



Universiteit
Leiden
The Netherlands

Advances in Survival Analysis and Optimal Scaling Methods

Willems, S.J.W.

Citation

Willems, S. J. W. (2020, March 19). *Advances in Survival Analysis and Optimal Scaling Methods*. Retrieved from <https://hdl.handle.net/1887/87058>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/87058>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/87058> holds various files of this Leiden University dissertation.

Author: Willems, S.J.W.

Title: Advances in Survival Analysis and Optimal Scaling Methods

Issue Date: 2020-03-19

ADVANCES IN SURVIVAL ANALYSIS AND OPTIMAL SCALING METHODS

SANNE J. W. WILLEMS

Advances in Survival Analysis and Optimal Scaling Methods

Author: Sanne J. W. Willems

Cover design: Sanne J. W. Willems & Tonny Hartog (Drukkerij Haveka)

Printing: Drukkerij Haveka

© 2020 Sanne J. W. Willems, Leiden, the Netherlands.

All rights reserved. No part of this publication may reproduced without prior permission of the author. The copyright of the papers that have been published or have been accepted for publication has been transferred to the respective journals.



**Universiteit
Leiden**

Mathematical Institute

ADVANCES IN SURVIVAL ANALYSIS AND OPTIMAL SCALING METHODS

PROEFSCHRIFT

TER VERKRIJGING VAN
DE GRAAD VAN DOCTOR AAN DE UNIVERSITEIT LEIDEN
OP GEZAG VAN RECTOR MAGNIFICUS PROF. MR. C. J. J. M. STOLKER,
VOLGENS BESLUIT VAN HET COLLEGE VOOR PROMOTIES
TE VERDEDIGEN OP DONDERDAG 19 MAART 2020
KLOKKE 15:00 UUR

DOOR

SANNE JOSEPHINA WILHELMINA WILLEMS

GEBOREN TE RAVENSTEIN
IN 1989

Promotor:

Prof. dr. J. J. Meulman (Universiteit Leiden)

Copromotor:

Dr. M. Fiocco (Universiteit Leiden & Leids Universitair Medisch Centrum)

Samenstelling van de promotiecommissie:

Prof. dr. H. Irth (Universiteit Leiden, voorzitter)

Prof. dr. A. W. van der Vaart (Universiteit Leiden, secretaris)

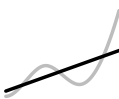
Prof. T. J. Hastie, PhD (Stanford University)

Prof. dr. H. Putter (Leids Universitair Medisch Centrum)

Dr. E. M. L. Dusseldorp (Universiteit Leiden)

CONTENTS

Introduction	7
1 Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator	13
1.1 Introduction	14
1.2 Dependent censoring in survival analysis	15
1.3 Application	21
1.4 Simulation study	25
1.5 Discussion	28
1.6 Supplementary material	32
2 Combining optimal scaling and survival analysis techniques to identify possible predictors for unemployment duration	43
2.1 Introduction	43
2.2 Development of UWV's Work Profiler 1.0	46
2.3 Alternative method to analyze reemployment data	51
2.4 Statistical analysis	54
2.5 Discussion	63
3 Optimal scaling for survival analysis with ordinal data	67
3.1 Introduction	68
3.2 Current practice	70
3.3 Optimal scaling in survival analysis	80
3.4 Simulation study	85
3.5 Discussion	94
3.6 Supplementary material	97



4 Optimal scaling transformations to model nonlinear relations in GLMs with ordered and unordered predictors 101

4.1 Introduction 102

4.2 Optimal scaling in linear regression 104

4.3 Optimal scaling in generalized linear models 107

4.4 Application of GLM with optimal scaling: logistic regression . . . 114

4.5 Discussion 131

4.6 Supplementary material 133

5 Variability in the interpretation of probability phrases used in Dutch news articles - a risk for miscommunication 135

5.1 Introduction 135

5.2 Background 138

5.3 Methods 141

5.4 Results 144

5.5 Discussion 150

5.6 Conclusion 152

5.7 Supplementary material 153

Bibliography 157

Index 165

Summary 167

Samenvatting 179

Acknowledgements 193

Curriculum vitae 195

INTRODUCTION

This thesis is based on five papers on several topics: survival analysis, optimal scaling transformations and statistics communication. Each topic is introduced briefly and then the outline of the thesis is provided.

Survival analysis

In survival analysis the time to the occurrence of an event of interest is studied. This type of analysis is regularly used in medical statistics to estimate the survival time of patients (i.e. time until death). However, it can also be used to estimate other time frames, like recovery time or unemployment duration.

Survival time is defined as the time between a prespecified time origin (e.g. birth, diagnosis, end of employment) and the time of occurrence of the event of interest (e.g. death, recovery, reemployment). One of the main aims in survival analysis is to estimate the *survival function* $S(t)$ which gives the probability that an individual does not experience the event of interest before time t .

A characteristic of survival data is the presence of *censoring*. This occurs when either the time origin or event time are unobserved. For example, if a patient i under treatment moves from one city to another and therefore changes hospital, researchers at the initial hospital will not observe the recovery time x_i . The only information available is that the event had not occurred yet at the patient's last hospital visit. Hence, the observed time for this patient is $t_i = \min(x_i, c_i)$, where c_i denotes the time between the time origin and the last hospital visit. Although less informative than the actual event time, the censoring time is valuable information since it indicates that the recovery time was *at least* more than the time to censoring, i.e. $x_i > c_i$.

Because censoring occurs regularly in survival data, survival analysis methods are designed to also include subjects with unknown event times. In many of these techniques, *independent censoring* is an important assumption. This means that within any subgroup of interest, censored subjects are representative of all individuals who remain at risk of experiencing the event, with respect to their survival experience. This assumption implies that the censored subjects

are randomly selected from the subgroup.

Usually, more information is known about subjects besides their event or censoring time, like gender, age and treatment. A way to compare, for example, the effect of different treatments, is to estimate and compare the survival curves for each treatment. However, in this way, only the categories of one variable can be compared in each analysis.

A more elaborate model which also allows to incorporate several covariates is Cox' proportional hazards model (Cox, 1972). This model focuses on estimating the *hazard function* which gives the rate at which an individual, who has survived until time t , will experience the event in the next instant of time. This hazard function is modeled as

$$h(t|\mathbf{Z}) = h_0(t) \exp \left[\sum_{k=1}^p \beta_k Z_k \right].$$

The survival chances of subjects with covariate values $\mathbf{Z}$ and $\mathbf{Z}^*$ can then be compared by looking at the proportion of their hazards, i.e.

$$\frac{h(t|\mathbf{Z})}{h(t|\mathbf{Z}^*)} = \frac{h_0(t) \exp \left[\sum_{k=1}^p \beta_k Z_k \right]}{h_0(t) \exp \left[\sum_{k=1}^p \beta_k Z_k^* \right]} = \exp \left[\sum_{k=1}^p \beta_k (Z_k - Z_k^*) \right],$$

which is a constant.

Each regression coefficient β_k in Cox' proportional hazards model indicates the change in relative risk. Linearity is assumed for continuous variables, so the regression coefficients of continuous variables indicate the change in the relative risk if the corresponding covariate is increased by one unit. The levels of categorical covariates are represented by dummy variables. Hence, the regression coefficients corresponding to a category level represent the relative risk between that specific level and the reference level.

Optimal scaling transformations

Linear regression is commonly used to model the relation between an outcome variable and a set of predictor variables. In linear regression the outcome is modeled as a linear combination of the predictor variables, i.e.

$$\mathbf{y} = \sum_{k=1}^p \beta_k \mathbf{x}_k + \epsilon.$$

Hence, this model assumes a linear relation between the outcome variable and the set of predictor variables.

However, in many cases the linearity assumption is too strict. A typical example of a nonlinear relation is the relation between the number of accidents caused by a driver and the age of the driver. In general, younger and older drivers cause more accidents than drivers in the middle-age group. Hence, this relation has a u-shape which cannot be captured by standard linear regression.

Since the linearity assumption is too strict, interest has grown in less restrictive models. Model adaptations and extensions that have been developed can be classified into three groups. The first group of nonlinear models is *nonlinear regression* in which the outcome variable is modeled as a nonlinear function of the predictors. The second group (*Generalized Linear Models*, McCullagh and Nelder (1989)) consists of models in which the outcome is modeled as a nonlinear link function on the linear combination of predictor variables. In the third group the predictor variables in the regression model are transformed such that the relation between predictors and the outcome is linearized. In some of these models the outcome is transformed as well.

