. Harrell, 2001]

Because the probability density functions belongs to an exponential family, the canonical link function yields a concave likelihood function. With a concave likelihood the maximum found by the procedure is a global maximum. Hence the choice of $b^{(0)}$ only have little influence on the result of the procedure, if the procedure has converged. To make sure the procedure has converged keep an eye on the number of iterations, a rule of thumb is that under 10 iterations is fine [Green, 1984].

Considering the Taylor expansion of the link function $g(y_i)$ at μ_i

$$\begin{aligned} g(y_i) &\approx g(\mu_i) + (y_i - \mu_i) g'(\mu_i) \\ &= \eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} \\ &= z_i, \end{aligned}$$

it is seen that z_i is a local approximation to $g(y_i)$. For this reason the algorithm can be initiated by letting $z_i^{(0)} = g(y_i)$ and $W^{(0)}$ be the identity matrix. From these $b^{(1)}$ can be computed and the algorithm is started.
For some models the value of z_i can not be computed. An example is the logistic regression model, where $g(y_i) = \log \frac{y_i}{n_i - y_i}$ and one of the observations $y_i = 0$ or $y_i = n_i$. A solution to this can be to make a slight adjustment to the approximation by

$$z_i^{(0)} = \log \frac{y_i + 0.5}{n_i - y_i + 0.5}.$$

This approximation is called the empirical logit and reasons for the choosing 0.5 is given in [Cox and Snell, 1989].
The estimation and sample distributions are based on asymptotic results, so for small data sets the parameter estimates may result in a poor model. For logistic regression the case where every covariate pattern has few observations or where the frequencies of success are close to one or zero, the model may be poor.

Estimating the variance of the parameters

This section is written on the basis of [Frank E. Harrell, 2001].
The observed information, defined as

$$I(\beta) = \mathbf{E}\left[\frac{-\partial^2}{\partial\beta\ \partial\beta^\top}\ell(\beta)\right],$$

describes how much 'information' the observations contain about the parameters. Information in the sense, that the "bigger" the observed information is, the more distinct the peak of the log likelihood function is (the peak is at the maximum likelihood estimate). With a distinct peak of the log likelihood follows an estimate with a small variance and thereby high precision. Note that in general the more observations the "bigger" the observed information is and thereby the higher the precision.

Estimating the covariance matrix of the parameters can be done with the observed information by

$$V = I^{-1}(\hat{\beta}).$$

This estimation is consistent and it is assumed, that the model is specified correctly in terms of distribution, regression assumptions and independence assumptions.

2.1.4 Hypothesis tests

This section is written on the basis of [Frank E. Harrell, 2001] and contains a general introduction of likelihood based test statistics.

After fitting a model by estimating its unknown parameters, it should be investigated if the model can reveal anything about the response variable. This can be tested with a global test introduced in the next section. If the global test does not show, that the model is significant, interpretation of individual parameters and associations is risky.

The relevance of each of the parameters should be evaluated, by testing if the model tells more about the response with the parameter or without it. This is done by testing for the significance of parameters. [David W. Hosmer, 2000]
When testing the significance of a model or a subset of its parameters, the null hypothesis is H_0: $\beta = 0$, where β is the vector with the model parameters or the vector with the parameters of interest. The general null hypothesis H_0: $\beta = \beta^0$ is used in the next sections in the name of generality.
Remember that test conclusions are only concerning statistical evidence and factors concerning clinical importance should be evaluated separately.

Global test statistics

When testing the global hypothesis that the model parameters β are known, H_0: $\beta = \beta^0$ two test statistics arise from likelihood theory. For a large number of observations they are both χ^2 distributed under the null hypothesis with $p+1$ degrees of freedom, where $p+1$ is the number of entries in the vector of parameters β and $\hat{\beta}$ is the maximum likelihood estimate of β.

The likelihood ratio test statistic is given by

$$LR = -2[\ell(\beta^0) - \ell(\hat{\beta})].$$

This test statistic is based on the ratio of the likelihood of the hypothesis value and the likelihood of the maximum likelihood parameter estimates. Applying the log function and multiplying with -2 ensures an appropriate distribution.

The other statistic is the Wald test statistic

$$W = (\hat{\beta} - \beta^0)^\top V^{-1} (\hat{\beta} - \beta^0).$$

This test statistic is a generalization of the z statistic from the normal distribution. It is a function of the difference between the maximum likelihood parameter estimates and the hypothesis value, normalized by an estimate of the variance of the maximum likelihood parameter estimates.

A special likelihood ratio test statistic is

$$D = -2[\ell(\text{fitted model}) - \ell(\text{saturated model})].$$

Note that the saturated model is the "full model" with as many parameters as observations.
This likelihood ratio statistic is called the deviance and plays a central role when dealing with goodness-of-fit. As shown later the deviance computed for a linear regression is equal to the residual sum of squares, so the deviance has similar properties to the residual sum of squares. For the logistic regression, where all observations have different covariate patterns and the response is binary, the log likelihood of the saturated model becomes:

$$\ell(\text{saturated model}) = \log\left[\prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{(1-y_i)}\right] = \log[1] = 0. \quad (2.18)$$

So the deviance is $D = -2\ell(\text{fitted model})$.

Testing of a subset of parameters

If a subset, β_1, of the model parameters, $\beta = \{\beta_1, \beta_2\}$ needs testing, the null hypothesis H_0: $\beta_1 = \beta_1^0$ can be tested with the following test statistics. In this case, β_2 is treated as a nuisance parameter. This might be useful when adjusting for confounding or evaluating the relevance of a specific parameter. Under the null hypothesis and for a large number

of observations the test statistics are all χ^2 distributed with k degrees of freedom, where k is the number of parameters of interest, that is k is the number of entries in β_1.
The likelihood ratio test statistics is then given by

$$LR = -2[\ell(\beta_1^0, \hat{\beta}_2^*) - \ell(\hat{\beta})],$$

where $\hat{\beta}_2^*$ is the maximum likelihood estimate of β_2 under the null hypothesis. This test statistic is the same as the change in the log likelihood ratio, when including the parameters in question and not, since

$$\begin{aligned} & LR[H_0 : \beta = \beta^0] - LR[H_0 : \beta_2 = \beta_2^0 \mid \beta_1 = \beta_1^0] \\ = & -2[\ell(\beta^0) - \ell(\hat{\beta})] + 2[\ell(\beta_1^0, \beta_2^0) - \ell(\beta_1^0, \hat{\beta}_2^*)] \\ = & -2[\ell(\beta_1^0, \hat{\beta}_2^*) - \ell(\hat{\beta})]. \end{aligned}$$

The Wald test statistic is now

$$W = (\hat{\beta}_1 - \beta_1^0)^\top V_{11}^{-1} (\hat{\beta}_1 - \beta_1^0), \text{ where } V = \begin{bmatrix} V_{11} & V_{12} \\ V_{12}^\top & V_{22} \end{bmatrix}.$$

This test statistic is a global Wald statistic limited to the parameters in question.
Note, that when dealing with non-binary discrete explanatory variables the test of significance has to include all design variables concerning the explanatory variable. Statistical programs may give the significance level for each design variable, but a separate analysis must be made to have a valid significance conclusion.

Choosing a test statistic

When choosing between the statistics, both statistical properties and computational expenses should be taken into account. The likelihood ratio test statistic has the best statistical properties followed by the score and Wald test statistic.

When testing the global hypothesis that no effects are significant, LR is often used, because the log likelihood evaluated at the model in question is available from the fitting process and the log likelihood at the model containing only the intercept is easy to compute.

When testing a subset of parameters, the likelihood ratio test statistic requires estimation of all $p+1$ model parameters and of the k parameters of interest under the null hypothesis. The Wald test statistic requires estimation of the $p+1$ model parameters, so the Wald test is the obvious choice, if estimations of the $p+1$ model parameters are made in advance. If there are any problems with the Wald test statistic the likelihood ratio test statistic should be used instead.

Testing the logistic regression

The major statistical problem with W is that it is sensitive to potential problems with the estimated covariance matrices. This is a problem in logistic regression, where the covariance matrix generally is overestimated as effects increase. The result of an overestimated covariance matrix is an underestimated W and it is thereby more difficult for the statistic to be significant.
The Wald test statistic also have a problem with large parameter estimates. If the difference between the parameter estimate and the null value increases, the Wald statistic for H_0: $\beta = \beta^0$ becomes larger, but after a certain point, it then drops and becomes smaller again. If the parameter estimate increases to $\pm\infty$, W drops to zero. Infinite estimates might occur in logistic regression if the mean of the predictor is close to 0 or 1 for one or more of the covariate patterns. In this case the likelihood ratio is preferable.
In the special case where all the explanatory variables are discrete the logistic model statistics are equivalent to a contingency table χ^2 statistics. As an example the global likelihood ratio statistic for all design variables in a k-sample model is the same as the $k \times 2$ contingency table likelihood ratio χ^2 statistic.

Confidence intervals

Pointwise confidence intervals for the parameters can be found with the introduced test statistics. The confidence interval based on the Wald test statistic is

$$\hat{\beta} \pm z_{1-\alpha/2} s,$$

where s is the vector of the diagonal entries in V. This interval is often used because of its simplicity.

The Wald based confidence interval for fitted values corresponding to a given covariate pattern, x, of the explanatory variables is

$$\frac{e^{x\hat{\beta}\pm z_{1-\alpha/2}SE[x\hat{\beta}]}}{1+e^{x\hat{\beta}\pm z_{1-\alpha/2}SE[x\hat{\beta}]}},$$

where $SE[x\hat{\beta}]$ is the estimated standard deviation found from the estimated variance given by

$$\begin{aligned}\hat{V}ar[x\hat{\beta}] &= \sum_{j=0}^{p} x_j^2 \hat{V}ar(\hat{\beta}_j) + \sum_{j=0}^{p}\sum_{k=j+1}^{p} 2x_j x_k \hat{C}ov(\hat{\beta}_j, \hat{\beta}_k) \\ &= x\hat{V}x^\top.\end{aligned}$$

[David W. Hosmer, 2000]
Note that the Wald based confidence intervals are strictly symmetrical intervals on the scale, they are found.
Nonsymmetrical confidence intervals can be found on the basis of the likelihood ratio statistic and the score statistic. They are not as computationally easy as the Wald test statistic though.

2.1.5 Testing model assumptions

In this section methods used to test the model assumptions in a logistic regression are given. The section is written on the basis of
[Frank E. Harrell, 2001].
The logistic regression model is a direct probability model, so there is no assumptions about the parameters, only about the regression equation. The regression equation assumptions are possible to verify both graphically and with tests. To do this, the concept of interaction is needed. Interaction can be seen as a variable describing the effect of combining two or more covariates. This is described further in section 2.1.5. When such a combined effect is absent, it is assumed, that the relationship between the log odds and a continuous explanatory variable is linear, when holding the other explanatory variables constant. This means, that the parameter β_j is the change in the log odds ratio per unit change in X_j.

Consider the simple model with a binary variable X_1 and a continuous variable X_2.

$$\text{logit}(Y=1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

In this case the model assumptions can be evaluated graphically by plotting $\text{logit}(Y = 1|X)$ versus X_2 for both values of X_1. If this plot shows two straight and parallel lines, the assumptions are met. In a more complex setting a similar approach can be used, where a number of different combinations of the explanatory variables are plotted.
Evaluating the assumptions can also be done in a more strict manor by testing for linearity and interaction. Testing for interaction is done by adding an interaction variable, so the model is

$$\text{logit}(Y = 1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$

and then testing the null hypothesis $\beta_3 = 0$. If the interaction is insignificant, the assumptions are met. Testing for linearity can by done by adding transformations or the like of the variable and testing for their significance. In a more complex setting the transformations for all the continuous explanatory variables and all possible interactions should be tested in the same way.

Interaction and confounding

In this section the concept of interaction and confounding is introduced and it is shown how to control for their effects in a logistic regression model. The section is based on [David W. Hosmer, 2000].

The term confounder is used to describe a covariate that is associated with both the outcome variable and a primary independent variable or risk factor. When both associations are present, the relationship between the risk factor and the outcome is said to be confounded.

The term interaction is used when a model contains a risk factor and a covariate and the effect of these two variables are not additive, that is when the effect of the risk factor on the outcome depends on the level of the covariate and vice versa. Epidemiologists use the term effect modifier to describe a variable that interacts with a risk factor.

Interaction can take many different forms. Consider a model containing a binary risk factor variable and a continuous covariate. If the association between the covariate and the outcome variable is the same within each level of the risk factor, then there is no interaction between the covariate and the risk factor. Graphically, the absence of interaction

yields a model with two parallel lines, one for each level of the risk factor.
When interaction is present, the association between the risk factor and the outcome variable differs, or depends on the level of the covariate. The simplest and most commonly used model for including interaction is one where the logit is linear in the confounder for all the levels of the risk factor but with different slopes. In any model, interaction is incorporated by the inclusion of second or higher order terms involving two or more variables.

Determining whether a covariate, X, is an effect modifier and/or a confounder involves several issues. Determining effect modification status involves the parametric structure of the logit, while determination of confounder status involves two things. First the covariate must be associated with the outcome variable. This means that the logit must have a nonzero slope in the covariate. Secondly the covariate must be associated with the risk factor variable. The association between the covariate and the risk factor may be very complex, but the essence is that there must be incomparability in the risk factor groups. This incomparability must be accounted for in the model if a correct, unconfounded, estimate of the effect for the risk factor is to be obtained.

In practice, one method to check the confounder status of a covariate is to compare the estimated coefficient for the risk factor variable from models with and without the covariate. If there is a clinically important change in the estimated coefficient for the risk factor this suggests that the covariate is a confounder and should be included in the model, regardless of the statistical significance of its estimated coefficient.

On the other hand, a covariate is an effect modifier only when the interaction term added is both clinically meaningful and statistically significant. When a covariate is an effect modifier, its status as a confounder is of secondary importance since the estimate of the effect of the risk factor depends on the specific value of the covariate.

The concept of confounding, interaction and effect modification, may be extended to the situations involving any number of variables on any measurement scale. The principals for identification and inclusion of

confounder and interaction variables in the model are the same regardless of the number of variables and their measurement scales.

2.1.6 Variable selection

In this section various methods for selecting the variables that result in a "best" model within the scientific context of the problem is presented. The section is based on [David W. Hosmer, 2000].

In statistical model building it is tradition to seek the most simple model that still explains the data. When the number of variables in the model are minimized, the model is more likely to be numerically stable, and is more easily generalized. The more variables included in a model, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data. Epidemiologic methodologists suggest including all clinically and intuitively relevant variables in the model, regardless of their statistical significance. This is done in order to provide as complete control of confounding as possible within the given dataset.
The major problem with this approach is that the model may be overfitted and produce numerically unstable estimates. Overfitting is typically characterized by unrealistically large estimated coefficients and/or estimated standard errors.

There are several steps that can be followed to aid the selection of variables for a logistic regression model.

Univariate analysis

Begin with a careful univariate analysis of each variable. For nominal, ordinal, and continuous variables with few integer values, this can be done with a contingency table of outcome versus the k levels of the independent variable. The Pearson χ^2 test can then be used to test for association.

For continuous variables, the most desirable univariate analysis involves fitting a univariate logistic regression model to obtain the estimated coefficient, the estimated standard error and the likelihood ratio test for significance of the coefficient.

Select variables for multivariate analysis

Upon completion of the univariate analysis, the variables for the multivariate analysis is selected. Any variable whose univariate test has a p-value < 0.25 is a candidate for the multivariate model along with all variables of known clinically importance. The level of 0.25 is chosen since traditional levels, such as 0.05, often fails to identify variables known to be important. Use of higher levels has the disadvantage of including variables that are of questionable importance
[David W. Hosmer, 2000].

One problem with any univariate approach is that it ignores the possibility that a collection of variables, each of which is weekly associated with the outcome, can become an important predictor of outcome when taken together. The chosen significance level should then be large enough to include the suspected variables in the multivariate model.

If the overall sample size and the number in each outcome group relative to the total number of candidate variables are large enough, it may be useful to begin with the multivariate model containing all possible variables. However, when the data are inadequate, this approach can produce a numerically unstable multivariate model.

Examine whether the included variables are significant

Following the fit of the multivariate model, the importance of each variable included in the model should be verified. This should include an examination of the Wald statistic for each variable and a comparison of each estimated coefficient with the coefficient from the model containing only that variable. Variables that do not contribute to the model based on this criteria should be eliminated and a new model should be fitted. The estimated coefficients for the remaining variables should be compared to those from the full model. If a variable coefficient have changed markedly in magnitude, this indicates that one or more of the excluded variables was important in the sense of providing a needed adjustment of the effect of the variables that remained in the model.
This process of deleting, refitting and verifying continues until it appears that all of the important variables are included in the model.

Check the linearity assumptions for continuous variables

Once a model that contains the essential variables is obtained, the variables in the model should be examined closer. For continuous variables the assumption of linearity in the logit should be checked.

If a continuous variable is represented as X_1 in a model, the model is assumed to be linear in X_1. Often, however, the outcome of interest, Y, does not behave linearly in all the predictors. The simplest way to describe a nonlinear effect of X_1 is to include a nonlinear term, $X_2 = X_1^2$ in the model

$$g(\mathbf{E}[Y|X_1]) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2,$$

where g is a monotonous and continuous function called the link function, described in chapter 2.

If the model is linear in X_1 then β_2 will be zero. When including X_1^2 it is assumed that the model is parabolic in X_1, however nonlinear effects will often not be parabolic. If a transformation of the predictor is known to induce linearity, this transformation may be used. This transformation is, however, seldom known. Higher power of X_1 may be included in the model to approximate many types of relationships, but polynomials have some undesirable properties and will not adequately fit many functional forms. For instance, polynomials do not adequately fit logarithmic functions or threshold effects [Frank E. Harrell, 2001].

Linear spline

Instead of including different transformations of the continuous predictor directly in the model, spline functions can be used. Spline functions are piecewise polynomials used in curve fitting. This means that they are polynomials within intervals of the continuous variable, X, that are connected.
The simplest spline function is a linear spline function, i.e. a piecewise linear function. If the x-axis is divided into intervals with endpoints at a, b, and c, called knots, the linear spline function is given by

$$f(X) = \beta_0 + \beta_1 X + \beta_2 (X - a)_+ + \beta_3 (X - b)_+ + \beta_4 (X - c)_+,$$

where

$$(u)_+ = \begin{cases} u, & u > 0; \\ 0, & u \leq 0. \end{cases}$$

The number of knots can vary depending on the amount of data available for fitting the function.

The general linear regression model can be written assuming only piecewise linearity in X by incorporating constructed variables X_2, X_3 and X_4

$$f(X) = X\beta$$

where $X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$, and $X_1 = X$, $X_2 = (X-a)_+$, $X_3 = (X-b)_+$, $X_4 = (X-c)_+$. By modeling a slope increment for X in an interval $(a, b]$ in terms of $(X-a)_+$ the function is constrained to join at the knots. Overall linearity in X can be tested by testing $H_0 : \ \beta_2 = \beta_3 = \beta_4 = 0$. [Frank E. Harrell, 2001]

Cubic spline functions
Although the linear spline is simple and can approximate many common relationships, it is not smooth and will not fit highly curved functions well. These problems can be overcome by using piecewise polynomials of order higher than linear. Cubic polynomials have been found to have nice properties with good ability to fit highly curved functions. Cubic splines can also be constructed so they are smooth at the join points, by forcing the first and second order derivatives of the function to agree at the knots.

Such a smooth cubic spline function with three knots (a, b, c) is given by

$$f(x) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X-a)_+^3 + \beta_5 (X-b)_+^3 + \beta_6 (X-c)_+^3$$

If a cubic spline function has k knots, it is necessary to estimate $k+3$ regression coefficients besides the intercept. [Frank E. Harrell, 2001]

Restricted cubic splines
Even though the cubic spline function has good ability to fit highly curved functions, it do have some drawbacks. The cubic spline function can behave poorly in the tails, i.e. before the first knot and after the last knot. To handle this problem the function is restricted to be linear

in the tails. The restricted cubic spline function has the additional advantage that only $k-1$ parameters besides the intercept needs to be estimated as opposed to $k+3$ parameters in the unrestricted cubic spline. The restricted cubic spline function with k knots $t_1, \ldots, t_k$ is given by

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-1} X_{k-1}$$

where $X_1 = X$ and for $j = 1, \ldots, k-2$

$$X_{j+1} = (X - t_j)_+^3 - \frac{(X - t_{k-1})_+^3 (t_k - t_j)}{t_k - t_{k-1}} + \frac{(X - t_k)_+^3 (t_{k-1} - t_j)}{t_k - t_{k-1}}$$

Once $\beta_0, \ldots, \beta_{k-1}$ are estimated, the restricted cubic spline can be stated as

$$f(X) = \beta_0 + \beta_1 X + \beta_2 (X - t_1)_+^3 + \beta_3 (X - t_2)_+^3 + \ldots + \beta_{k+1} (X - t_k)_+^3$$

by computing

$$\begin{aligned} \beta_k &= \frac{\beta_2 (t_1 - t_k) + \beta_3 (t_2 - t_k) + \ldots + \beta_{k-1} (t_{k-2} - t_k)}{t_k - t_{k-1}} \\ \beta_{k+1} &= \frac{\beta_2 (t_1 - t_{k-1}) + \beta_3 (t_2 - t_{k-1}) + \ldots + \beta_{k-1} (t_{k-2} - t_{k-1})}{t_{k-1} - t_k}. \end{aligned}$$

A test of linearity in X can be obtained by testing

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_{k-1} = 0.$$

[Frank E. Harrell, 2001]

Choosing number and position of knots

The location of knots in a restricted cubic spline model is not very important in most situations. The fit depends much more on the choice of the number of knots, k. Placing knots at fixed quantiles of a predictor's marginal distribution is a good approach in most datasets. This ensures that enough points are available in each interval, and it also guards against letting outliers overly influence the placement of the knots.

The number of knots chosen is determined by the sample size available to estimate the unknown parameters. More than 5 knots are seldom required in a restricted cubic spline model. [Frank E. Harrell, 2001] The decision is then between $k = 3, 4,$ or 5. For many datasets, $k = 4$ offers an adequate fit of the model and is a good compromise between flexibility and loss of precision caused by overfitting a small sample. When the sample size is large $k = 5$ is a good choice, and with small samples $k = 3$ may be adequate. [Frank E. Harrell, 2001]

Interaction terms are included

When the model is refined and all the continuous variables are scaled correctly, the model is checked for interactions among the variables. The final decision as to whether an interaction term should be included in the model should be based on statistical as well as practical considerations. Any interaction term in the model must make sense from a clinical perspective.

Before the model is used for inferences the fit of the model must be checked. Methods for assessment of fit are described in chapter 2.2.

2.2 Assessment of model fit

In this section various methods used to describe the goodness-of-fit of a logistic regression model are presented. The section is based on [David W. Hosmer, 2000].

Assessment of the fit of a model or the goodness-of-fit of a model is a measure that tells how effectively the model describes the outcome variable. To assess the fit of the model it is necessary to know exactly what it means that the model fits.

Let the observed sample values of the outcome variable be denoted as $\mathbf{y}$ where $\mathbf{y}^\top = (y_1, y_2, \ldots, y_n)$ and the values predicted by the model, the fitted values, as $\hat{\mathbf{y}}$ where $\hat{\mathbf{y}}^\top = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$. The model fits well if

- summary measures of the distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$ are small, and

- the contribution of each pair $(y_i, \hat{y}_i)$, $i = 1, 2, 3, \ldots, n$ to these summary measures is unsystematic with respect to the covariates or the outcome, and it is small relative to the error structure of the model.

So to get a complete assessment of the fit of the model both calculations of the summary measures of the distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$ and a thorough examination of the individual components should be made.

2.2.1 Summary measures of goodness-of-fit

Goodness-of-fit is assessed over the combinations of fitted values determined by the predictors in the model, not the total collection of predictors. For example, suppose that the model contains p independent variables, $\mathbf{x}^\top = (x_1, x_2, \ldots, x_p)$, and let J denote the number of distinct values of $\mathbf{x}$ observed. If some of the n observations have the same value of $\mathbf{x}$ then $J < n$. Let the total number of observations with $\mathbf{x} = \mathbf{x}_j$ be denoted by m_j, $j = 1, 2, \ldots, J$, then $\sum_{j=1}^{J} m_j = n$. Let y_j denote the total number of positive responses among the m_j observations with $\mathbf{x} = \mathbf{x}_j$. The distribution of the goodness-of-fit statistic is then obtained by letting n become large. If the number of covariate patterns increases with n then the number of observations in each pattern, m_j, becomes small. Distributional results obtained when only $n \to \infty$ are said to be based on n-asymptotics. If $J < n$ is fixed then $m_j \to \infty$ when $n \to \infty$. Distributional results based on $m_j \to \infty$ are said to be based on m-asymptotics.
Initially it is assumed that $J \approx n$, as is expected when there is at least one continuous predictor in the model.

Pearson χ^2 statistic and the deviance

In logistic regression there are many ways to measure the difference between the observed and fitted values. To emphasize the fact that the fitted values in a logistic regression are calculated for each covariate pattern and depend on the estimated probability of that pattern, the fitted value for the j'th covariate pattern is denoted as $\hat{y}_j$,

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(x_j)}}{1 + e^{\hat{g}(x_j)}}$$

where $\hat{\pi}_j$ is the estimated probability of the j'th covariate pattern and $\hat{g}(x_j)$ is the estimated logit.

Two measures of the difference between the observed and the fitted values are now considered, the Pearson residual and the deviance residual. For a given covariate pattern the Pearson residual is defined as

$$r(y_j, \hat{\pi}_j) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}. \tag{2.19}$$

The Pearson chi-square statistic is then the summary statistic based on these residuals

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2.$$

The deviance residual is defined as

$$d(y_j, \hat{\pi}_j) = \pm \left\{ 2 \left[y_j \log\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\log\left(\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{1/2} \tag{2.20}$$

where the sign + or − is decided by sign$(y_j - m_j\hat{\pi}_j)$. The summary statistic based on the deviance residuals is the deviance

$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2.$$

Under the assumption that the fitted model is correct in all aspects, the distribution of the statistics X^2 and D is chi-square with $J-(p+1)$ degrees-of-freedom. For the deviance this follows from the fact that D is the likelihood ratio test statistic of the fitted model with $p+1$ parameters versus a saturated model with J parameters, as seen in section 2.1.4. For the X^2 this follows from the fact that X^2 is asymptotically equivalent to D. The prof of the relationship between X^2 and D uses

the Taylor series expansion of $s \log s/t$ about $s = t$.

$$
\begin{aligned}
& D \\
& = 2 \sum_{j=1}^{J} \left\{ (y_j - m_j\hat{\pi}_j) + \frac{1}{2} \frac{(y_j - m_j\hat{\pi}_j)^2}{m_j\hat{\pi}_j} + [(m_j - y_j) - (m_j - m_j\hat{\pi}_j)] \right. \\
& \left. + \frac{1}{2} \frac{[(m_j - y_j) - (m_j - m_j\hat{\pi}_j)]^2}{m_j - m_j\hat{\pi}_j} + \ldots \right\} \\
& \approx 2 \sum_{j=1}^{J} \frac{1}{2} \left(\frac{(y_j - m_j\hat{\pi}_j)^2}{m_j\hat{\pi}_j} + \frac{[(m_j - y_j) - (m_j - m_j\hat{\pi}_j)]^2}{m_j - m_j\hat{\pi}_j} \right) \\
& = \sum_{j=1}^{J} \frac{(y_j - m_j\hat{\pi}_j)^2}{m_j\hat{\pi}_j} \\
& + \frac{(m_j - y_j)^2 + (m_j - m_j\hat{\pi}_j)^2 - 2(m_j - y_j)(m_j - m_j\hat{\pi}_j)}{m_j(1 - \hat{\pi}_j)} \\
& = \sum_{j=1}^{J} \frac{(y_j - m_j\hat{\pi}_j)^2}{m_j\hat{\pi}_j} + \frac{y_j^2 + (m_j\hat{\pi}_j)^2 - 2y_j m_j\hat{\pi}_j}{m_j(1 - \hat{\pi}_j)} \\
& = \sum_{j=1}^{J} \frac{m_j(1 - \hat{\pi}_j)(y_j - m_j\hat{\pi}_j)^2 + m_j\hat{\pi}_j(y_j - m_j\hat{\pi}_j)^2}{m_j^2\hat{\pi}_j(1 - \hat{\pi}_j)} \\
& = \sum_{j=1}^{J} \frac{(y_j - m_j\hat{\pi}_j)^2}{m_j\hat{\pi}_j(1 - \hat{\pi}_j)} \\
& = X^2.
\end{aligned}
$$

However, if $J \approx n$ the chi square distribution is obtained under n-asymptotics, meaning that the number of parameters is increasing with the sample size. Thus, p-values calculated for these two statistics when $J \approx n$ using the $\chi^2(J - p - 1)$ distribution, are incorrect. One way to avoid the above difficulties with the distributions of X^2 and D when $J \approx n$ is to group the data such that m-asymptotics can be used.