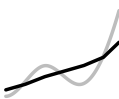
In this thesis we focus on a member of the third group, namely on Optimal Scaling regression (Gifi, 1990; Van der Kooij and Meulman, 1999; Young et al., 1976). This method uses the concepts of linear regression, and finds optimal transformations of the predictor variables while simultaneously estimating the regression coefficients. The aim of optimal scaling is to find optimal numeric values (called *quantifications*) that replace the original predictor values. These quantifications will have numerical properties and linearize the relation between the outcome and transformed predictors. The model is defined as

$$\mathbf{y} = \sum_{k=1}^p \beta_k \varphi_k(\mathbf{x}_k) + \epsilon,$$

where $\mathbf{x}_k$ are the original observed values of predictor k and $\varphi_k(\mathbf{x}_k)$ their quantifications.

A different type of transformation can be chosen for each variable in the model, and this choice depends on which data properties should be preserved. The restrictions applied are specified by choosing a *scaling level*, which can be either nominal, ordinal, nonmonotone spline, monotone spline, or numeric (Meulman et al., 2019).

The nominal and ordinal scaling levels are usually applied to categorical data. The former only preserves the grouping property of a predictor, and the latter preserves both the grouping and ordering properties. For continuous data or categorical variables with many category levels, a smooth spline transformation is more suitable. Spline functions allow for nonlinear transformations which can either preserve the ordering property of the data (monotone spline) or not (nonmonotone spline). If a linear relation is expected, linearity constraints can



be applied by choosing the numeric scaling level; this will give the same result as in linear regression.

Optimal Scaling transformations are not only applied in linear regression but can also be used to introduce nonlinearity in other types of models. For example, these transformations have been introduced in *Principal Components Analysis*. This analysis method reduces the dimensionality of a dataset by summarizing its original variables into a set of linearly uncorrelated variables (the *principal components*) which are linear combinations of the original observed values. PCA with OS transformations is the nonlinear equivalent of PCA (Meulman et al., 2004; Takane et al., 1978).

Statistics communication

Especially with the increasing amount of collected data, statistical results are more frequently used in decision making. Therefore, communicating their findings to decision makers in a clear manner is an important task of statisticians. If the final results are misunderstood, then all the work that was put into collecting the data, developing analysis methods and applying them, is wasted.

Since statistical models are often used to make predictions, clear communication of the estimated probabilities is important. Previous research shows that the persons who communicate estimated probabilities prefer to express these verbally by using probability expressions as *unlikely*, *usually* and *maybe* because these expressions convey some amount of uncertainty (Druzdzel, 1989). This preference indicates that a translation step is needed from the estimated numerical probability to an appropriate verbal phrase. In some cases probability scales are used to standardize this translation step. For example, a probability scale may state that the phrase *very likely* should be used for probabilities in the range of 90–95% and *extremely likely* for 95–99%. Usually these scales are symmetrical. So, if *very likely* represents the range 95–99%, then *very unlikely* indicates 1–5%.

Extensive research has been done on the interpretation of English verbal probability phrases. This research showed that there is huge variation in the interpretation of verbal probability phrases and that interpretation is often asymmetric. For example, the mean perceived percentages of mirrored expressions as *likely* and *unlikely* do not sum to 100% (Lichtenstein and Newman, 1967; Reagan et al., 1989; Stheeman et al., 1993). Additionally, interpretation is influenced by the base rate expectation of the statement in which the phrase is placed. For example, the numerical interpretation of *likely* in the statement “*It is likely that it will rain in Manchester, England, next June*” is usually higher than in “*It is*

likely that it will rain in Barcelona, Spain, next June” (Wallsten et al., 1986).

These research results show that it is impossible to summarize verbal probability expressions into (symmetrical) probability scales that would be supported by everyone. Yet, still many organizations are using scales like this.

Outline of this thesis

This thesis consists of five chapters that are published or submitted papers.

The first chapter focuses on the issue of dependent censoring in survival analysis. It was motivated by a clinical research question from the Department of Psychiatry of Leiden University Medical Center, where it is expected that most patients with anxiety problems stop coming to their appointments when they start to feel better. Therefore, censored patients are not representative for the whole group and the independent censoring assumption is violated. In this chapter, we discuss the *Inverse Probability Censoring Weighted Estimator (IPCW)* and propose a new user friendly algorithm in the statistical software package R (R Core Team, 2018). This algorithm is applied to the data on the anxiety patients from the Department of Psychiatry of Leiden University Medical Center. Furthermore, the performance of IPCW is studied in a simulation study.

The second chapter combines PCA with Optimal Scaling transformations and survival analysis techniques. The research revolves around survey data that were provided by the Dutch Employee Insurance Agency (UWV) in the Netherlands. The dataset contains many categorical and continuous variables that may predict unemployment duration. Nonlinear PCA is first applied to reduce the dimensionality of the data by finding uncorrelated composite scores, which are weighted sums of the original variables. Then the Cox proportional hazards model is fit on the composite scores to study the association between possible predictors and unemployment duration.

In the third and fourth chapter, additional nonlinearity is introduced in GLMs by applying Optimal Scaling transformations on the predictor variables in these models. As a result, a GLM’s link function is no longer applied to the linear combination of predictor variables, but on the weighted sum of their quantifications. In this way nonlinearity is introduced via both the transformations of the variables and the link function. First (chapter 3) this technique is applied to Cox’ proportional hazards model in survival analysis. The aim of combining these techniques is to preserve the ordering in the levels of categorical predictor variables. The model is studied in a simulation study. Next (chapter 4), it is shown how the Optimal Scaling technique can be extended to the family of GLMs. Three different datasets are used to demonstrate the method. All datasets contain a binary outcome variable and a set of categorical and/or

continuous predictor variables. Since the outcome variables in the datasets are binary, the chapter's demonstrations focus on logistic regression. It is shown how the different scaling levels can be applied to the predictors with different properties. It is also discussed how the quantifications of the variables can enhance the visualization and interpretation of the model, which will simplify the communication of the results.

The final chapter continues on the topic of the interpretation of probability expressions. A study was conducted on probability phrases from the Dutch language like *waarschijnlijk* (*probably*) and *misschien* (*maybe*), and frequency phrases as *soms* (*sometimes*) and *doorgaans* (*usually*), of which many have not been studied before. Although extensive research has been done on English expressions, it is important to study them in other languages as well. Namely, many international organizations publish their documents in more than one language and then the meaning of verbal probability expressions may get lost in translation.

CORRECTING FOR DEPENDENT CENSORING IN ROUTINE OUTCOME MONITORING DATA BY APPLYING THE INVERSE PROBABILITY CENSORING WEIGHTED ESTIMATOR

1

Censored data makes survival analysis more complicated because exact event times are not observed. Statistical methodology developed to account for censored observations assumes that patients' withdrawal from a study is independent of the event of interest. However, in practice some covariates might be associated to both lifetime and censoring mechanism, inducing dependent censoring. In this case, standard survival techniques, like the Kaplan-Meier Estimator, give biased results.

The Inverse Probability Censoring Weighted Estimator (IPCW) was developed to correct for bias due to dependent censoring. In this paper we explore the use of IPCW and describe why it is effective in removing the bias. Since implementing this method is highly time consuming and requires programming and mathematical skills, we propose a user friendly algorithm in R. Applications to a toy example and to a medical dataset illustrate how the algorithm works.

A simulation study was carried out to investigate the performance of the IPCW estimator in situations where dependent censoring is present in the data. Different sample sizes, strengths of the censoring model, and percentages of censored individuals were chosen. Results show that in each scenario the IPCW estimator reduces the bias induced in the traditional Kaplan-Meier approach where dependent censoring is ignored.

This chapter is published as Willems, S. J. W., Schat, A., van Noorden, M., and Fiocco, M. (2018). *Correcting for dependent censoring in routine outcome monitoring data by applying the inverse probability censoring weighted estimator*. *Statistical Methods in Medical Research*, 27(2), 323-335.



1.1 Introduction

Survival analysis is the study of the distribution of life times, i.e. the times from a pre-specified initiating event (e.g. birth, diagnosis, start of treatment) to some terminal event of interest (e.g. death, relapse, remission). It is most prominently (but not exclusively) used in the biomedical sciences. A special feature of survival studies is that it takes time to observe the event of interest. As a result, for a number of subjects the event is not observed during follow-up, and the only available information is that the event of interest has not taken place yet at the last observation time. This phenomenon is called *censoring*. Methodologies are developed to include censored subjects in analysis.