The Hosmer-Lemeshow test

Hosmer and Lemeshow suggested grouping based on the values of the estimated probabilities. Suppose, for sake of discussion that $J = n$. In this case the statistics are obtained from a $2 \times n$ table with the rows corresponding to the two outcomes and the columns to the n

values of the estimated probabilities, with the smallest value as the first column, and the largest value as the n'th column. There are two different grouping strategies:

1. Collapse the table based on percentiles of the estimated probabilities.

2. Collapse the table based on fixed values of the estimated probabilities.

In both strategies it is common to collapse the table into 10 groups. In the first strategy the first group contains the $n'_1 = n/10$ observations having the smallest estimated probabilities, and the last group contains the $n'_{10} = n/10$ observations with the largest estimated probabilities.

The second strategy obtains cutpoints at the values $k/10, k = 1, 2, \ldots, 9$, and the groups contain all observations with an estimated probability between the adjacent cutpoints.

For the $y = 1$ row the expected probability for a given group is obtained by summing the estimated probability of all the observations in the group. For the $y = 0$ row, the expected probabilities are obtained by summing one minus the estimated probability for all the observations in the groups. For both grouping methods, the Hosmer-Lemeshow goodness-of-fit statistic, χ^2_{HL} is calculated as follows

$$\chi^2_{HL} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

where g is the number of groups and n'_k is the total number of observations in the k'th group. The number of covariate patterns in the k'th group is denoted as c_k, the number of responses in the k'th group as

$$o_k = \sum_{j=1}^{c_k} y_j,$$

and

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k}$$

is the average estimated probability.

If $J \approx n$ and the fitted logistic regression model is the correct one, the test statistic is approximately χ^2 distributed with $g-2$ degrees of freedom. Research has shown that the grouping method based on percentiles are to be preferred because it better adherence to the χ^2_{g-2} distribution, especially when many of the estimated probabilities are small, i.e. less than 0.2 [David W. Hosmer, 2000]. Thus unless stated otherwise the Hosmer-Lemeshow test statistic is based on the percentile grouping.

Because the distribution of χ^2_{HL} depends on m-asymptotics, the appropriateness of the p-value depends on whether the estimated expected frequencies are large. If all the expected frequencies are greater than 5, then there is reason to believe that the calculation of the p-value is accurate enough to support the hypothesis of modelfit.
Additional comments on the calculations of χ^2_{HL} are needed. When the number of covariate patterns is less than n some of the patterns has $m_j > 1$ and there is therefore a possibility that a pattern will occur in more than one probability group. The value of χ^2_{HL} will then, to some extent, depend on how these ties are assigned to the groups. Different statistical packages handles ties differently, but the use of different methods is not likely to be an issue unless the number of covariate patterns is so small that assigning all tied values to one group results in a huge imbalance in group size, or worse in fewer than 10 groups. In addition, when too few groups are used to calculate χ^2_{HL} the sensitivity may be too small to distinguish between observed and expected frequencies. If χ^2_{HL} is calculated from fewer then 6 groups it will almost always indicate that the model fits.

The advantage of a summary goodness-of-fit statistic like χ^2_{HL} is that it gives a single value that can be used to assess the fit. The disadvantage is however, that in the grouping process important deviation from fit due to a small number of individual datapoints, may be missed. Another disadvantage is that the test is fairly dependent on the choice of how the predictions are grouped and therefore the choice of the number of groups should be independent of n [Frank E. Harrell, 2001]. Hence before finally accepting the modelfit, an analysis of the individual residuals and relevant diagnostic statistics should be performed.

This is described further in section 2.2.2.
However the table containing the expected and observed frequencies of the different groups, contains descriptive information about the adequacy of the fitted model in the different groups. Comparing observed and expected frequencies within each group may indicate where the model does not perform satisfactorily.

Power in detecting lack of fit
As mentioned before a complete assessment of fit involves summary tests and measures as well as diagnostic statistics. This is especially important to keep in mind when using overall statistics. The desired outcome for most investigations is not to reject the null hypothesis that the model fits. With this decision one is subject to the possibility of type II error and hence the power of the test becomes an issue. To get powerful goodness-of-fit tests one should have a sample size of $n > 400$.

Another way to obtain more power for detecting lack of fit is to test specific alternatives to the model. To test the assumptions of linearity and additivity, the model may be expanded with cubic splines for each of the continuous predictors and with interaction terms for each possible interaction, and then the new coefficients are tested. There are virtually no departure from linearity and additivity that cannot be detected from this expansion. [Frank E. Harrell, 2001]

Another measure of model performance that often is a useful supplement to the overall test of fit, will now be presented before the diagnostic statistics are discussed.

Area under the ROC curve

Sensitivity and specificity rely on a single cutpoint to classify a test result as being positive. A more complete description of classification ability is given by the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve maps the probability of true positives versus false positives, i.e. the sensitivity versus 1-specificity, for the entire range of cutpoints. The area under the ROC curve, which ranges from 0 to 1, gives a measure of the model's ability to discriminate between those observations with $y = 1$ and those with $y = 0$.

As a general rule for the area A

- $A = 0.5$: suggests no discrimination
- $0.7 \leq A < 0.8$: is acceptable discrimination
- $0.8 \leq A < 0.9$: is excellent discrimination
- $A \geq 0.9$: is outstanding discrimination.

In practice it is extremely unusual to observe areas under the ROC curve greater than 0.9. This is because when there is complete separation it is impossible to estimate the coefficients of a logistic regression model, and to obtain an area greater than 0.9 almost complete separation is required.

One should keep in mind that a poorly fitted model may still have good discrimination. For example, if 0.25 is added to every probability in a good fitted logistic model with good discrimination abilities, the new model would now fit poorly but the discrimination would be unaffected. The model performance should then be assessed by considering both calibration and discrimination.

Other summary measures

A short discussion of R^2 measures are now presented. In general, these measures are based on various comparisons of the predicted values from the fitted model to those from the null-model, the no data or intercept only model, and as a result are not goodness-of-fit measures. A true measure of fit is one based on a comparison of observed values and values from the fitted model. However, there are situations where the R^2 measure can be useful when comparing the fit of competing models on the same data.
Hosmer and Lemeshow, [David W. Hosmer, 2000] propose the following as criteria for a good measure:

- The measure has an easily understood interpretation.
- The squared measure has a lower bound of 0 and an upper bound of 1.
- The measure is not changed by a linear transformation of model covariates.

The linear regression-like sum-of-squares R^2 satisfies these three criteria.

When there are n covariate patterns the linear regression-like measure is

$$R^2_{SS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where $\hat{\pi}_i$ is the estimated probability of the i'th covariate pattern and $\bar{y} = \bar{\pi} = \frac{\sum_{i=1}^{n} y_i}{n}$.
The extension of this measures to the setting of $J < n$ covariate patterns is

$$R^2_{SSC} = 1 - \frac{\sum_{j=1}^{J}(y_j - m_j\hat{\pi}_j)^2}{\sum_{j=1}^{J}(y_j - m_j\bar{y})^2}.$$

Another version of R^2_{SSC} is obtained when the log-likelihoods are used instead of sums-of-squares. If we let L_0 and L_p denote the log-likelihoods for the model containing only the intercept and the model containing the intercept and the p covariates respectively, then the log-likelihood-based R^2 is

$$R^2_L = \frac{L_0 - L_p}{L_0} = 1 - \frac{L_p}{L_0}.$$

The maximum value for R^2_L is obtained when the saturated model is fitted. If $J = n$ then the log-likelihood for the saturated model is zero, $L_S = 0$, as seen in line (2.18) and then $R^2_L = 1$. However, if $J < n$ then the maximum is less than 1. A modification of the statistic that can attain 1 in the $J < n$ case is

$$R^2_{LS} = \frac{L_0 - L_p}{L_0 - L_S}.$$

The value of the log-likelihood from the saturated model, L_S, can be calculated from the deviance for the model with p covariates:

$$L_S = L_p + 0.5D,$$

where $D = \sum_{j=1}^{J} 2\left[y_j \log\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)\log\left(\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right)\right]$. Unfortunately low R^2 values in logistic regression are the norm and this presents a problem when reporting their values to an audience accustomed to seeing linear regression values. Thus routinely publishing of R^2 values with results from fitted logistic regression models are not recommended. However, they may be helpful in the model building stage as a statistic to evaluate competing models.

2.2.2 Logistic regression diagnostics

As mentioned before summary statistics based on Pearson chi-square residuals provides a single value that summarizes the agreement between observed and fitted values. However these statistics do not provide information about deviation from fit due to a small number of individual data points. Therefore it is important to examine other measures to see if fit is supported over the entire set of covariate patterns before concluding that the model fits. This is done through a number of specialized measures that falls under the general heading of regression diagnostics.

The key quantities for logistic regression diagnostics are the components of the residual sum-of-squares. In linear regression a key assumption is that the error variance does not depend on the conditional mean, $\mathbf{E}[Y_j|\mathbf{x}_j]$. However, in logistic regression the error has a binomial distribution and, as a result, the error variance is a function of the conditional mean

$$\begin{aligned}\text{Var}[Y_j|\mathbf{x}_j] &= m_j\mathbf{E}[Y_j|\mathbf{x}_j]\left(1 - \mathbf{E}[Y_j|\mathbf{x}_j]\right) \\ &= m_j\pi(\mathbf{x}_j)\left[1 - \pi(\mathbf{x}_j)\right].\end{aligned}$$

Thus when looking at the residuals in line (2.19) and (2.20), it is seen that they are divided by estimates of their standard errors. Let r_j and d_j denote the values of the Pearson residual in (2.19) and the deviance residual in (2.20) respectively, for covariate pattern $\mathbf{x}_j$. Since the residuals have been divided by an estimate of the standard error, it is expected that if the model is correct these quantities have a mean approximately equal to zero and a variance approximately equal to 1.

In addition to the residuals for each covariate pattern, other quantities important to the interpretation of linear regression diagnostics are the hat matrix and the leverage values derived from it. In linear regression the hat matrix is the matrix that provides the fitted values as the projection of the outcome variable into the covariate space. Let $\mathbf{X}$ denote the $J \times (p+1)$ design matrix containing the values for all J covariate patterns formed from the observed values of the p covariates. In linear regression the hat matrix is $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. The hat matrix for logistic regression is shown in [Pregibon, 1981] to be

$$\mathbf{H} = \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}^\top\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{V}^{1/2}$$

where $\mathbf{V}$ is a $J \times J$ diagonal matrix with $v_j = m_j\hat{\pi}(\mathbf{x}_j)[1-\hat{\pi}(\mathbf{x}_j)]$.
In linear regression the diagonal elements of the hat matrix are called the leverage values and are proportional to the distance from $\mathbf{x}_j$ to the mean of the data. The concept of leverage is important in linear regression, since points far from the mean may have considerable influence on the values of the estimated parameters.
Let h_j denote the j'th diagonal element in the matrix $\mathbf{H}$ for logistic regression. It can be shown that

$$h_j = m_j\hat{\pi}(\mathbf{x}_j)[1-\hat{\pi}(\mathbf{x}_j)]\mathbf{x}_j^\top(\mathbf{X}^\top\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_j = v_j b_j,$$

where $b_j = \mathbf{x}_j^\top(\mathbf{X}^\top\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_j$ and $\mathbf{x}_j^\top = (1, x_{1j}, x_{2j}, \ldots, x_{pj})$ is the vector of covariate values defining the j'th covariate pattern.