Standard methods used to analyse data with censored observations assume that censoring is *noninformative*. This means that censoring carries no prognostic information about the survival experience. Therefore, individuals who are censored at a specific time point should be as likely to experience an event as those subjects who remain in the study, i.e. the probability of being censored is the same for all subjects at risk. *Informative censoring* may occur when time to event and time to censoring are dependent, either directly or through covariates. In the latter case, dependence between event and censoring times is induced through covariates associated to time to event and time to censoring. This type of informative censoring is called *dependent censoring* and it is the focus of this paper. For example, young patients may be more likely to quit a treatment than older patients. This implies that the event is observed more often in older than in young patients, leading to biased estimates of survival probabilities, because the observed event times are not representative for the event times of the whole population.

Since dependent censoring can cause bias in the results, it is crucial to consider this aspect in the analysis. The Inverse Probability Censoring Weighted Estimator (IPCW) was proposed (Robins, 1993; Robins and Finkelstein, 2000; Robins and Rotnitzky, 1992) to correct for the presence of dependent censoring. This method is based on the idea of compensating for censored subjects by giving extra weight to subjects who are not censored. More specific, IPCW assigns extra weight to subjects with similar characteristics as the ones that are censored.

Employing IPCW is very time consuming. Programming and mathematical skills are needed to apply the IPCW procedure, and careful bookkeeping is required. In this paper a user friendly implementation of IPCW in R (Therneau, 2013; Therneau and Grambsch, 2000) is proposed. The algorithm makes clever use of the existing `survival` package (Therneau, 2013; Therneau and Grambsch, 2000). It makes the IPCW method more accessible to researchers.

As illustration, IPCW is applied to a toy dataset and to a real dataset from

the Department of Psychiatry of the Leiden University Medical Center (LUMC) in the Netherlands. A simulation study is performed to compare IPCW with the traditional methodology in case dependent censoring is present in the data. Several scenarios were generated with different sample sizes, percentages of censoring, and strengths of the censoring mechanism. Simulations show that correcting for the presence of dependent censoring by using IPCW is crucial and reduces bias in the estimation results.

This paper is organized as follows. Notation and a short introduction to techniques concerning survival analysis and details about the concept of dependent censoring are outlined in section 1.2. The IPCW method is illustrated and applied to a real dataset in section 1.3. In section 1.4 a simulation study is presented in which the performance of IPCW is investigated. Details concerning the proposed user friendly implementation of the IPCW method in **R** are outlined in subsection 1.6.1, and details concerning the simulation study are described in subsection 1.6.2.

1.2 Dependent censoring in survival analysis

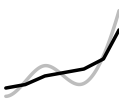
1.2.1 Definition dependent censoring

The aim of survival analysis is to estimate time to a certain event of interest, e.g. death or recovery. However, event times are often incompletely observed, i.e. time to the occurrence of the event of interest is not known for some subjects. This phenomenon is called *censoring*. Survival data for a subject j is represented as triplets $(t_j, \delta_j, [\mathbf{Z}_j(t), 0 \leq t \leq t_j])$, where the observed variable t_j represents the survival time x_j if the event is observed ($\delta_j = 1$), or the censoring time c_j if the event is not observed ($\delta_j = 0$), i.e. $t_j = \min(x_j, c_j)$. The vector of p (time-dependent) covariates is denoted by $\mathbf{Z}_j(t) = (Z_{j1}(t), \dots, Z_{jp}(t))^t$. It is assumed that the value of $\mathbf{Z}_j(t)$ is known at any time point in the subject's observation time. Kaplan and Meier (1958) introduced the *Product-Limit Estimator* as an estimator for the marginal survival function (also referred to as *Kaplan-Meier Estimator*). This estimator is defined as

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t} \left[1 - \frac{d_i}{Y_i}\right] & \text{if } t \geq t_1 \end{cases}, \quad (1.1)$$

where d_i and Y_i are respectively the number of events and the number of individuals at risk at time t_i . This estimator does not consider the covariates $\mathbf{Z}_j(t)$ in the survival curve estimation.

The Kaplan-Meier Estimator assumes that censoring times are noninformative, i.e. that there is no dependence between lifetime X and censoring



mechanism C . Hence, the hazard of X among subjects at risk is the marginal hazard of X , i.e.

$$h_X(t|C \geq t) = h_X(t), \quad (1.2)$$

or, equivalently, the hazard of censoring time C at time t does not depend on event time X

$$h_C(t|X, X > t) = h_C(t|X > t). \quad (1.3)$$

This assumption may, however, be questionable in practice. For example in situations where the covariates history $\bar{\mathbf{Z}}(t) = \{\mathbf{Z}(x); 0 \leq x \leq t\}$ associated with event times also predicts censoring, i.e.

$$h_C(t|\bar{\mathbf{Z}}(t), X > t) \neq h_C(t|X > t). \quad (1.4)$$

A time-dependent Cox proportional hazard model for censoring can be used to test this condition. If (1.4) holds, event time X and censoring time C are dependent through covariates, a phenomenon called *dependent censoring*, and, as a consequence, (1.2) and (1.3) do not hold (Robins and Finkelstein, 2000). In this case, estimators based on the independent censoring assumption, like Kaplan-Meier Estimator, should not be used (Kleinbaum and Klein, 2012). A proper methodology that accounts for dependent censoring should then be employed to avoid biased results.

To illustrate the possible bias when dependent censoring is present in the data under investigation, simulated data in which covariates are associated to both time to event and time to censoring is used to estimate the survival function. Results are given in Figure 1.1 where the traditional Kaplan-Meier estimator and the real survival (which simply is the Kaplan-Meier Estimator in case all events times are observed) are shown together. For details concerning the simulations see subsection 1.6.2.

As expected the Kaplan-Meier Estimator (KM) overestimates the real survival probabilities. This indicates that correction for the presence of dependent censoring is important in order to obtain a good estimator. More examples are shown in the simulation study described in section 1.4.

1.2.2 IPCW

The IPCW Estimator (Robins, 1993; Robins and Finkelstein, 2000; Robins and Rotnitzky, 1992) was developed to correct for dependent censoring. It corrects for censored subjects by giving extra weight to subjects who are not censored. Typically, these weights are chosen in such a way that individuals who best match the censored subjects will receive more weight. At every observed time point t each subject j is given a weight which is inversely proportional to the estimated probability of having remained uncensored until time t . The estimated

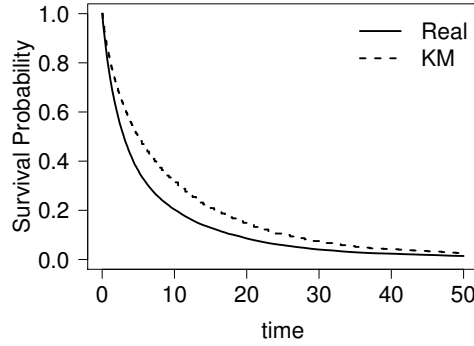


Figure 1.1: Real survival curve and Kaplan-Meier estimate in case dependent censoring is present in the data.

probability is based on the fit of a Cox model for censoring with risk factors for failure and censoring. When a subject is censored at time t_c , individuals who remain at risk should be given extra weight from t_c onward. Hence, IPCW weights have to be recalculated for each subject at risk at each censoring time. This procedure can be summarized in the four steps below.

IPCW to correct for dependent censoring

Step 1: Fit a model for the censoring mechanism that incorporates covariates associated with event and censoring time.

Step 2: Estimate the probability of remaining uncensored at each observed time point t for all subjects at risk at that time point. Denote this estimated probability for subject j at time t as $\hat{K}_j^Z(t)$.

Step 3: Compute the IPCW weights as $\hat{W}_j(t) = 1/\hat{K}_j^Z(t)$.

Step 4: Estimate the survival probabilities $\hat{S}_{IPCW}(t_\tau)$ for time to event in the absence of censoring with subjects weighted according to the IPCW methodology at each observed time point t_τ of interest.