When looking at the Pearson residuals it seems that they have to be further standardized in order for them to have a variance of 1. This can be seen from the following calculations. If a Taylor approximation of the residual is made, the result is

$$y - \hat{y} \approx y - \pi(\hat{\beta}) + \frac{\partial\pi}{\partial\beta^\top}(\beta - \hat{\beta}).$$

This can be rewritten as

$$\begin{aligned}\mathbf{V}^{-\frac{1}{2}}(y-\hat{y}) &\approx \mathbf{V}^{-\frac{1}{2}}(y-\pi(\hat{\beta})) + \mathbf{V}^{-\frac{1}{2}}(\frac{\partial\pi}{\partial\beta^\top}(\beta-\hat{\beta})) \\ &= \mathcal{X} + \mathbf{V}^{\frac{1}{2}}\mathbf{X}(\beta-\hat{\beta}),\end{aligned}$$

since $\frac{\partial \pi}{\partial \beta^\top} = \frac{\partial \pi}{\partial \theta^\top}\frac{\partial \theta}{\partial \beta^\top} = \mathbf{VX}$, where $\theta = \text{logit}(\pi)$ and where $\mathcal{X}$ is the Pearson residual. By further expansion one gets:

$$\begin{aligned}(\mathbf{I}-\mathbf{H})\mathbf{V}^{-\frac{1}{2}}(y-\hat{y}) &\approx (\mathbf{I}-\mathbf{H})\mathcal{X} + (\mathbf{I}-\mathbf{H})\mathbf{V}^{\frac{1}{2}}\mathbf{X}(\beta-\hat{\beta}) \\ &= \mathcal{X}\end{aligned}$$

where $\mathbf{H}\mathcal{X} = 0$ since $\mathbf{H}\mathcal{X} = \mathbf{V}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^\top\mathbf{VX})^{-1}\mathbf{X}^\top V^{\frac{1}{2}}\mathcal{X}$ and $\mathbf{X}^\top\mathbf{V}^{\frac{1}{2}}\mathcal{X} = \mathbf{X}^\top(y-\hat{y}) = 0$. The variance of the Pearson residual $\mathcal{X}$ is then

$$\begin{aligned}\text{Var}(\mathcal{X}) &= (\mathbf{I}-\mathbf{H})\mathbf{V}^{-\frac{1}{2}}\text{Var}(Y-\hat{y})((\mathbf{I}-\mathbf{H})\mathbf{V}^{-\frac{1}{2}})^\top \\ &= (\mathbf{I}-\mathbf{H})\mathbf{V}^{-\frac{1}{2}}\mathbf{V}\mathbf{V}^{-\frac{1}{2}}(\mathbf{I}-\mathbf{H})^\top \\ &= (\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})^\top = (\mathbf{I}-\mathbf{H}),\end{aligned}$$

where the last equality follows from the fact that $(\mathbf{I}-\mathbf{H})$ is idempotent. The variance of the j'th covariate pattern is then $1-h_j$. Let r_j denote the Pearson residual in equation (2.19), then the standardized Pearson residual for covariate pattern $\mathbf{x}_j$ is

$$r_{sj} = \frac{r_j}{\sqrt{1-h_j}}.$$

Another useful diagnostic statistic is one that examines the effect deletion of the j'th covariate pattern has on the value of the estimated coefficient. The basic formula for the change in the estimate $\hat{\beta}_j$ is given according to [Pregibon, 1981] by,

$$\Delta\hat{\beta}_j = \hat{\beta} - \hat{\beta}_{-j} = \frac{(\mathbf{X}^\top\mathbf{VX})^{-1}\mathbf{x}_j(y_j - \hat{y_j}}{1-h_j}$$

where $\hat{\beta}$ and $\hat{\beta}_{-j}$ are the maximum likelihood estimates computed using all observations and excluding the j'th covariate pattern. A scalar measure that summarizes the effect of deleting the j'th covariate pattern over all the coefficients is

$$\begin{aligned}c_j &= (\hat{\beta}-\hat{\beta}_{-j})^\top(\mathbf{X}^\top\mathbf{VX})(\hat{\beta}-\hat{\beta}_{-j}) \\ &= \frac{r_j^2 h_j}{(1-h_j)^2} \\ &= \frac{r_{sj}^2 h_j}{1-h_j}.\end{aligned}$$

In [Pregibon, 1981] it is shown that the decrease in the value of the Pearson chi-square statistic due to deletion of the observations with covariate pattern $\mathbf{x}_j$ is

$$\Delta X_j^2 = \frac{r_j^2}{1 - h_j} = r_{sj}^2.$$

A similar quantity may be obtained for the change in the deviance, which also is shown in [Pregibon, 1981]

$$\Delta D_j = d_j^2 + \frac{r_j^2 h_j}{1 - h_j}.$$

By replacing r_j^2 with d_j^2 this approximation is obtained:

$$\Delta D_j = \frac{d_j^2}{1 - h_j}.$$

These diagnostic statistics are appealing, since large values of ΔX_j^2 and/or ΔD_j help identify those covariate patterns that are poorly fitted, and since large values of c_j identifies those observations that have a great deal of influence on the values of the estimated parameters.

The approach used to interpret the values of the diagnostic are graphical. Large values of diagnostics either appear as spikes or reside in the extreme corners of the plot.

A number of different plots have been suggested for use, each directed at a particular aspect of fit. A few easily obtained plots that are meaningful in logistic regression are

- the plot of ΔX_j^2 versus $\hat{\pi}_j$,
- the plot of ΔD_j versus $\hat{\pi}_j$,
- the plot of $\Delta\hat{\beta}_j$ versus $\hat{\pi}_j$.

If there are some overinfluential observations or covariate patterns these may be deleted if they are not clinical relevant.

2.3 Model validation

In this section model validation methods are described. If nothing else is noted this section is based on [David W. Hosmer, 2000].

Model validation is done to ascertain whether predicted values from the model are likely to accurately predict responses on future observations or observations not used to develop the model.

There are two major methods of model validation, external and internal. Within the clinical world, the most stringent external validation involves testing a final model developed in one country on observations from another country at another time. Testing a finished model on new observations from the same geographic area but from a different institution as observations used to fit the model is a less stringent form of external validation. The least stringent form of external validation involves using the first m of n observations for model training and using the remaining $n - m$ observations as a test sample.
Internal validation involves fitting and validating the model by carefully using one series of observations. The one dataset is used in this way to estimate the likely performance of the final model on new observations. [Frank E. Harrell, 2001]

The use of validation data amounts to an assessment of goodness-of-fit where the fitted model is considered to be known, and no estimation is performed.

The methods for assessment of fit in the validation sample parallel those described in section 2.2.1 and 2.2.2 for the developmental sample. The major difference is that the values of the coefficients in the model are regarded as fixed constants rather than estimated values.

Suppose that the validation sample consists of n_v observations, which may be grouped into J_v covariate patterns. In keeping previous notation, let y_j denote the number of positive responses among the m_j observations with covariate pattern $\mathbf{x} = \mathbf{x}_j$ for $j = 1, 2, \ldots, J_v$. The probability for the j'th covariate pattern is π_j, the value of the previously estimated logistic model using the covariate pattern $\mathbf{x}_j$, from the validation sample. These quantities become the basis for the com-

putation of the summary measures of fit, X^2, D, and χ^2_{HL} from the validation sample.

The computation of the Pearson chi-square statistic follows directly from equation (2.19), with proper substitution of quantities from the validation sample. In this case X^2 is computed as the sum of J_v independent terms. If each $m_j\pi_j$ is large enough to use the normal approximation to the binomial distribution, then X^2 is distributed as $\chi^2(J_v)$ under the hypothesis that the model is correct. In practice the observed number of observations within each covariate pattern is often small, with most $m_j = 1$. Hence m-asymptotics cannot be employed. In this case the Hosmer-Lemeshow test should be used.

The same line of reasoning discussed in section 2.2.1 to develop the Hosmer-Lemeshow test may be used to obtain an equivalent statistic for the validation sample. Assume that 10 groups composed of the deciles of risk is used. Any other grouping strategy could be used with proper modifications in the calculations. Let n_k denote the approximately $n_v/10$ observations in the k'th decile of risk. Let $o_k = \sum_{j=1}^{J_{vk}} y_j$ be the number of positive responses among the covariate patterns falling in the k'th decile of risk. The estimate of the expected value of o_k under the assumption that the model is correct is $e_k = \sum_{j=1}^{J_{vk}} m_j\pi_j$, where J_{vk} is the number of covariate patterns in the k'th decile of risk. The Hosmer-Lemeshow statistic is as the Pearson χ^2 statistic computed from the observed and expected frequencies

$$\chi^2_{HL} = \sum_{k=1}^{g} \frac{(o_k - e_k)^2}{n_k\bar{\pi}_k(1-\bar{\pi}_k)},$$

where $\bar{\pi}_k = \sum_{j=1}^{J_{vk}} m_j\pi_j/n_k$. Under the hypothesis that the model is correct, and the assumption that each e_k is sufficiently large for each term in χ^2_{HL} to be distributed as $\chi^2(1)$, it follows that χ^2_{HL} is distributed as $\chi^2(10)$. In addition to calculating a p-value to assess overall fit, it is recommended that each term in χ^2_{HL} is examined to assess the fit within each decile of risk.

To assess the models ability to discriminate on the validation sample the ROC curves may be plotted as described in section 2.2.1.

Naive Bayes method

This chapter is written on the basis of [Elomaa and Rousu, 2003] and [Tom M. Mitchell (2010),] and contains a short introduction to the naive Bayes method.
The naive Bayes is a simple and surprisingly effective classification algorithm. Its classification performance is comparable to state-of-the-art classifiers [Friedman et al., 1997]. This is a surprise, because the naive Bayes algorithm has extreme independence assumptions, which seldom are met in practice.

The general idea behind classification is to find a systematic way of predicting what class a subject is in, given a set of measurements on the subject. A classification rule is a systematic way of predicting what class a subject belongs to.
Assume $X = (X_1, X_2, \ldots, X_n)$ is a vector of n stochastic variables, discrete or real numbered, where x_i is a realization of X_i. Let Ψ, the sample space, be the set of all possible samples $x = (x_1, x_2, \ldots, x_n)$. The binary variable Y is the classification variable, with y being the realization of Y.
According to Bayes' theorem the probability of a classification variable is then

$$P(Y = y|X = x) = \frac{P(Y = y)P(X = x|Y = y)}{P(X = x)}.$$

The naive Bayes assumes, that the stochastic variables are independent given the class $Y = y$, so this becomes:

$$\begin{aligned} P(Y = y|X = x) &= \frac{P(Y = y)\prod_{i=1}^{n} P(X_i = x_i|Y = y)}{P(X = x)} \\ &= \frac{P(Y = y)\prod_{i=1}^{n} P(X_i = x_i|Y = y)}{\sum_{j=0}^{1} P(Y = j)\prod_{k=1}^{n} P(X_k = x_k|Y = j)}. \end{aligned} \tag{3.1}$$

This means that a classification rule based on Bayes' theorem is

$$\arg \max_{y \in \{0,1\}} P(Y = y|X = x).$$

The values of $P(Y = y)$ and $P(X_i = x_i|Y = y)$ can be estimated using a training dataset.

Consider the training dataset consisting of m observations $(x, y) \in (\Psi, \{0, 1\})$ and let m_y be the number of observations with the class y. Estimating $P(Y = y)$ can then be done with $p(y) = m_y/m$. When X_i is discrete the estimation of the conditional marginal probability $P(X_i = x_i|Y = y)$ can be done with $p(x_i|y) = m_{x_iy}/m_y$, where m_{x_iy} is the number of observations with $X_i = x_i$ and class y. Handling the continuous X_i can be done by assuming normality and then estimating the mean, μ_{iy} and standard deviation, σ_{iy} with the x_i's from the training dataset. These estimations, are then used to calculate the conditional marginal probabilities

$$p(x_i|y) = \frac{1}{\sqrt{2\pi}\sigma_{iy}} \exp\left(-\frac{(x_i - \mu_{iy})^2}{2\sigma_{iy}^2}\right).$$

In naive Bayes continuous variables are often handled this way, but doing so, the discrete and continuous variables are treated differently. Discretization of the continuous variable has been seen to improve the classification performance and make naive Bayes more efficient [Dougherty et al., 1995].
An example of one of the most simple discretization methods is equal-width binning. The method is applied by setting a number of cut points evenly over the range of X_k. A discrete variable defined by a number of intervals with equal width and by the assumption that X_k is uniformly distributed within the intervals, called bins is made.

Linking naive Bayes to logistic regression

Both logistic regression and the naive Bayes can be used to describe $P(Y = y|X = x)$. To be able to compare the estimates from logistic regression and naive Bayes later, the following computations are made. First the log odds ratio given by $\log[\text{Odds}(Y = 1|X = x)/\text{Odds}(Y = 1)]$ is rewritten as:

$$
\begin{aligned}
&\operatorname{logit}(P(Y=1|X=x)) - \operatorname{logit}(P(Y=1)) \\
=&\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} \frac{P(Y=0)}{P(Y=1)} \\
=&\log \frac{P(Y=1)P(X=x|Y=1)}{P(Y=0)P(X=x|Y=0)} \frac{P(Y=0)}{P(Y=1)} \\
=&\log \frac{\prod_{i=1}^{n}\left[P(X_i=x_i|Y=1)\right]}{\prod_{j=1}^{n}\left[P(X_j=x_j|Y=0)\right]} \\
=&\log \prod_{i=1}^{n}\left[\frac{P(X_i=x_i|Y=1)P(Y=1)P(Y=0)}{P(X_i=x_i|Y=0)P(Y=0)P(Y=1)}\right] \\
=&\log \prod_{i=1}^{n}\left[\frac{P(Y=1|X_i=x_i)P(X_i=x_i)}{P(Y=0|X_i=x_i)P(X_i=x_i)} \frac{P(Y=0)}{P(Y=1)}\right] \\
=&\log \prod_{i=1}^{n}\left[\frac{P(Y=1|X_i=x_i)}{P(Y=0|X_i=x_i)} \frac{P(Y=0)}{P(Y=1)}\right] \\
=&\sum_{i=1}^{n}\left[\operatorname{logit}P(Y=1|X_i=x_i) - \operatorname{logit}P(Y=1)\right].
\end{aligned}
$$

This means, that naive Bayes can express the probability of belonging to class $Y=1$ given $X=x$ for binary X_i's as

$$
\begin{aligned}
&\operatorname{logit}(P(Y=1|X=x)) \\
=&(1-n)\operatorname{logit}P(Y=1) + \sum_{i=1}^{n}\left[\operatorname{logit}P(Y=1|X_i=0) + \right. \\
&\left.\left(\operatorname{logit}P(Y=1|X_i=1) - \operatorname{logit}P(Y=1|X_i=0)\right)x_i\right] \\
=&\beta_0 + \sum_{i=1}^{n}\beta_i x_i, \qquad (3.2)
\end{aligned}
$$

where

$$
\beta_0 = (1-n)\operatorname{logit}P(Y=1) + \sum_{i=1}^{n}\operatorname{logit}P(Y=1|X_i=0)
$$

and

$$
\beta_i = \operatorname{logit}P(Y=1|X_i=1) - \operatorname{logit}P(Y=1|X_i=0)
$$

for $i = 1, \ldots, n$.
If X_l is a discrete variable with the possible values $\alpha_1, \alpha_2, \ldots, \alpha_m$, $m > 2$, then a reference value, say α_m, is chosen and $m - 1$ indicator variables are constructed. The indicator values are constructed, so that X_l corresponds to the indicator variables $X_{l_1}, \ldots, X_{l_{m-1}}$, where $X_{l_i} = 1$ if $X_l = \alpha_i$ and zero otherwise. With use of these indicator variables the function (3.2) can also be used when one or more X's are discrete. When estimating $\beta_0, \ldots, \beta_n$ with the naive Bayes method, the function (3.2) corresponds to the function from logistic regression, (2.6). This function is suitable for the classification task, but may not have as good predictability performance as the logistic regression.

Theory of classification trees

In this chapter the concept of classification trees is introduced. When wanting to classify subjects a classification tree is an alternative method to logistic regression and the naive Bayes method. Classification trees have the advantages that their results are simple to understand and to interpret, and that they perform well with large data in a short time. This chapter is based on [Leo Breiman, 1984].

The first thing is to define a classification rule for the classification trees. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a vector containing the measurements of each subject, and let $\mathcal{X}$ be a set of all possible measurement vectors. Suppose that the subjects fall into K classes and let $\mathcal{C} = \{1, 2, \ldots, K\}$ be the set of classes.
A classification rule is then a function $d(\mathbf{x})$ defined on $\mathcal{X}$ taking values in $\mathcal{C}$. Another way of looking at a classifier is to define A_k as the subset of $\mathcal{X}$ on which $d(\mathbf{x}) = k$, that is $A_k = \{\mathbf{x}|d(\mathbf{x}) = k\}$. The sets $A_1, \ldots, A_K$ are disjoint and $\mathcal{X} = \bigcup_k A_k$, so the A_j form a partition of $\mathcal{X}$.

Classifiers are constructed based on past experiences. These past experiences are summarized by a learning sample, which consists of the measurements on N subjects together with their actual classes. That is a learning sample $\mathcal{L}$ consists of data $(\mathbf{x_1}, k_1), \ldots, (\mathbf{x}_N, k_N)$ where $\mathbf{x}_n \in \mathcal{X}$ and $k_n \in \{1, \ldots, K\}$, $n = 1, 2, \ldots, N$.

Estimating accuracy

When constructing classifiers, the concept of estimating accuracy is essential.

Given a classifier, that is, given a function $d(\mathbf{x})$ defined on $\mathcal{X}$ taking values in $\mathcal{C}$ let $R^*(d)$ denote the "true misclassification rate" of the classifier. The value of $R^*(d)$ can be conceptualized in the following way. Define the space $\mathcal{X} \times \mathcal{C}$ as the set of all couples $(\mathbf{x}, k)$ where $\mathbf{x} \in \mathcal{X}$ and $k \in \mathcal{C}$. Let $P(A, k)$ be a probability on $\mathcal{X} \times \mathcal{C}$, i.e. $P(A, k)$ is the probability that a measurement vector is in A and its class is k.

Assume that the learning sample $\mathcal{L}$ is drawn at random from the distribution $P(A, k)$, and construct $d(\mathbf{x})$ using $\mathcal{L}$. Then define $R^*(d)$ as the probability that d will misclassify a new sample drawn from the same distribution as $\mathcal{L}$, $R^*(d) = P(d(\mathbf{X}) \neq Y|\mathcal{L})$, where $(\mathbf{X}, Y)$, $\mathbf{X} \in \mathcal{X}$, $Y \in \mathcal{C}$ is the new sample from the probability distribution $P(A, k)$.

There are three ways of estimating the true misclassification rate given a learning sample. The first is the resubstitution estimate. The resubstitution estimate, $R(d)$, is

$$R(d) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[d(\mathbf{x}_n) \neq k_n].$$

The problem with the resubstitution estimate is that it is computed using the same data used to construct d, instead of an independent sample. $R(d)$ can therefore give an overly optimistic estimate of the accuracy of d.

The second method is test sample estimation. Here $\mathcal{L}$ is divided into two sets $\mathcal{L}_1$ and $\mathcal{L}_2$ where $\mathcal{L}_1$ is used to construct d and $\mathcal{L}_2$ is used to estimate $R^*(d)$. Let N_2 be the number of subjects in $\mathcal{L}_2$, then the test sample estimate, $R^{ts}(d)$, is given by

$$R^{ts}(d) = \frac{1}{N_2} \sum_{(\mathbf{x},k) \in \mathcal{L}_2} \mathbb{I}[d(\mathbf{x}) \neq k]$$

Care needs to be taken so that the subjects in $\mathcal{L}_2$ can be considered as independent of those in $\mathcal{L}_1$ and drawn from the same distribution. The test sample approach has the drawback that it reduces effective sample size.

For smaller sample sizes the third approach, called V-fold cross validation is preferred. The cases in $\mathcal{L}$ are randomly divided into V subsets of nearly equal size. Denote these subsets as $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_V$. For every v, $v = 1, \ldots, V$, use the learning sample $\mathcal{L} - \mathcal{L}_v$ to find $d^{(v)}(\mathbf{x})$. Then $\mathcal{L}_v$ can be used to calculate a test sample estimate for $R^*(d^{(v)})$:

$$R^{ts}(d^{(v)}) = \frac{1}{N_v} \sum_{(\mathbf{x},k) \in \mathcal{L}_v} \mathbb{I}[d^{(v)}(\mathbf{x}) \neq k]$$

where N_v is the number of cases in $\mathcal{L}_v$. The basic assumption of cross-validation is that the classifier $d^{(v)}, v = 1, \ldots, N$ have misclassification rates that approximates $R^*(d)$. The V-fold cross-validation estimate is then defined as

$$R^{cv}(d) = \frac{1}{V} \sum_{v=1}^{V} R^{ts}(d^{(v)}).$$

Bayes rule

The major guide that has been used in the construction of classifiers is the concept of the Bayes rule. Suppose that $(\mathbf{X}, Y)$, $\mathbf{X} \in \mathcal{X}$, $Y \in \mathcal{C}$, is a random sample from the probability distribution $P(A, k)$ on $\mathcal{X} \times \mathcal{C}$. Then $d_B(\mathbf{x})$ is a Bayes rule if for any other classifier $d(\mathbf{x})$, $P(d_B(\mathbf{X}) \neq Y) \leq P(d(\mathbf{X}) \neq Y)$. The Bayes misclassification rate is

$$R_B = P(d_B(\mathbf{X}) \neq Y).$$

4.1 Construction of the tree classifier

Tree structured classifiers are constructed by repeating splits of subsets of $\mathcal{X}$ into two descendant subsets, beginning with $\mathcal{X}$. In the terminology of tree theory, a node t is a subset of $\mathcal{X}$, the root node t_1 is $\mathcal{X}$, and the leaves of the tree are called the terminal nodes.

The construction of a classification tree includes three elements:

- The selection of the splits
- When to stop splitting nodes
- The assignment of each terminal node to a class

4.1.1 Selection of the splits

The first problem in tree construction is how to use $\mathcal{L}$ to find the binary splits that divides $\mathcal{X}$ into smaller and smaller subsets. The fundamental idea is to select each split of a subset, so that the data in each of the descendant subsets are purer than the data in the parent subset.

For any node t, suppose that there is a candidate split s of the node which divides it into t_L and t_R such that a proportion p_L of the subjects

in t go into t_L and a proportion p_R go into t_R. Then the goodness of the split is defined to be the decrease in impurity

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R),$$

where $i(t)$ is a measure of impurity of t.

A tree is then grown in the following way. At the root note t_1 a search is made through all candidate splits to find that split s^* which gives the largest decrease in impurity, i.e.

$$\Delta i(s^*, t_1) = \max_{s \in \mathcal{S}} \Delta i(s, t_1)$$

where $\mathcal{S}$ is a set of all candidate splits. Then t_1 is split into t_2 and t_3 using the split s^* and the same search procedure for the best split is repeated on both t_2 and t_3 separately.
When a node t is reached such that no significant decrease in impurity is possible, then t is not splitted and becomes a terminal node.

4.1.2 Initial tree growing methodology

The initial methods used for constructing tree classifiers is now formulated in more detail.

In the learning sample $\mathcal{L}$ for a K class problem, let N_k be the number of subjects in class k. The set of priors $\{\pi(k)\} = \{P(Y = k)\}$ are either estimated from the data as $\{N_k/N\}$ or supplied by the analyst.

In a node t, let $N(t)$ be the total number of subjects in $\mathcal{L}$ with $\mathbf{x} \in t$, and $N_k(t)$ the number of class k subjects in t. The proportion of class k subjects in $\mathcal{L}$ falling into t is $N_k(t)/N_k$. For a given set of priors, $\pi(k)$ is interpreted as the probability that a class k subject will be presented to the tree. Therefore,

$$p(k,t) = \pi(k) N_k(t)/N_k$$

is taken as the resubstitution estimate for the probability that a subject will both be in class k and fall into node t. Note that from this point on lower case p denotes an estimated probability and upper case P a theoretical probability.

The resubstitution estimate $p(t)$ of the probability that any subject falls into node t is defined by

$$p(t) = \sum_{k=1}^{K} p(k,t).$$

The four elements needed in the initial tree growing procedure is:

- A set $\mathcal{Q}$ of binary questions.
- A goodness of split criterion $\Phi(s,t)$ that can be evaluated for any split s of any node t.
- A stop-splitting rule.
- A rule for assigning every terminal node to a class.

The splitting and stop-splitting rule

The goodness of split criterion is derived from an impurity function.

An impurity function is a function Φ defined on the set of all K-tuples of numbers $(p_1, \ldots, p_K)$ satisfying $p_k \geq 0$, $k = 1, \ldots, K$, $\sum_{k=1}^{K} p_k = 1$ with the properties

(i) Φ is at maximum only at the point $(\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k})$,

(ii) Φ achieves its minimum only at the points $(1,0,\ldots,0), (0,1,0,\ldots,0), \ldots, (0,0,\ldots,0,1)$,

(iii) Φ is a symmetric function of $p_1, \ldots, p_k$.

Given an impurity function, the impurity measure $i(t)$ is then defined as

$$i(t) = \Phi(p(1|t), p(2|t), \ldots, p(K|t)).$$

If a split s of a node t sends a proportion p_R of the subjects in t to t_R and the proportion p_L to t_L, then the decrease in impurity is defined as

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R).$$

Then the goodness of split $\Phi(s,t)$ is the decrease in impurity $\Delta i(s,t)$. Suppose some splitting has been done and a current set of terminal nodes is made. Denote the current set of terminal nodes by $\tilde{T}$, set $I(t) = i(t)p(t)$, and define the tree impurity $I(T)$ by

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t)p(t).$$

The decrease in tree impurity by splits on t is defined as

$$\Delta I(s,t) = I(t) - I(t_L) - I(t_R).$$

To stop splitting one sets a threshold $\beta \geq 0$, and declare a node t terminal if $\max_{s \in \mathcal{S}} \Delta I(s,t) < \beta$.

The class assignment rule

Suppose a tree T has been constructed and has terminal nodes $\tilde{T}$.

A class assignment rule assigns a class $k \in \{1, \ldots, K\}$ to every terminal node $t \in \tilde{T}$. The class assigned to node $t \in \tilde{T}$ is denoted by $k(t)$.

In particular, we focus on the class assignment rule $k^*(t)$ defined as follows. If $p(k|t) = \max_i p(i|t)$, then $k^*(t) = k$. Using this rule gives that, the resubstitution estimate, $r(t)$ of the probability of misclassification, given that a subject falls into node t, is $r(t) = 1 - \max_k p(k|t)$.
Denote $R(t) = r(t)p(t)$, then the resubstitution estimate for the overall misclassification rate $R^*(T)$ of the tree classifier T is

$$R(T) = \sum_{t \in \tilde{T}} R(t).$$

Initial Tree growing algorithm

At this point when one wants to grow a classification tree one should do as follows:

- Choose a set of binary questions.
- Define the impurity measure $I(t) = i(t)p(t)$ and calculate the decrease in tree impurity

  $$\Delta I(s,t) = I(t) - I(t_L) - I(t_R)$$

  for each split s and terminal node t.

- Stop splitting when $\Delta I(s,t) < \beta$ for all splits and terminal nodes t.
- Assign a class to each terminal node by using the rule: If $p(k|t) = \max_i p(i|t)$, then assign class k to the node.
- Calculate the overall misclassification rate for the tree classifier T by

$$R(T) = \sum_{t \in \tilde{T}} R(t) = \sum_{t \in \tilde{T}} r(t)p(t) = \sum_{t \in \tilde{T}} (1 - \max_k p(k|t))p(t).$$

4.1.3 Methodological development

In the initial tree growing method there are some deficiencies.

Growing right sized trees

The most significant difficulty is that the trees often gives dishonest results. When using a threshold as a stopping rule, the splitting is either stopped too soon at some terminal nodes or continued too far in other parts of the tree. So instead of attempting to stop the splitting at the right set of terminal nodes, continue the splitting until all terminal nodes are very small. Then selectively prune this large tree upward, to get a decreasing sequence of subtrees. Finally use cross-validation or test sample estimates to pick out the subtree that has the lowest estimated misclassification rate. This procedure will be described further in section 4.3.