IPCW is based on the assumption that, given $\bar{\mathbf{Z}}(t)$, the hazard of C at time t is independent of the event time X , i.e.

$$h_C(t|\bar{\mathbf{Z}}(t), X, X > t) = h_C(t|\bar{\mathbf{Z}}(t), X > t). \quad (1.5)$$

This means that all covariates that might be associated with event or censoring time must be measured, i.e. there should be no unmeasured confounders. If

this assumption holds, in theory, IPCW can fully correct for bias due to dependent censoring. Therefore, researchers should always gather enough prognostic values that are expected to influence censoring time, such that (1.5) may be approximately true (Robins and Finkelstein, 2000). Even if all prognostic values are included in the model, a crucial step in the IPCW method is to have good estimates for the parameters in the censoring model. Small sample sizes, measurement errors, and other causes for a bad fit for the censoring model may reduce the precision of the estimated weights. However, when the fitted censoring model is accurate enough, the IPCW method will reduce the bias due to dependent censoring.

In the remaining of this section, details concerning the IPCW procedure are outlined. Data for subject j is denoted by $(t_j, \delta_j, \delta_{C_j}, [\mathbf{Z}_j(t), 0 \leq t \leq t_j])$, where t_j is the observed time point, δ_j is the event indicator, $\delta_{C_j} = 1 - \delta_j$ the censoring indicator and $\mathbf{Z}_j(t)$ the vector of covariate values at time t . To implement the IPCW procedure, careful bookkeeping, mathematical and programming skills are required, which makes implementation time consuming. In subsection 1.6.1, a user friendly implementation of IPCW in R (R Core Team, 2018) that uses the `survival` package (Therneau, 2013; Therneau and Grambsch, 2000) is proposed. Toy data in Table 1.1 is used to illustrate how the algorithm works.

id	t	δ	Z
1	18	0	1
2	23	1	2
3	27	0	1
4	32	1	2
5	57	0	3
6	64	1	2

Table 1.1: Toy survival data for 6 subjects. For each subject, the observed time point t , event indicator δ and time-independent covariate Z are given.

Step 1: Fit censoring model

To assess the influence of covariates $\mathbf{Z}(t)$ on the probability of being censored, these covariates are included in the Cox model for time to censoring

$$h_C(t|\bar{\mathbf{Z}}(t)) = h_{C_0}(t) \exp(\boldsymbol{\beta}_C^t \mathbf{Z}(t)), \quad (1.6)$$

where h_{C_0} is the baseline hazard, $\boldsymbol{\beta}_C$ is the vector of model parameters, $\mathbf{Z}(t)$ represents the covariates at time t , and $\bar{\mathbf{Z}}(t)$ the covariate history. Since this is

a standard Cox model, existing methods to fit Cox models can be used for this step.

Fitting the Cox model for time to censoring on the toy data gives a hazard ratio of 0.281 for the covariate Z , with 95% confidence interval $[0.031, 2.538]$. This effect is not significant since no dependence between covariates and censoring times was deliberately included in the data.

id	t_{start}	t_{stop}	δ	δ_C	z	$\hat{K}_j^Z(t_{start})$	$\widehat{W}_j(t_{start})$
1	0	18	0	1	1	1.0000	1.0000
2	0	18	0	0	2	1.0000	1.0000
2	18	23	1	0	2	0.8890	1.1249
3	0	18	0	0	1	1.0000	1.0000
3	18	23	0	0	1	0.6577	1.5205
3	23	27	0	1	1	0.6577	1.5205
4	0	18	0	0	2	1.0000	1.0000
4	18	23	0	0	2	0.8890	1.1249
4	23	27	0	0	2	0.8890	1.1249
4	27	32	1	0	2	0.6827	1.4649
5	0	18	0	0	3	1.0000	1.0000
5	18	23	0	0	3	0.9675	1.0336
5	23	27	0	0	3	0.9675	1.0336
5	27	32	0	0	3	0.8984	1.1131
5	32	57	0	1	3	0.8984	1.1131
6	0	18	0	0	2	1.0000	1.0000
6	18	23	0	0	2	0.8890	1.1249
6	23	27	0	0	2	0.8890	1.1249
6	27	32	0	0	2	0.6827	1.4649
6	32	57	0	0	2	0.6827	1.4649
6	57	64	1	0	2	0.2828	3.5365

Table 1.2: *Estimated IPCW weights for the toy example.*

Step 2: Estimate probabilities of remaining uncensored

The estimated hazard for censoring, $\hat{h}_C(t|\bar{\mathbf{Z}}(t))$ in (1.6) can be used to estimate probabilities of remaining uncensored for each individual at each time point. These probabilities can be estimated by using the analogue of the Kaplan-Meier Estimator for survival probabilities

$$\hat{S}_{KM}(t|\mathbf{Z}_j(t)) = \prod_{\{i; t_i < t, \delta_i = 0\}} [1 - \hat{h}_{C_0}(t_i) \exp(\hat{\beta}_C^t \mathbf{Z}_j(t_i))]. \quad (1.7)$$

The estimated probabilities are denoted by $\hat{K}_j^{\mathbf{Z}}(t)$, where $\mathbf{Z}$ is used to emphasize the dependence on $\mathbf{Z}_j(t)$. Results for the toy example are shown in Table 1.2. Note that the estimated probabilities $\hat{K}_j^{\mathbf{Z}}(t)$ are the same for subjects with equal covariate values.

Step 3: Compute IPCW weights

The weights for each subject j are computed as $\widehat{W}_j(t) = 1/\hat{K}_j^{\mathbf{Z}}(t)$, which is inversely proportional to the estimated probability of remaining uncensored until time t . The estimated IPCW weights for the toy example are given in Table 1.2.

At the end of the study time, when most subjects have experienced the event or have been censored, the probabilities of remaining uncensored become very small. Hence, the IPCW weights will become large and unstable. Weights can be stabilized by dividing the marginal probability of remaining uncensored by the estimated probability of remaining uncensored ($\hat{K}_j^{\mathbf{0}}(t)/\hat{K}_j^{\mathbf{Z}}(t)$).

Step 4: Estimate the IPCW version of the survival curve

In Steps 1-3 it was described how to compute weights for all subjects at risk at each observed time point. By weighting the subjects, a model for time to event in the absence of censoring can be fitted. The Kaplan-Meier Estimator for time to event is adjusted to include the weighted subjects as follows

$$\hat{S}_{IPCW}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t} \left[1 - \frac{\sum_{\{j: \delta_j(t_i)=1\}} \widehat{W}_j(t_i)}{\sum_{k=1}^n R_k(t_i) \widehat{W}_k(t_i)} \right] & \text{if } t \geq t_1 \end{cases} \quad (1.8)$$

The numerator $\sum_{\{j: \delta_j(t_i)=1\}} \widehat{W}_j(t_i)$ is the sum of weights of all subject who experience the event at time point t_i . It represents the estimated number of events that would have occurred at t_i in the absence of censoring. It can be compared to the term d_i in (1.1). The denominator $\sum_{k=1}^n R_k(t_i) \widehat{W}_k(t_i)$ is the sum over weights of all subjects who are at risk at t_i . It represents the estimated number of subjects at risk at t_i in the absence of censoring, as Y_i in (1.1). Note that in case the stabilized weights are implemented, the numerator $\hat{K}_j^{\mathbf{0}}(t)$ is cancelled out in (1.8), i.e. stabilized and unstabilized weights result in the same survival probability estimates.

Survival probabilities estimated with IPCW ($\hat{S}_{IPCW}(t)$) for the toy example are compared to the survival probabilities estimated with standard Kaplan-Meier ($\hat{S}(t)$) in Table 1.3. The difference between these probabilities is very small because there are only few subjects and no dependent censoring was induced in the data.

t	$\hat{S}(t)$	$\hat{S}_{IPCW}(t)$
18	1.000	1.000
23	0.800	0.810
27	0.800	0.810
32	0.533	0.517
57	0.533	0.517
64	0.000	0.000

Table 1.3: Survival probabilities for the toy dataset estimated with standard Kaplan-Meier ($\hat{S}(t)$), and the IPCW version ($\hat{S}_{IPCW}(t)$).

Figure 1.1 showed the poor performance of traditional Kaplan-Meier in the presence of dependent censoring in data. To correct for dependent censoring the IPCW version of the survival curve was estimated on the same dataset (denote by KM_{IPCW}). Results are shown in Figure 1.2 where the classical Kaplan-Meier (KM) is plotted along with KM_{IPCW} and the real survival curve. From this figure we can conclude that IPCW reduced the bias caused by dependent censoring. More simulation results will be shown in section 1.4.