Splitting rules

Many different criteria can be defined for selecting the best split at each node. In section 4.2 an often used rule is defined. However within a wide range of splitting criteria the properties of the final tree are insensitive to the choice of splitting rule. The criterion used to prune upward is much more important.

4.2 The Gini splitting rule

In the next section, a method is given for selecting the right sized tree assuming that a large tree $T_{\max}$ is already grown. So in this section the

Gini splitting rule which can be used to construct $T_{\max}$ is introduced. Assuming that a set $\mathcal{S}$ of splits at every node t has been specified, then the fundamental ingredient in growing $T_{\max}$ is a splitting rule.

The impurity function given by

$$\Phi(p_1, \ldots, p_K) = \sum_{k \neq j} p(k|t)p(j|t)$$

is known as the Gini index. It can also be written as

$$\Phi(p_1, \ldots, p_K) = \left(\sum_{k=1}^{K} p(k|t) \right)^2 - \sum_{k=1}^{K} p(k|t)^2 = 1 - \sum_{k=1}^{K} p(k|t)^2$$

and has an interesting interpretation: Instead of using the plurality rule to classify subjects in a node t, use the rule that assigns a subject selected at random from the node to class j with probability $p(j|t)$. The estimated probability that the subject is actually in class k is $p(k|t)$. Therefore, the estimated probability of misclassification under this rule is the Gini index

$$\sum_{j \neq k} p(j|t)p(k|t).$$

Note that the Gini index considered as a function $\Phi(p_1, \ldots, p_K)$ is a quadratic polynomial with nonnegative coefficients. Hence, it is concave in the sense that for $r + s = 1,\ r \geq 0,\ s \geq 0$

$$\Phi(rp_1 + sp'_1, \ldots, rp_K + sp'_k) \geq r\Phi(p_1, \ldots p_K) + s\Phi(p'_1, \ldots, p'_K).$$

This ensures that for any split s

$$\Delta i(s, t) \geq 0.$$

Actually, it is strictly concave, so that $\Delta i(s, t) = 0$ only if $p(k|t_R) = p(k|t_L) = p(k|t), k = 1, \ldots, K$.

4.3 Right sized trees and honest estimates

In this section methods to get the right sized tree T and to get more accurate estimates of the true probability of misclassification or of the true expected misclassification rate $R^*(T)$ is provided.

At each step the stepwise tree structure does an optimization over a large number of possible splits. The usual results when resubstitution estimates are used, are much splitting, trees that are larger than the data warrant, and a resubstitution estimate $R(T)$ that is biased downward.

Another problem arises when using a threshold β as a stopping rule. If β is set too low, then there is too much splitting and the tree is too large. Increasing β leads to the following difficulty: There may be nodes t where the decrease in impurity is small, but the descendant nodes t_L, t_R of t may have splits with large decreases in impurity. By declaring t terminal the good splits on t_L or t_R are lost. So instead of using a stopping rule, a more satisfactory procedure can be used. The procedure consisting of two key elements

1. Prune instead of stopping. Grow a tree that is much too large and prune it upward in the right way until you finally cut back to the root node.

2. Use more accurate estimates of $R^*(T)$ to select the right sized tree among the pruned subtrees.

4.3.1 Getting ready to prune

Recall that the resubstitution estimate of the overall misclassification rate is given by

$$R(T) = \sum_{t \in \tilde{T}} r(t)p(t) = \sum_{t \in \tilde{T}} R(t),$$

where $r(t) = 1 - \max_k p(k|t)$.
The first step is to grow a very large tree, $T_{\max}$ by letting the splitting procedure continue until all terminal nodes are either small, contain only identical measurement vectors or contain only subjects of the same class. The size of the initial tree is not critical as long as it is large enough.

To define the pruning process more precisely, call a node t' lower down the tree a descendant of a higher node t if there is a connected path from t to t', and t an ancestor to t'.

A branch T_t of T with root node $t \in T$ consists of the node t and all descendants of t in T. Then pruning a branch T_t from a tree T consists of deleting from T all descendants of t. The tree pruned in this way will be denoted $T - T_t$.

If T' is made from T by successively pruning off branches, then T' is called a pruned subtree of T and denoted by $T' \prec T$. Even for a moderate sized $T_{\max}$ there is an extremely large number of subtrees and an even larger number of distinct ways of pruning up to $\{t_1\}$. A selective pruning procedure is necessary. That is, a selection of a reasonable number of subtrees, decreasing in size, such that roughly speaking, each subtree selected is the best subtree with its size.
The word best indicates the use of some criterion for determining how good a subtree T is. Such a criterion is introduced next.

4.3.2 Minimal cost-complexity pruning

The idea behind minimal cost-complexity pruning is as follows. For any subtree $T \preceq T_{\max}$ define its complexity, $|\tilde{T}|$ as the number of terminal nodes in T. Let $\alpha \geq 0$ be a real number called the complexity parameter and define the cost-complexity measure $R_\alpha(T)$ as

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|.$$

Thus, $R_\alpha(T)$ is a linear combination of the cost of the tree and its complexity.

Now a pruning process that both gives a unique subtree $T \preceq T_{\max}$ which minimizes $R_\alpha(T)$ and that ensures that the nesting $T_1 \succ T_2 \succ \ldots \succ \{t_1\}$ holds, is constructed.

Begin with: The smallest minimizing subtree $T(\alpha)$ for complexity parameter α is defined by the conditions

1. $R_\alpha(T(\alpha)) = \min_{T \preceq T_{\max}} R_\alpha(T)$
2. If $R_\alpha(T) = R_\alpha(T(\alpha))$, then $T(\alpha) \preceq T$.

This definition breaks ties in minimal cost-complexity by selecting the smallest minimizer of R_α.

The jumping-off point for the pruning procedure is in this case not $T_{\max}$ but rather $T_1 = T(0)$. That is, T_1 is the smallest subtree of $T_{\max}$ that satisfy $R(T_1) = R(T_{\max})$. To get T_1 from $T_{\max}$, let t_L, t_R be any two terminal nodes in $T_{\max}$ resulting from a split of the immediate ancestor node t. Breiman et al., [Leo Breiman, 1984] showed that $R(t) \geq R(t_L) + R(t_R)$. If $R(t) = R(t_L) + R(t_R)$, then prune off t_L and t_R. Continue this process until no more pruning is possible. The resulting tree is T_1.

For T_t any branch of T_1, define $R(T_t)$ by

$$R(T_t) = \sum_{t' \in \tilde{T}_t} R(t')$$

where $\tilde{T}_t$ is the set of terminal nodes of T_t.

The minimal cost-complexity pruning works by weakest-link cutting. For any node $t \in T_1$, denote by $\{t\}$ the subbranch of T_t consisting of the single node t. Set $R_\alpha(\{t\}) = R(t) + \alpha$. For any branch T_t, define

$$R_\alpha(T_t) = R(T_t) + \alpha|\tilde{T}_t|.$$

As long as $R_\alpha(T_t) < R_\alpha(\{t\})$ the branch T_t has a smaller cost-complexity than the single node $\{t\}$. But at some critical value of α, the two cost-complexities become equal. At this point the subbranch $\{t\}$ is smaller than T_t, has the same cost-complexity, and is therefore preferable. To find this critical value of α, solve the inequality

$$R_\alpha(T_t) < R_\alpha(\{t\}),$$

getting

$$\alpha < \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1},$$

and since $R(t) > R(T_t)$ for any nonterminal node ([Leo Breiman, 1984]) then $\alpha > 0$.

Define a function $g_1(t)$, $t \in T_1$, as

$$g_1(t) = \begin{cases} \frac{R(t) - R(T_t)}{|\tilde{T}_t| - 1}, & t \notin \tilde{T}_1 \\ +\infty, & t \in \tilde{T}_1. \end{cases}$$

Then define the weakest link $\bar{t}_1$ in T_1 as the node such that

$$g_1(\bar{t}_1) = \min_{t \in T_1} g_1(t)$$

and put $\alpha_2 = g_1(\bar{t}_1)$. The node $\bar{t}_1$ is the weakest link in the sense that as the parameter α increases, it is the first node where $R_\alpha(\{t\})$ becomes equal to $R_\alpha(T_t)$. Then $\{\bar{t}_1\}$ becomes preferable to $T_{\bar{t}_1}$, and α_2 is the value of α at which equality occurs.

Define a new tree $T_2 \prec T_1$ by pruning away the branch $T_{\bar{t}_1}$, that is,

$$T_2 = T_1 - T_{\bar{t}_1}.$$

Now, using T_2 instead of T_1, find the weakest link in T_2. More precisely, letting T_{2t} be any branch of T_2, define

$$g_2(t) = \begin{cases} \frac{R(t) - R(T_{2t})}{|\tilde{T}_{2t}| - 1}, & t \in T_2,\ t \notin \tilde{T}_2 \\ +\infty, & t \in \tilde{T}_2. \end{cases}$$

and $\bar{t}_2 \in T_2$, α_3 by

$$g_2(\bar{t}_2) = \min_{t \in T_2} g_2(t), \qquad \alpha_3 = g_2(\bar{t}_2).$$

Repeat the procedure by defining $T_3 = T_2 - T_{\bar{t}_2}$ and find the weakest link $\bar{t}_3$ in T_3 and the corresponding parameter value α_4.

If at any stage there is a multiplicity of weakest links, then prune away both branches. Continuing this way, gives a decreasing sequence of subtrees

$$T_1 \succ T_2 \succ \ldots \succ \{t_1\}.$$

Starting with T_1, the algorithm initially tends to prune off large subbranches with many terminal nodes. As the tree gets smaller, it tends to cut off fewer at a time.[Leo Breiman, 1984]

4.3.3 The best pruned subtree: an estimation problem

The method of pruning results in a decreasing sequence of subtrees $T_1 \succ T_2 \succ \ldots \{t\}$, where $T_h = T(\alpha_h)$, $\alpha_1 = 0$. The problem is now to select one of these as the optimum-sized tree. The optimum-sized tree is the tree that minimizes the misclassification cost.
If the resubstitution estimate $R(T_h)$ is used as a criterion, the largest tree T_1 would be selected. But if one had an honest estimate $\hat{R}(T_h)$ of the misclassification cost, then the best subtree T_{h0} could be defined as the subtree that minimizes $\hat{R}(T_h)$.

In this section relatively unbiased estimates of the true misclassification cost are constructed.
To study the bias or the standard error of an estimate, a probability model is necessary. Assume as previously that the subjects in $\mathcal{L}$ are N independent draws from the probability distribution $P(A, k)$ on $\mathcal{X} \times \mathcal{C}$, and $(\mathbf{X}, Y)$ is a random sample with distribution $P(A, k)$, independent of $\mathcal{L}$.
In the general case, with variable misclassification costs $C(i|k)$, where $C(i|k)$ is the cost of misclassifying a class k subject as a class i subject, define

(i) $Q^*(i|k) = P(d(\mathbf{X}) = i|Y = k)$ so that $Q^*(i|k)$ is the probability that a subject in k is classified as i by d.

(ii) $R^*(k) = \sum_i C(i|k)Q^*(I|k)$ so that $R^*(k)$ is the expected cost of misclassification for class k subjects.

(iii) $R^*(d) = \sum_{k=1}^{K} R^*(k)\pi(k)$ as the expected misclassification cost for the classifier d.

Both test sample and cross validation provides estimates of $Q^*(i|k)$, $R^*(k)$ and $R^*(d)$. The basic idea is that $Q^*(i|k)$ can be estimated using simple counts of class misclassification. Then $R^*(k)$ and $R^*(T_h)$ are estimated through (ii) and (iii).

Test sample estimates

Select a fixed number $N^{(2)}$ of cases at random from $\mathcal{L}$ to form the test sample $\mathcal{L}_2$. The remainder $\mathcal{L}_1$ form the new learning sample.

The tree $T_{\max}$ is grown using $\mathcal{L}_1$ and pruned upward to give the sequence $T_1 \succ T_2 \succ \cdots \succ \{t_1\}$.
Now take the subjects in $\mathcal{L}_2$ and drop them through T_1. Each tree T_h assigns a predicted classification to each subject in $\mathcal{L}_2$. Since the true class of each subject in $\mathcal{L}_2$ is known, the misclassification cost of T_h operating on $\mathcal{L}_2$ can be computed. This produces the estimate $R^{ts}(T_h)$, called the test sample estimate.

In more detail, denote by $N_k^{(2)}$ the number of class k subjects in $\mathcal{L}_2$. For T any one of the trees $T_1, T_2, \ldots$, take $N_{ik}^{(2)}$ to be the number of class k subjects in $\mathcal{L}_2$ whose by T predicted class is i. Then $Q^*(i|k)$ is estimated as

$$Q^{ts}(i|k) = N_{ik}^{(2)}/N_k^{(2)}.$$

That is, as the proportion of test sample class k subjects that the tree T classifies as i.
The expected cost of misclassification for class k is estimated as

$$R^{ts}(k) = \sum_i C(i|k)Q^{ts}(i|k).$$

For the priors this gives the estimate

$$R^{ts}(T) = \sum_{k=1}^{K} R^{ts}(k)\pi(k).$$

If the priors are data estimated, $\mathcal{L}_2$ can be used to estimate them as $\pi(k) = N_k^{(2)}/N^{(2)}$. In this case the estimate simplifies to

$$R^{ts}(T) = \frac{1}{N^{(2)}} \sum_{i,k} C(i|k)N_{ik}^{(2)}.$$

Using the assumed probability model, it is easy to show that the estimates $Q^{ts}(i|k)$ are biased only if $N_k^{(2)} = 0$ For any reasonable distribution of reasonable sample size, the probability that $N_k^{(2)} = 0$ is so small that these estimates may be taken as unbiased. As a consequence, so are the estimators $R^{ts}(T)$, in fact, in the estimated prior case, there is cancellation and $R^{ts}(T)$ is exactly unbiased.

The test sample estimates can be used to select the right sized tree T_{h0} by the rule

$$R^{ts}(T_{h0}) = \min_h R^{ts}(T_h).$$

After selection of T_{h0}, $R^{ts}(T_{h0})$ is used as an estimate of its expected misclassification cost.

Cross-validation estimates

Another method used to estimate the true misclassification cost is cross-validation. If the sample size in $\mathcal{L}$ is large the test sample estimate should be used, but if it is small cross-validation is the preferred estimation method.
In V-fold cross-validation, the original learning sample $\mathcal{L}$ is divided into V subsets $\mathcal{L}_v$, $v = 1, \ldots, V$, each containing approximately the same number of subjects. The v'th learning sample is

$$\mathcal{L}^{(v)} = \mathcal{L} - \mathcal{L}_v.$$

In V-fold cross-validation, V auxiliary trees are grown together with the main tree grown on $\mathcal{L}$. The v'th auxiliary tree is grown using the learning sample $\mathcal{L}^{(v)}$.
Start by growing V overly large trees $T^{(v)}_{\max}$, $v = 1, \ldots, V$ as well as $T_{\max}$. For each value of the complexity parameter α, let $T(\alpha)$ and $T^{(v)}(\alpha)$, $v = 1, \ldots, V$ be the corresponding minimal cost-complexity subtree of $T_{\max}$ and $T^{(v)}_{\max}$ respectively. For each v the trees $T^{(v)}_{\max}$ and $T^{(v)}(\alpha)$ have been constructed without the subjects in $\mathcal{L}_v$. Thus, the subjects in $\mathcal{L}_v$ can serve as an independent test sample for the tree $T^{(v)}(\alpha)$.

Put $\mathcal{L}_v$ down the tree $T^{(v)}_{\max}$ for $v = 1, \ldots, V$. Fix the value of the complexity parameter α. For every v, i, k define

$N^{(v)}_{ik}$ = the number of class k subjects in $\mathcal{L}_v$ classified as i by $T^{(v)}(\alpha)$,

and set $N_{ik} = \sum_v N^{(v)}_{ik}$, so N_{ik} is the total number of class k test subjects classified as i. The total number of class k subjects in all test samples is N_k, the number of class k subjects in $\mathcal{L}$.

The idea is now that for V large, $T^{(v)}(\alpha)$ should have about the same classification accuracy as $T(\alpha)$. Hence the estimate of $Q^*(i|k)$ for $T(\alpha)$ is

$$Q^{cv}(i|k) = N_{ik}/N_k.$$

For the prior $\{\pi(k)\}$ given or estimated, set

$$R^{cv}(k) = \sum_i C(i|k)Q^{cv}(i|k)$$

and put

$$R^{cv}(T(\alpha)) = \sum_k R^{cv}(k)\pi(k). \tag{4.1}$$

The implementation is simplified by the fact that although α may vary continuously, the minimal cost-complexity tree grown on $\mathcal{L}$ are equal to T_h for $\alpha_h \leq \alpha < \alpha_{h+1}$. Put

$$\alpha'_h = \sqrt{\alpha_h \alpha_{h+1}}$$

so that α'_h is the geometric midpoint of the interval where $T(\alpha) = T_h$. Then put

$$R^{cv}(T_h) = R^{cv}(T(\alpha'_h)),$$

where $R^{cv}(T(\alpha'_h))$ is defined by (4.1).

Now the rule for selecting the right sized tree is: Select the tree T_{h_0} such that

$$R^{cv}(T_{h_0}) = \min_h R^{cv}(T_h).$$

Then use $R^{cv}(T_{h_0})$ as an estimate of the misclassification cost.

4.4 Class probability trees

In some situations, given a measurement vector $\mathbf{x}$, what is wanted is an estimate of the probability that the subject is in class k, $k = 1, 2, \ldots, K$.
In terms of a probability model, suppose that data are drawn from the probability distribution

$$P(A, k) = P(\mathbf{X} \in A, Y = k).$$

Then estimates of the probabilities

$$P(k|\mathbf{x}) = P(Y = k|\mathbf{X} = \mathbf{x}),\ k = 1, 2, \ldots, K$$

are to be constructed. For this type of problem, instead of constructing classification rules, rules of the type

$$\mathbf{d}(\mathbf{x}) = (d(1|\mathbf{x}), \ldots, d(K|\mathbf{x}))$$

with $d(k|\mathbf{x}) \geq 0,\ k = 1, \ldots, K$ and

$$\sum_{k=1}^{K} d(k|\mathbf{x}) = 1, \text{for all } \mathbf{x}$$

is constructed. Such rules are called class probability estimators.
The best estimator for this problem is called the Bayes estimator defined by

$$\mathbf{d}_B(\mathbf{x}) = (P(1|\mathbf{x}), \ldots, P(K|\mathbf{x})).$$

The accuracy of a class probability estimator $\mathbf{d}(x)$ is defined by the value

$$\mathbf{E}\left[\sum_{k=1}^{K}(P(k|\mathbf{X}) - d(k|\mathbf{X}))^2\right].$$

However, this criterion poses a problem, since its value depends on the unknown $P(k|\mathbf{x})$ that is to be estimated. This problem can be solved by putting it into a different setting. Let $\mathbf{X}$, Y on $\mathcal{X} \times \mathcal{C}$ have the distribution $P(A, k)$ and define new variables $Z_k,\ k = 1, \ldots, K$ by

$$Z_k = \begin{cases} 1, & \text{if } Y = k \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\mathbf{E}[Z_k|\mathbf{X} = \mathbf{x}] = P(Y = k|\mathbf{X} = \mathbf{x}) = P(k|\mathbf{x}).$$

Let $\mathbf{d}(\mathbf{x}) = (d(1|\mathbf{x}), \ldots, d(K|\mathbf{x}))$ be any class probability estimator. The mean square error, $R^*(\mathbf{d})$ of $\mathbf{d}(\mathbf{x})$ is defined as

$$\mathbf{E}\left[\sum_{k=1}^{K}(Z_k - d(k|\mathbf{X}))^2\right].$$

Thus, the mean square error of $\mathbf{d}(\mathbf{x})$ is the sum of its mean squared errors as a predictor of the variables $Z_k,\ k = 1, \ldots, K$.
The key identity is that for any class probability estimator $\mathbf{d}(\mathbf{x})$,

$$R^*(\mathbf{d}) - R^*(\mathbf{d}_B) = \mathbf{E}\left[\sum_{k=1}^{K}(P(k|\mathbf{X}) - d(k|\mathbf{X}))^2\right].$$

From this it is seen that among all class probability estimators, $\mathbf{d}_B$ has minimum mean square error. It is also seen that the accuracy of $\mathbf{d}$ differs from $R^*(\mathbf{d})$ only by the constant term $R^*(\mathbf{d}_B)$. Therefore, to compare the accuracy of two estimators $\mathbf{d}_1$ and $\mathbf{d}_2$, the values of $R^*(\mathbf{d}_1)$ and $R^*(\mathbf{d}_2)$ can be compared.
The significant advantage gained here is that $R^*(\mathbf{d})$ can be estimated from data, while accuracy cannot.

The next step is then to use trees to produce class probability estimates with minimal values of R^*.

4.4.1 Growing and pruning class probability trees

Assume that a tree T has been grown on a learning sample $(\mathbf{x}_n, k_n)$, $n = 1, 2, \ldots, N$, using an unspecified splitting rule and has the set of terminal nodes $\tilde{T}$.
Associated with each terminal node t are the resubstitution estimates $p(k|t),\ k = 1, \ldots, K$ for the conditional probability of being in class k given node t. The natural way to use T as a class probability estimator is by defining $\mathbf{d}(\mathbf{x}) = (p(1|t), \ldots, p(K|t))$ if $\mathbf{x} \in t$.

For each subject $(\mathbf{x}_n, k_n)$ in the learning sample, define K values $\{z_{n,i}\}$ by

$$z_{n,i} = \begin{cases} 1, & \text{if } k_n = i \\ 0 & \text{otherwise.} \end{cases}$$

Then the resubstitution estimate $R(T)$ of $R^*(T)$ can be formed as follows. For all $(\mathbf{x}_n, k_n)$ with $\mathbf{x}_n \in t,\ k_n = k$, are

$$\sum_i (z_{n,i} - d(i|\mathbf{x}_n))^2 = (1 - p(k|t))^2 + \sum_{i \neq k} p(i|t)^2 = 1 - 2p(k|t) + S,$$

where $S = \sum_i p(i|t)^2$. Then put

$$\begin{aligned} R(\mathbf{d}) &= \sum_{t\in\tilde{T}} \sum_{k=1}^{K} (1 - 2p(k|t) + S)p(k,t) \\ &= \sum_{t\in\tilde{T}} \sum_{k=1}^{K} (1 - 2p(k|t) + S)p(k|t)p(t) \\ &= \sum_{t\in\tilde{T}} (1 - S)p(t). \end{aligned}$$

Then $1 - S = 1 - \sum_{k=1}^{K} p(k|t)^2$ is exactly the Gini index. So growing a tree by using the Gini splitting rule continually minimizes the re-substitution estimate $R(T)$ for the MSE. As a consequence, the Gini splitting rule is used as the best strategy for growing a class probability tree.

The major difference between classification trees grown using the Gini rule and class probability trees is in the pruning and selection process. Classification trees are pruned using the criterion $R(T) + \alpha|\tilde{T}|$, where $R(T) = \sum_{t\in\tilde{T}} r(t)p(t)$ and $r(t) = 1 - \max_k p(k|t)$ is the within-node misclassification cost.
Class probability trees are pruned upward using $R(T) + \alpha|\tilde{T}|$ but with $r(t) = 1 - \sum_k p(k|t)^2$, the within node Gini diversity index.

The pruning is then done as follows. Grow $T_{\max}$ as before, and prune upward, to get the sequence
$T_1 \succ T_2 \succ \cdots \succ \{t_1\}$. To get test sample estimates $R^{ts}(T)$ of $R^*(T)$ for T any of the T_h, run all the $N_k^{(2)}$ class k subjects in the test sample down the tree T. Define

$$R_k^{ts}(T) = \frac{1}{N_k^{(2)}} \sum_{i=1}^{K} \sum_{n=1}^{N_k^{(2)}} (z_{n,i} - d(i|\mathbf{x}_n))^2,$$

where the sum is over the $N_k^{(2)}$ test sample subjects. Then

$$R^{ts}(T) = \sum_{k=1}^{K} R_k^{ts}(T)\pi(k).$$

Appendix

A.1 The development of the Charlson comorbidity index

This section is a summary of the original article introducing the Charlson comorbidity index, [Charlson, 1987].
The aim for Mary Charlson and colleagues was to develop a method for classifying comorbid conditions which might alter the risk of short term mortality for use in longitudinal studies.

The patients used for development

The training population consisted of all patients admitted to the medical service at New York hospital during a 1-month period in 1984. This resulted in 604 patients being included.
At admission the reason for admission and the severity of the illness were recorded. After discharge the number and severity of comorbid diseases at the time of admission were recorded.
Follow-up information was obtained for 559 of the patients after one year.
Survival was measured as months from the admission date to the date of death or to 1 year after admission. Where follow-up information was not found, the patients were considered as withdrawn alive at the last date of contact with the hospital.

The patients used for validation

The validation population consisted of 685 women with histologically proven primary carcinoma of the breast, who received their first treatment at Yale New Haven Hospital between 1 January 1962 and 31 December 1969.
From medical records a chronology of each patient's illness was compiled. The number and severity of comorbid diseases were also noted.
A 5 year follow-up was obtained for 684 patients and 10 year follow-up information was obtained for 680 patients.
Deaths were then registered as due to either breast cancer or comorbid diseases. To be considered as a comorbid death, the patient must have been free from metastatic diseases at the last examination before the

time of death. Survival was measured in months and calculated from the start of anti-neoplastic therapy to the primary site or if no such treatment was given from the date of the first therapy to a metastatic site.

Classification of comorbidity

All comorbid diseases were recorded. For more common conditions data were collected characterizing the disease severity.

Statistical methods

The relationship of potential important variables to survival in both the training and validation population was assessed using Cox's regression method for life table data.
The proportional hazards analysis was performed using the PHGLM procedure in SAS. The stepwise procedure was used and dummy variables were set up for nominal data. Comorbid diseases were coded as 0,1, severity was codes as 1 to 5, 1 being not ill and 5 being moribund and age was coded in decades.
Unadjusted relative risks were calculated as the proportion of patients with the condition who died divided by the proportion of patients without the disease who died.
The adjusted relative risks were calculated from the beta coefficients generated by the stepwise backward proportional hazards model, as the ratio of those with the disease to those without. The adjusted relative risks estimated the risk of death with a given comorbid disease controlling for all coexistent comorbid diseases as well as illness severity and reason for admission.
In the validation population, survival rates were calculated by the life table method, with cancer deaths handled by regarding the patient as withdrawn alive at the time of death.

The development of the weighted index

A weighted index that takes both the number and the seriousness of comorbid diseases into account was developed. The adjusted relative risks were used as weights for the different comorbid diseases. To simplify the index, diseases with a relative risk of < 1.2 were assigned the weight 0, relative risks of $\geq 1.2 \ < 1.5$ were assigned 1; $\geq 1.5 \ < 2.5$ a

weight of 2; ≥ 2.5 < 3.5 a weight of 3, and those with a relative risk of 6 or more were assigned a weight of 6.
The analysis was performed first with all the diseases with a relative risk of 1.3 or more and secondly with only those diseases which had a significant or near significant independent impact on mortality. The results were similar.