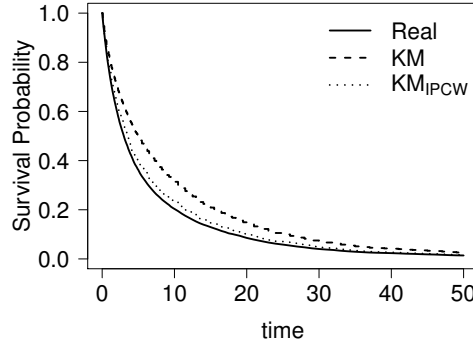


Figure 1.2: Real survival curve and its Kaplan-Meier estimates with (KM_{IPCW}) and without (KM) implementing IPCW.

1.3 Application

1.3.1 Routine outcome monitoring data

IPCW is applied to a dataset from the Department of Psychiatry at the Leiden University Medical Center (LUMC) and Rivierduinen, a local healthcare provider,

in the Netherlands. All patients diagnosed with Diagnostic Statistical Manual-fourth edition- text revision (DSM-IV-TR) mood and/or anxiety disorder and with suicidal ideation were included in the dataset. Suicidal ideation was defined as a score of 2 or higher on item 10 of the Montgomery-Åsberg Depression Rating Scale (MADRS, Montgomery and Åsberg (1979)). Data were collected through a procedure known as Routine Outcome Monitoring (ROM) used to gather information concerning treatment progress by repeatedly measuring symptom severity. The goal is to diagnose patients, and to inform clinicians and patients about the treatment progress. The dataset was used to identify baseline predictors for remission of suicidal ideation (A. Schat et al. *Predicting remission of suicidal ideation in depressed and anxious outpatients: the Leiden routine outcome monitoring study*. Submitted). Remission of suicidal ideation was defined as a score below 2 on item 10 of the MADRS. Socio-demographic variables, functional and clinical scores on several other psychometric instruments were considered as possible predictors for remission. All patients were followed from diagnosis until remission or loss to follow-up, with a maximum of two years.

Standard survival analysis was performed on the dataset of 769 depressed and/or anxious patients with suicidal ideation at baseline. Baseline scores were determined by the Dutch Dimensional Assessment of Personality Pathology-Short Form (DAPP-SF) and the 36-Item Short Form Health Survey (SF-36). DAPP-SF is a short form of the Dimensional Assessment of Personality Pathology - Basic Questionnaire (DAPP-BQ) used for the assessment of personality pathology. SF-36 is a set of quality-of-life measures. To detect baseline covariates with a significant effect on remission of suicidal ideation, the Cox model was fitted. Table 1.4 shows the hazard ratios and the corresponding 95% confidence intervals.

Risk factor	HR (95% CI)
Education level	
- low	1
- high	1.169 (0.990, 1.380)
Depression severity (adjusted MADRS)	0.841 (0.774, 0.913)
Self-harm (DAPP-SF)	0.773 (0.711, 0.841)
General health (SF-36)	0.913 (0.841, 0.992)

Table 1.4: *Hazard ratios for remission of suicidal ideation.*

Covariate *self harm* has the largest effect on remission. To compare survival curves for patients with different levels of self-harm, clinicians defined 3 groups for this variable: *low*, *medium*, and *high*. In Figure 1.3 the estimated Kaplan-Meier curves are shown for these three groups. Recall that the event of interest is remission, hence low survival probability is favorable for patients, since it

indicates a high probability of remission.

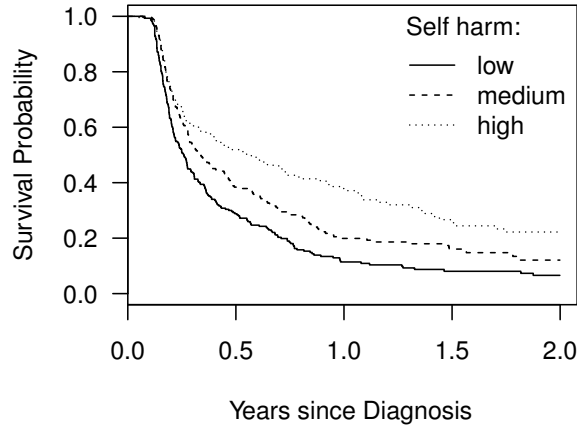


Figure 1.3: *Kaplan-Meier curves for remission from suicidal ideation.*

1.3.2 Correcting for dependent censoring in ROM data

In this study, researchers were only interested in time to remission within the first two years after diagnosis. Therefore, each patient who did not experience the event of interest within two years was censored at 2 years, as is shown in Figure 1.4. This is called *administrative censoring* and was applied to all patients who were still at risk two years after baseline measurements (5.2% of the included patients). Censoring at two years was independent of patient characteristics. Many patients in the ROM dataset (18.2%) were censored during the first two years after baseline measurements. These patients are represented by the steps in the inverse Kaplan-Meier curve for time to censoring (Figure 1.4) during the first two years.

Clinicians believe that patients' withdrawal is likely to be related to their health status. As ROM sessions took place approximately every three months and a final measurement session was not obligatory, it is conceivable that patients who achieved remission, and therefore ended the treatment, did not have a final ROM measurement reflecting their improvement. This suggests a dependence between time to event and time to censoring through health status, i.e. presence of dependent censoring. If true, this would result in overestimation of survival probabilities, since patients who are more likely to experience the event (remission) are also more likely to be censored. Therefore, less events will be observed, and the survival probabilities will be overestimated. IPCW was applied to this dataset to try to correct for dependent censoring. Unfortunately, there is no covariate that directly represents the health status of patients. Instead,

several covariates related to health status were used to fit the censoring model. In this example, we rely on the assumption that patients who do not return for a ROM assessment due to their remission, were close to remission on their last assessment. If this assumption does not hold, it is impossible to observe whether patients that are almost in remission are more likely to be censored than patients with a bad health status.

Since the administrative censoring is independent of patient's characteristics, patients who were censored at two years after diagnosis should not be included in the Cox model for the censoring mechanism. Therefore, the IPCW method described in this section is based on a censoring model for time to non-administrative censoring.

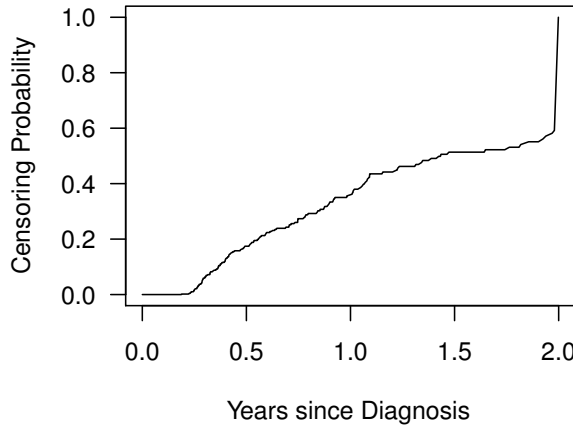


Figure 1.4: *Inverse Kaplan-Meier curves for time to censoring representing the probability of being censored at each time point since diagnosis, given that the subject was not censored until that time point.*

In the censoring model socio-demographic variables and baseline scores considered to be possible predictors for remission were included. In Table 1.5 hazard ratios for time to censoring and their corresponding 95% confidence intervals are shown.

The censoring model to be fitted in step 1 of the IPCW algorithm includes all covariates given in Table 1.5, and those that are significantly associated to time to event (Table 1.4). This censoring model was used to estimate the conditional probabilities of remaining uncensored (step 2), and IPCW weights (step 3). The resulting IPCW survival curve (step 4) is almost identical to the original Kaplan-Meier curve estimated without applying IPCW (not shown). This suggests that the dependent censoring has hardly any influence on the estimated survival probabilities at population level. However, by looking at

Risk factor	HR (95% CI)
Depression severity (adjusted MADRS)	1.201 (0.975, 1.478)
Anxiety severity	0.825 (0.665, 1.024)
Affective lability (DAPP-SF)	0.875 (0.740, 1.035)
Self-harm (DAPP-SF)	1.198 (0.999, 1.436)

Table 1.5: *Cox Model for time to censoring.*

the individual level difference between the prediction for the model with and without IPCW ($|\hat{\beta}_X^{IPCW} Z_j - \hat{\beta}_X^0 Z_j|$), some difference between the survival curves estimated with (KM_{IPCW}) and without IPCW weights (KM) can be found. Results are shown in Figure 1.5. Hence, IPCW does have an impact on individual level. As Figure 1.5 shows, Kaplan-Meier estimator overestimates the survival probabilities for some individuals, which confirms clinicians' experience.

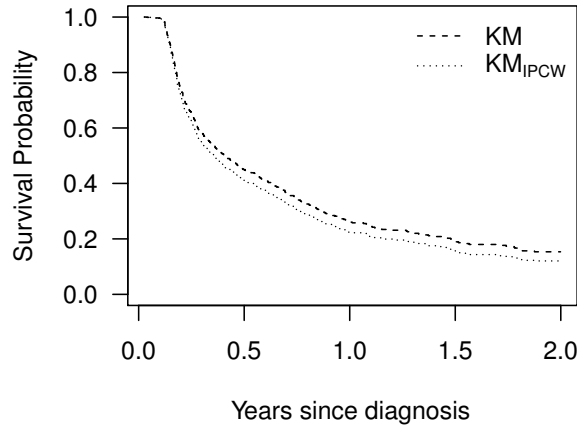


Figure 1.5: *Survival curves estimated with and without the IPCW method for the subject for whom $|\hat{\beta}_X^{IPCW} Z_j - \hat{\beta}_X^0 Z_j|$ is the largest.*

1.4 Simulation study

A simulations study was performed to investigate the behavior of IPCW under different scenarios. In the simulation process different sample sizes, strength of the censoring model, and percentage of censored individuals were chosen. In this section all steps in the simulation process are illustrated and part of the results coming from a large simulation study are discussed.

Steps in the simulations process

Step 1: For each subject j , $j = 1, \dots, n$, generate a continuous variable Z_1 from a standard normal distribution and generate a categorical variable Z_2 from a Bernoulli distribution with $p = 0.5$.

Step 2: Determine the hazards for event and censoring for each individual in the dataset. Hazards depend on covariates Z_1 and Z_2 :

$$h_X(t|\mathbf{Z}) = h_{0_X}(t) \exp(\beta^t \mathbf{Z}), \quad (1.9)$$

$$h_C(t|\mathbf{Z}) = h_{0_C}(t) \exp(\phi^t \mathbf{Z}). \quad (1.10)$$

Step 3: Sample the event times x_j and censoring times c_j from an exponential distribution, with rates $h_X(t|\mathbf{Z}_j)$ and $h_C(t|\mathbf{Z}_j)$ respectively.

Step 4: For each individual compute the observed time point $t_j = \min(x_j, c_j)$ and the corresponding event indicator δ_j .

Step 5: Estimate the true survival curve (Kaplan-Meier on x_j), the standard Kaplan-Meier result (Kaplan-Meier on (t_j, δ_j)), and the IPCW survival estimates (IPCW on (t_j, δ_j)).

All details concerning the method developed to generate survival data with dependent censoring and different scenarios are outlined in subsection 1.6.2.

1.4.1 Varying the sample size

To investigate the effect of sample size n on IPCW results, simulations were performed with different numbers of subjects. The three different sample sizes n were 100, 250, and 500. The other variables, like strength of the censoring model ($\phi = (1.5, 5)$), censoring percentage (35%), and strength of the time to event model ($\beta = (0.5, 1.5)$) remained constant. Results are illustrated in Figure 1.6. As expected, the survival curve estimated by employing IPCW gets closer to the real survival curve for increasing n , while the Kaplan-Meier Estimator's performance does not change when more subjects are observed. The improvement of the IPCW result is due to more precise regression coefficient estimates for the censoring model based on large sample sizes, which results in more accurate IPCW weights, hence in better estimation outcome. Figure 1.6a suggests that IPCW gives better results than standard Kaplan-Meier, even in case of a small sample size.

1.4.2 Varying the strength of the censoring model

The strength of the censoring model is defined by the parameter $\phi = (\phi_1, \phi_2)$, where ϕ_1 is the regression coefficient for Z_1 , and ϕ_2 indicates the effect of the

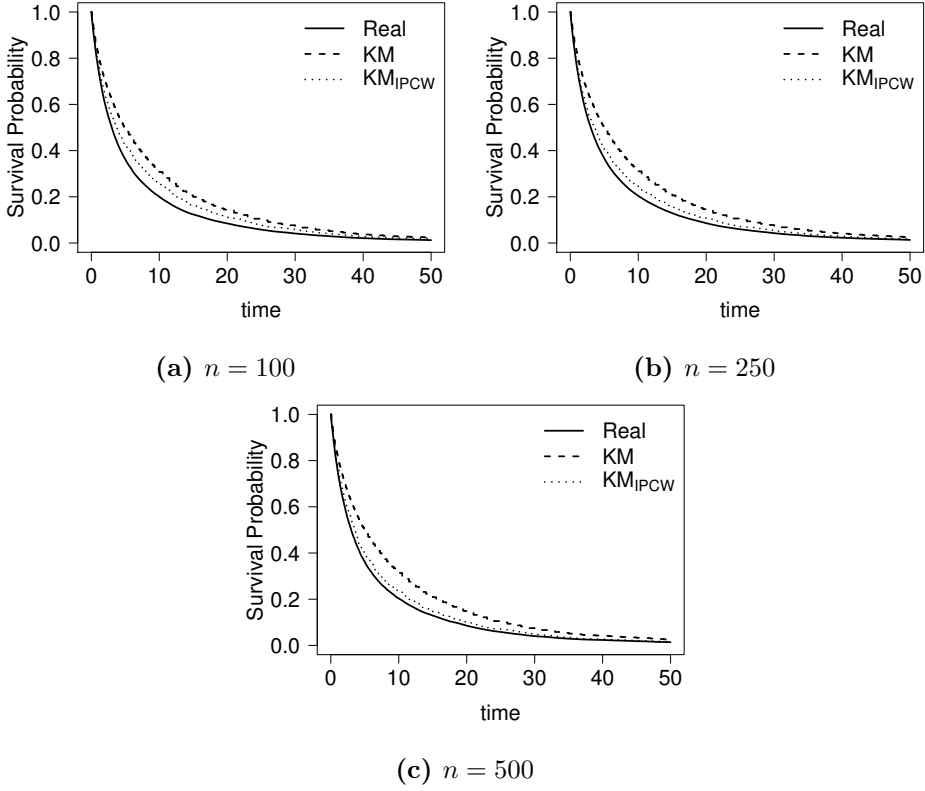


Figure 1.6: True survival probabilities (Real) and the survival curves estimated with standard Kaplan-Meier (KM) and IPCW (KM_{IPCW}) corresponding to different sample sizes n .

categorical variable Z_2 . For larger absolute values of ϕ_1 and ϕ_2 , the dependence between the censoring mechanism and the covariates become stronger. Values of ϕ equal to $(1.5, 5)$ and $(0.75, 2)$ correspond to strong and weak censoring mechanisms respectively. Values of ϕ equal to $(0, 0)$ means absence of dependent censoring.

In Figure 1.7 results corresponding to the different values of the censoring mechanism are shown. In these scenarios β is equal to $(0.5, 1.5)$, the sample size n is equal to 100 and 35% of the subjects were censored. The two covariates Z_1 and Z_2 were generated from a standard normal distribution and a Bernoulli distribution respectively. The first may, for example, represent the ages of subjects in the study and the second one may represent gender (e.g. 0 = male and 1 = female). In this case, the choice of $\phi = (1.5, 5)$ for the strong censoring mechanism indicates that $\exp(5)$ is the risk for women to experience the event compared to men. Furthermore, older individuals have higher censoring

hazards compared to younger individuals. In the presence of weak censoring, the hazard ratio between women and men is equal to $\exp(2)$, and the difference between the hazards of young and old subjects is also smaller. Hence, when weak censoring is present in the data, covariates have a smaller effect on the hazard, i.e. the censoring effect is smaller than in the former situation (strong censoring mechanism).

When considering these strong and weak censoring models, subjects who have a higher probability of experiencing the event also have a higher chance of being censored. Therefore, less events are observed and survival probabilities are overestimated by both standard Kaplan-Meier and IPCW (Figure 1.7b and Figure 1.7a). The stronger the censoring mechanism, the worse the fit for both methods. However, IPCW is less biased than standard Kaplan-Meier in both cases. In the absence of dependent censoring ($\phi = (0, 0)$), both IPCW and standard Kaplan-Meier give accurate results (Figure 1.7c). This suggests that the IPCW method gives good estimates even in case dependent censoring is not present.