Validation of the index

A comparison between the training and validation data of disease frequencies showed a remarkable low prevalence of the diseases in the validation study. Comparing 1-year survivals for patients with the same index values showed that the validation study had higher 1-year survivals for all index values.
The significance of potential predictors was found, and only age and CCI were significant. The relative risk for increasing level of comorbidity index was 2.3 (95% confidence interval [1.9-2.8]) and for each decade of age it was 2.4 (95% confidence interval [2.0-2.9])

A.1.1 The weights in the Charlson comorbidity index

Disease	Weight
Myocardial infarction	1
Congestive heart failure	1
Peripheral vascular disease	1
Cerebrovascular disease	1
Dementia	1
Chronic pulmonary disease	1
Connective tissue disease	1
Ulcer disease	1
Mild liver disease	1
Diabetes I and II	1
Hemiplegia	2
Moderate to severe renal disease	2
Diabetes with end organ damage	2
Any tumor	2
Leukemia	2
Lymphoma	2
Moderate to severe liver disease	3
Metastatic solid tumor	6
AIDS	6

Table A.1 *The weights from the Charlson comorbidity index.*

A.2 The grouping scheme for the Index of Coexistent Disease (ICED)

Diagnosis	Severity rating
Organic heart disease	1 2 3 4
Ischaemic heart disease	1 2 3 4
Primary arrhythmias and conduction problems	1 2 3 4
Congestive heart failure	1 2 3 4
Hypertension	1 2 3 4
Cerebral vascular accident	1 2 3 4
Peripheral vascular disease	1 2 3 4
Diabetes mellitus	1 2 3 4
Respiratory problems	1 2 3 4
Malignancies	1 2 3 4
Hepatobiliary disease	1 2 3 4
Renal disease	1 2 3 4
Arthritis	1 2 3 4
Gastro-intestinal disease	1 2 3 4

Table A.2 *Physical scale.*

Functional impairment		
Circulation 0 1 2	Respiration 0 1 2	Neurological 0 1 2
Mental Status 0 1 2	Urinary 0 1 2	Fecal 0 1 2
Feeding 0 1 2	Ambulation 0 1 2	Transfer 0 1 2
Vision 0 1 2	Hearing 0 1 2	Speech 0 1 2

Table A.3 *Functional scale.*

Peak intensity of physical scale	Peak intensity of functional impairment	ICED Levels
0	0	0
0	1	0
1	0	1
2	0	1
1	1	2
2	1	2
3+	any(0-2)	3
any(0-4)	2	3
	Total Score	_____

Table A.4 *Grouping System.*

A.3 ICD codes for comorbidities

Disease	ICD-8	ICD-10
Myocardial infarction	410	I21, I22, I23
Congestive heart failure	427.09, 427.10, 427.11, 427.19, 428.99, 782.49	I50, I11.0, I13.0, I13.2
Peripheral vascular disease	440, 441, 442, 443, 444, 445	I70, I71, I72, I73, I74, I77
Cerebrovascular disease	430-438	I60-I69, G45, G46
Dementia	290.09-290.19, 293.09	F00-F03, F05.1, G30
Chronic pulmonary disease	490-493, 515-518	J40-J47, J60-J67, J68.4, J70.1, J70.3, J84.1, J92.0, J96.1, J98.2, J98.3
Connective tissue disease	712, 716, 734, 446, 135.99	M05, M06, M08, M09, M30, M31, M32, M33, M34, M35, M36, D86
Ulcer disease	530.91, 530.98, 531-534	K22.1, K25-K28
Mild liver disease	571, 573.01, 573.04	B18, K70.0-K70.3, K70.9, K71, K73, K74, K76.0
Diabetes I and II	249.00, 249.06, 249.07, 249.09, 250.00, 250.06, 250.07, 250.09	E10.0, E10.1, E10.9, E11.0, E11.1, E11.9
Hemiplegia	344	G81, G82
Moderate to severe renal disease	403, 404, 580-584, 590.09, 593.19, 753.10-753.19, 792	I12, I13, N00-N05, N07, N11, N14, N17-N19, Q61
Diabetes with end organ damage	249.01-249.05, 249.08, 250.01-250.05, 250.08	E10.2-E10.8, E11.2-E11.8
Any tumor	140-194	C00-C75
Leukemia	204-207	C91-C95
Lymphoma	200-203, 275.59	C81-C85, C88, C90, C96
Moderate to severe liver disease	070.00, 070.02, 070.04, 070.06, 070.08, 573.00, 456.00-456.09	B15.0, B16.0, B16.2, B19.0, K70.4, K72, K76.6, I85
Metastatic solid tumor	195-198, 199	C76-C80
AIDS/HIV	079.83	B20-B24, Z21, Z219

Table A.5 *IDC codes for the diseases in the Charlson comorbidity index.*

Disease	ICD-8	ICD-10
Alcohol related disorders	291, 303, 979, 980, 577.10	F10, K860, Z721, R780, T51, K292, G621, G721, G312, I426
History of obesity	277.99	E65, E66
Hypertension	400-404	I10-I15

Table A.6 *IDC codes for the three new diseases.*

A.4 Kaplan-Meier curves

In this section Kaplan-Meier estimates are presented. The section is based on [Frank E. Harrell, 2001] and [Collett, 1994]. The Kaplan-Meier curves can be used to graphically assess the crude survival.

Survival function and hazard function

When summarizing survival data, there are two functions of central interest, namely the survival function and the hazard function. Let T be a random variable denoting the response, that is time until an event. The distribution function of T is given by

$$F(t) = P(T < t) = \int_0^t f(u)du$$

and represents the probability that the survival time is less than some value t.

The survival function, $S(t)$, is defined as the probability that the survival time is greater than or equal to t, so

$$S(t) = P(T \geq t) = 1 - F(t).$$

If the event is death, $S(t)$ is the probability that the subject will survive at least until time t. Since survival theory often is used to model the event death, this terminology will be used from this point on.

The hazard function is the probability that a subject dies at time t, conditioned on the subject having survived until then. The hazard at time t is related to the probability that the subject will die in a small interval around t, given that the subject is alive before time t. The hazard function is defined as

$$h(t) = \lim_{u \to 0} \frac{P(t \leq T < t + u | T \geq t)}{u}.$$

From this definition a useful relationship between the survival and the hazard function can be obtained. Using the law of conditional proba-

bility the hazard function becomes

$$\begin{aligned} h(t) &= \lim_{u \to 0} \frac{P(t \leq T < t+u)/P(T \geq t)}{u} \\ &= \lim_{u \to 0} \frac{(F(t+u) - F(t))/u}{S(t)} \\ &= \frac{\partial F(t)/\partial t}{S(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

Kaplan-Meier estimate

As the true form of the survival distribution is seldom known, it is useful to estimate this distribution without making any assumptions.

Let $S_n(t)$ denote the empirical survival function. $S_n(t)$ is given by the fraction of observed failure times that exceed t

$$S_n(t) = \frac{\text{number of subjects with } T \geq t}{\text{number of subjects in the dataset}}$$

When censoring is present, $S(t)$ can be estimated by the Kaplan-Meier estimator.

To determine the Kaplan-Meier estimate of the survival function from a sample of censored survival data, a series of time intervals is formed. Each of these intervals is constructed so that only one death time is contained in the interval, and so that this death time occurs at the beginning of the interval. For example, suppose that $t_{(1)}$, $t_{(2)}$ and $t_{(3)}$ are three observed survival times and that $t_{(1)} < t_{(2)} < t_{(3)}$. Let c be a censored survival time falling between $t_{(2)}$ and $t_{(3)}$. The constructed intervals then begin at times $t_{(1)}$, $t_{(2)}$ and $t_{(3)}$, and each interval contains only one death time, although there could be more than one subject who dies at a death time. The time origin is denoted by t_0.

In general, suppose that there are n subjects with observed survival times $t_1, t_2, \ldots, t_n$. Some of these observations may be censored, and there may be more than one subject with the same observed survival time. Suppose then, that there are r death times with $r \leq n$. After arranging these death times in ascending order, the j'th is denoted $t_{(j)}$,

for $j = 1, 2, \ldots, r$. The number of subjects who are alive just before time $t_{(j)}$ is denoted n_j, for $j = 1, 2, \ldots, r$, and d_j denotes the number of subjects who die at time $t_{(j)}$. Since there are n_j subjects who are alive just before $t_{(j)}$ and d_j deaths at $t_{(j)}$, the probability of a subject dying in the interval from $t_{(j)} - u$ to $t_{(j)}$ is estimated by d_j/n_j. The corresponding estimated probability of survival through that interval is then $(n_j - d_j)/n_j$.

From the way the time intervals are constructed, the interval from $t_{(j)}$ to $t_{(j+1)} - u$ contains no deaths. The probability of surviving from $t_{(j)}$ to $t_{(j+1)} - u$ is therefore unity, and the joint probability of surviving from $t_{(j)} - u$ to $t_{(j)}$ and from $t_{(j)}$ to $t_{(j+1)} - u$ can be estimated by $(n_j - d_j)/n_j$. As $n \to 0$, $(n_j - d_j)/n_j$ becomes an estimate of the probability of surviving from $t_{(j)}$ to $t_{(j+1)}$. It is now assumed that the deaths of the subjects in the sample occur independently of one another. Then, the estimated survival function at any time in the j'th time interval from $t_{(j)}$ to $t_{(j+1)}$, $j = 1, 2, \ldots, r$, where $t_{(r+1)}$ is defined to be ∞, will be the estimated probability of surviving beyond $t_{(j)}$. This is the Kaplan-Meier estimate of the survival function, which is given by

$$S_{KM}(t) = \prod_{k=1}^{j} \frac{n_k - d_k}{n_k}$$

for $t_{(j)} \leq t < t_{(j+1)}$, $j = 1, 2, \ldots, r$, with $S_{KM}(t) = 1$ for $t < t_{(1)}$.

References

[Andersen et al., 1999] Andersen, T., Madsen, M., Jørgensen, J., Mellemkjær, L., and Olsen, J. (1999).
The Danish National Hospital Register. A valuable source of data for modern health sciences. *Danish Medical Bulletin*, vol. 46:263–268.

[Azzalini, 2002] Azzalini, A. (2002).
Statistical Inference - Based on the likelihood. Chapman & Hall, first edition.

[Charlson, 1987] Charlson, M. E. (1987).
A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *J Chron Dis*, vol. 40 nr. 5:373–383.

[Christensen et al., 2007] Christensen, S., Jacobsen, J., Bartels, P., and Nørgaard, M. (2007).
Hospital standardised mortality ratios based on data from administrative registries. A pilot project. *Ugeskrift for læger*, Aug 20:2767–72.

[Collett, 1994] Collett, D. (1994).
Modelling survival data in medical research. Chapman & Hall/CRC, first edition.

[Cox and Snell, 1989] Cox, D. R. and Snell, E. J. (1989).
Analysis of binary data. Chapman & Hall/CRC, second edition.

[David W. Hosmer, 2000] David W. Hosmer, S. L. (2000).
Applied Logistic Regression. John Wiley & Sons, second edition.

[Dobson, 1990] Dobson, A. J. (1990).
An Introduktion to Generalized Linear Models. Chapman & Hall, first edition.

[Dougherty et al., 1995] Dougherty, J., Kohavi, R., and Sahami, M. (1995).
Supervised and unsupervised discretization of continuous features. *Procedings of the Twelfth International Conference on Machine Learning*, Morgan Kaufmann:194–202.

[Elomaa and Rousu, 2003] Elomaa, T. and Rousu, J. (2003).
On Decision Boundaries of Naive Bayes in Continuous Domains. *7th European Conference on Principles and Practice of knowledge Discovery in Databases*, Springer:144–155.

[Extermann, 2000] Extermann, M. (2000).
Measuring comorbidity in older cancer patients. *European Journal of Cancer*, vol. 36:453–471.

[Frank E. Harrell, 2001] Frank E. Harrell, J. (2001).
Regression Modeling Strategies. Springer, first edition.

[Frank E. Harrell and Lee, 1985] Frank E. Harrell, J. and Lee, K. L. (1985).
A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. *Biostatistics: Strategies in Biomedical, Public Health and Evironmental Sciences*, The Bernard G. Greenberg Volume:333–343.

[Friedman et al., 1997] Friedman, N., Geiger, D., and Goldszmidt, M. (1997).
Bayesian Network Classifiers. *Machine Learning*, 29:131–163.

[Green, 1984] Green, P. T. (1984).
Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *Journal of the Royal Statistical Society, series B (Methodological)*, vol. 46 nr. 2:149–192.

[Halperin et al., 1971] Halperin, M., Blackwelder, W. C., and Verter, J. I. (1971).
Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. *Journal of Chronic diseases*, vol. 24, No 2:125–158.

[Hulley et al., 2001] Hulley, S. B., Cumming, S. R., Browner, W. S., Grady, D., Hearst, N., and Newman, T. B. (2001).
Designing Clinical Research. An Epidemiologic Approach. Lippincott Williams & Wilkins, second edition.

[Leo Breiman, 1984] Leo Breiman, Jerome H. Friedman, R. A. O. C. J. S. (1984).
Classification and regression Trees. Wadsworth, first edition.

[Pregibon, 1981] Pregibon, D. (1981).
Logistic Regression Diagnostics. *The Annals of Statistics*, vol. 9 nr. 4:705–724.

[Press and Wilson, 1978] Press, S. J. and Wilson, S. (1978).
Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Assosiation*, vol. 73, No 364:699–705.

[Tom M. Mitchell (2010),] Tom M. Mitchell (2010).
Machine Learning, Chapter 1, Draft of January 19 2010.
http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf.

[Thomsen et al., 2006] Thomsen, R., Riis, A., Nørgaard, M., Jacobsen, J., Christensen, S., MCDonald, C., and Sørensen, H. (2006).
Rising incidence and persistently high mortality of hospitalized pneumonia: a 10-year population-based study in Denmark. *Journal of International Medicin*, vol. 259:410–417.