1.4.3 Varying the percentage of censored subjects

Varying the percentage of censored subjects while keeping the censoring model constant ($\phi = (1.5, 5)$), can be done by changing the baseline probability of being censored (h_{0C}). Simulations were performed to find h_{0C} corresponding with censoring percentages equal to 35%, 50%, and 65%. In each simulation, $\beta = (0.5, 1.5)$ and sample size is equal to 100. Results in Figure 1.8 show that the performance of both methods becomes worse with increasing censoring percentage. This was expected, since it is known that the precision of the estimate declines with increasing censoring percentage.

1.5 Discussion

Standard survival analysis techniques assume independence between time to event and censoring. This assumption is violated when covariates are associated to both event and censoring time. In case of dependent censoring, traditional methods, like the Kaplan-Meier estimator, may give biased estimates for survival probabilities. IPCW can be used in these situations; this method corrects for censored subjects by giving extra weight to those who remain at risk, and assigning most weight to subjects that are most similar to the censored one.

IPCW was applied to a clinical dataset where patients suffer from suicidal ideation. Time to remission was the event of interest. The IPCW method did not seem to have an effect on estimated survival curves at population level, but it did have an impact on individual level. This result does not necessarily imply

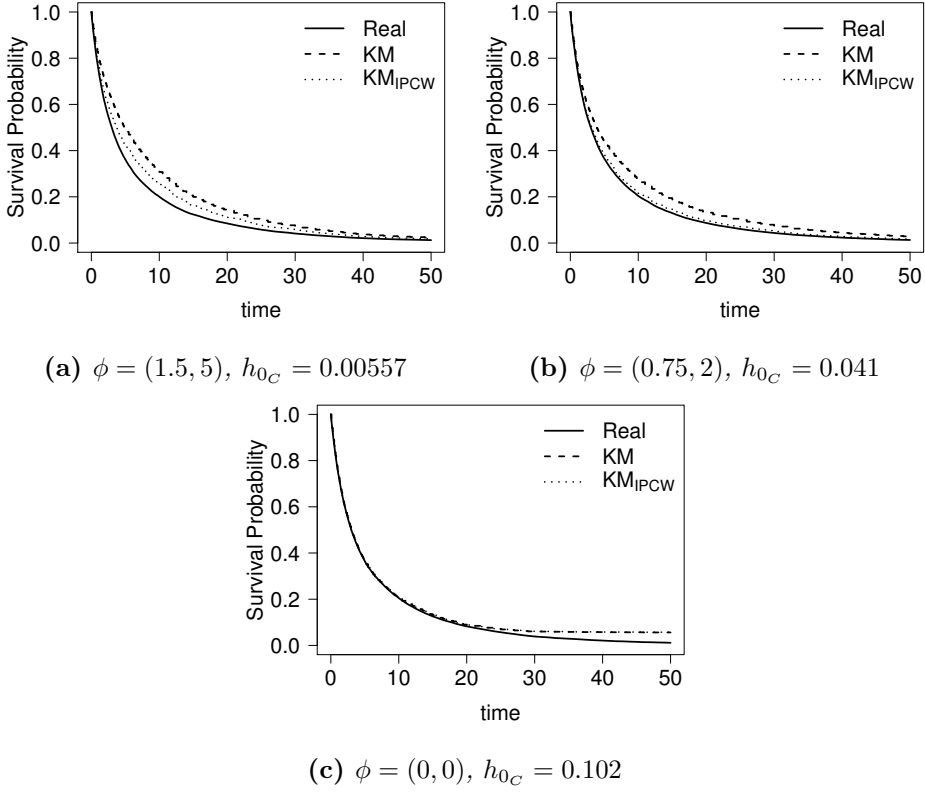


Figure 1.7: True survival probabilities (Real) and the survival curves estimated with standard Kaplan-Meier (KM) and IPCW (KM_{IPCW}) corresponding to different parameters ϕ in the censoring model.

that dependent censoring is not present in the data. There could be unmeasured covariates that influence both time to event and time to censoring, or there could be a mechanism that causes event and censoring times to be directly dependent, i.e. not only through covariates. In these cases, the censoring model estimated with IPCW does not fully describe the censoring mechanism. Therefore, the IPCW results do not completely correct for the censoring mechanism.

A simulation study was carried out to study the performance of IPCW in case of dependent censoring. Dependent censoring was induced in the generated survival data by simulating two time independent covariates that influence both time to event and time to censoring. Several different scenarios were generated by varying the sample size, the strength of the censoring mechanism, and the percentage of censored subjects. The simulation study showed that in each scenario the IPCW estimator performs better than the standard survival technique. The better the fit for the censoring model, the more accurate the

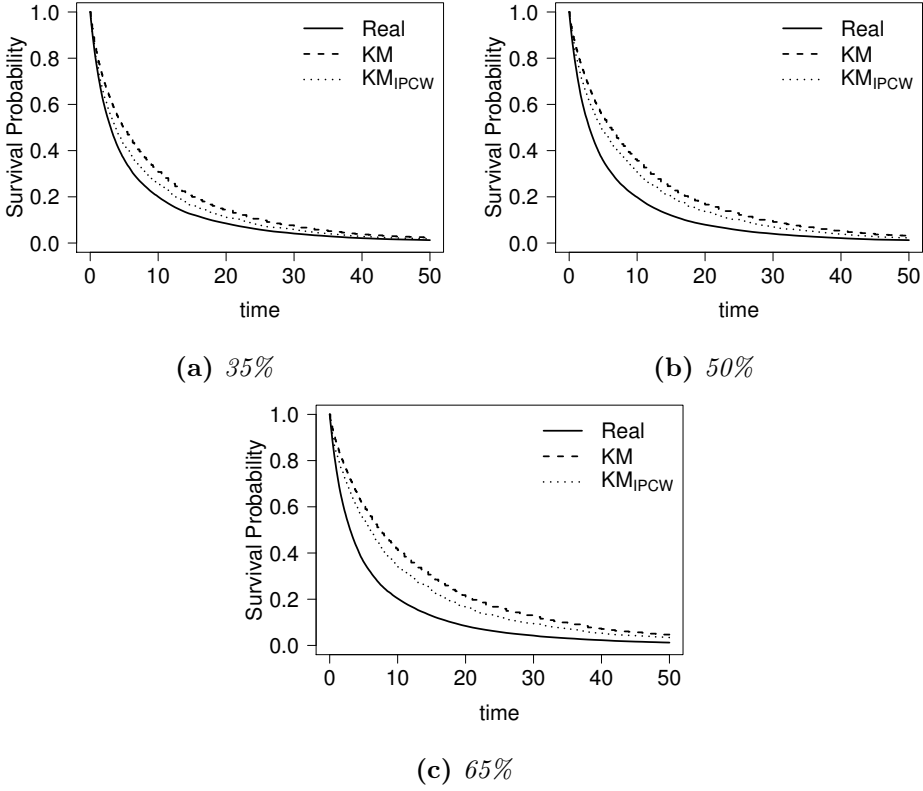


Figure 1.8: True survival probabilities (*Real*) and the survival curves estimated with standard Kaplan-Meier (*KM*) and IPCW (*KM_{IPCW}*) corresponding to different censoring percentages.

IPCW result. In this simulation study, event and censoring times were generated from an exponential distribution, i.e. constant hazard rates were assumed for both models. In practice, hazards may not be constant, but may vary over time.

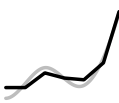
The Corrected Group Prognosis method (CGP) (Chang et al., 1982) might be used as an alternative to IPCW in case of dependent censoring. Here, the Cox model is fit to the whole dataset. Then, survival curves, conditional on the observed covariates, are estimated for each subject in the dataset. The marginal survival curve is then obtained by averaging over the covariate-specific curves. Since dependent censoring does not cause bias in the Cox model, this method will give an unbiased result for the marginal survival curve. The simulation study described in this paper was repeated to compare the performance of CGP with the performance of IPCW (not shown). Simulation results suggest that both methods perform well in case of dependent censoring. However, while IPCW can deal also with time varying covariates (covariates which value may change

over time), CGP cannot. In these situations CGP will not perform well, since it does not incorporate time varying covariates in the estimations of individual survival curves for each subject. Time varying covariates can be included in the Cox model for time to censoring, and can therefore be incorporated in the IPCW methodology.

In the analysis of survival data, attention is mainly given to the survival times and prognostic factors, but almost no attention is given to the censoring mechanism. If the probability of being censored is not the same for each individual at risk, standard survival analysis techniques may give biased results. Further investigation on the censoring mechanism may be needed and IPCW can be applied to adjust for dependent censoring.

Acknowledgements

The Department of Psychiatry of Leiden University Medical Center (LUMC) in the Netherlands is gratefully acknowledged for providing the dataset used in this chapter.



1.6 Supplementary material

1.6.1 A user friendly implementation of IPCW in R

In this subsection a user friendly implementation of IPCW in R is proposed. The IPCW method can be summarized as follows.

IPCW to correct for dependent censoring:

Step 1: Fit a model for the censoring mechanism that incorporates covariates associated with event and censoring time.

Step 2: Estimate the probability of remaining uncensored at each observed time point t for all subjects at risk at that time point. Denote this estimated probability for subject j at time t as $\hat{K}_j^Z(t)$.

Step 3: Compute the IPCW weights as $\hat{W}_j(t) = 1/\hat{K}_j^Z(t)$.

Step 4: Estimate the survival probabilities $\hat{S}_{IPCW}(t_\tau)$ for time to event in the absence of censoring with subjects weighted according to the IPCW methodology at each observed time point t_τ of interest.

Implementing steps 1-4 requires a lot of time and careful book keeping. Programming and mathematical skills are needed. In this appendix, a special way to prepare the data is proposed such that the standard `survival` package in R can be used to employ the IPCW method. In this way, IPCW can be easily used and the method becomes more accessible to researchers with limited mathematical and programming skills.

Data preparation

The IPCW estimator can be used to correct for dependent censoring by giving extra weights to subjects in the risk set. These weights have to be estimated at each observed time point, since they may change at each of these points. The idea is based on a special data format that enables the use of the `survival` package in R to easily estimate the IPCW weights at each of the observed time points. Namely, the proposed long data format includes one row for each weight that should be estimated.

Data for subject j is represented by $(t_j, \delta_j, \delta_{C_j}, \mathbf{Z}_j(t))$, where t_j is the observed time point, δ_j is the event indicator, $\delta_{C_j} = 1 - \delta_j$ the censoring indicator, and $\mathbf{Z}_j(t)$ the (time-dependent) covariates. The censoring indicator can be calculated with R as follows.

```
data$censored <- 1 - data$status
```

where **status** denotes δ_j .

Survival data are usually stored in a one-row-per-subject (wide) format. The step proposed here transforms the data from a wide to a long format. This format will then be used to estimate the function of interest, $S_{IPCW}(t)$.

To obtain the long format, the interval from time origin until event time or censoring time for each individual in the dataset is divided into subintervals. The boundaries of these subintervals are the event and censoring times of all subjects in the dataset, i.e. $t_1, t_2, \dots, t_m$, with m the number of different observed time points, where $t_l < t_{l+1}$, with $l = 1, \dots, m$. For each individual j one extra row for each subinterval $(t_l, t_{l+1}]$ where he is at risk is added to the data. The status values δ and δ_C take value 0 in each of these subintervals until the last one, where δ or δ_C changes into 1 respectively if an event has occurred or if the subject is censored.

By transforming the dataset from wide to long format, a data row is added for each time point at which IPCW weights have to be estimated. This makes it possible to give extra weight to the subjects that remain at risk from the moment that an other subject is censored. In Table 1.6 the long data format corresponding to the toy data set from this paper is illustrated. Cox' Proportional Hazards model can easily be fitted to this data (Andersen and Gill, 1982).

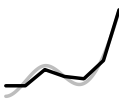
Transforming the dataset from wide to long format can be done by employing **survSplit** from the package **survival**. This function splits each time interval in subintervals with boundaries **times**, and automatically adjusts the event status δ . However, it does not adjust both event and censoring status simultaneously, therefore, the **survSplit** function has to be applied two times, the second time by using δ_C as event indicator.

The data can be transformed by using **survSplit** as follows. First, define a variable that indicates the beginning of an interval, and find all observed time points.

```
data$Tstart <- 0
times <- sort(unique(data$time))
```

Apply the function **survSplit** to the event indicator δ (**event** = "status"), and put in chronological order per subject.

```
data.long <- survSplit(data,
                      cut = times,
                      end = "time",
                      start = "Tstart",
                      event = "status",
                      id = "id")
```



```
data.long <- data.long[
  order(data.long$id,
        data.long$time),
]
```

Repeat for the censoring indicator δ_C (`event = "censored"`) and put in chronological order per subject.

```
data.long.cens <- survSplit(data,
                           cut = times,
                           end = "time",
                           start = "Tstart",
                           event = "censored",
                           id = "id")
data.long.cens <- data.long.cens[
  order(data.long.cens$id,
        data.long.cens$time),
]
```

Add the adjusted censoring indicator (in `data.long.cens`) to the long data format `data.long`.

```
data.long$censored <- data.long.cens$censored
```

It may be convenient to rename some of the columns, e.g. *Tstop* for the end of the intervals.

In case data includes time dependent variables, first data in a long format must be prepared, resulting in extra rows for each observation on a subject. The `unfold` function from the `RcmdrPlugin.survival` package (Fox and Carvalho, 2012) can be used for this. The R code above can then be applied to add extra rows for each time interval. The function `survSplit` will automatically keep track of the values of the covariates in each time interval.

Step 1: Fit censoring model

The censoring model can be fit with a Cox model by employing the `coxph` function in the `survival` package, and by using the censoring indicator δ_C as "event" indicator. In this way, subjects who experience the event of interest are now considered censored. Estimated model parameters $\hat{\beta}_C$ will indicate the effect of covariates $\mathbf{Z}_j(t)$ on the censoring time. R code for fitting the Cox model for time to censoring is provided here below.

id	t	δ	δ_C	z
1	18	0	1	1
2	23	1	0	2
3	27	0	1	1
4	32	1	0	2
5	57	0	1	3
6	64	1	0	2

(a) *Wide format.*

id	t_{start}	t_{stop}	δ	δ_C	z
1	0	18	0	1	1
2	0	18	0	0	2
2	18	23	1	0	2
3	0	18	0	0	1
3	18	23	0	0	1
3	23	27	0	1	1
4	0	18	0	0	2
4	18	23	0	0	2
4	23	27	0	0	2
4	27	32	1	0	2
5	0	18	0	0	3
5	18	23	0	0	3
5	23	27	0	0	3
5	27	32	0	0	3
5	32	57	0	1	3
6	0	18	0	0	2
6	18	23	0	0	2
6	23	27	0	0	2
6	27	32	0	0	2
6	32	57	0	0	2
6	57	64	1	0	2

(b) *Long format.*

Table 1.6: *Survival data in wide (a) and long format (b).*

```
CZ <- coxph(Surv(Tstart,
                  Tstop,
                  censored) ~ cov,
             data = data.long)
```

Step 2: Estimate probabilities of remaining uncensored

The probabilities of remaining uncensored given covariates can be estimated with the Kaplan-Meier Estimator (Kaplan and Meier, 1958)

$$\hat{S}_{CKM}(t|\mathbf{Z}_j(t)) = \prod_{\{i; t_i < t, \delta_i = 0\}} [1 - \hat{h}_{C_0}(t_i) \exp(\hat{\beta}_C^t \mathbf{Z}_j(t_i))]. \quad (1.11)$$

This estimator is implemented in the `survival` package and the probabilities of

remaining uncensored can be estimated by applying the functions `survfit` (with `type = "kaplan-meier"`) and `summary` on the fitted censoring model (CZ).

Recall that the probabilities of remaining uncensored must be estimated for each subject j at each observed time point where the subject is at risk. Hence, by using the data in long format, this probability is estimated as follows.

```
data.long$KZti <- NULL

for(i in 1:nrow(data.long))
{
  datai <- data.long[i,]

  sfiCZ <- survfit(CZ, newdata = datai, type = "kaplan-meier")
  ssfiCZ <- summary(sfiCZ,
                    times = datai$Tstart)
  data.long$KZti[i] <- ssfiCZ$surv
}
```

This step can be very time consuming for datasets with large sample sizes, since the model has to be fit many times for each subject. In case only time-fixed covariates are included in the analysis, it is enough to call the `survfit` function only one time per subject, since specific subject's covariates do not change over time.

```
IDs <- unique(data.long$id)
KZti <- NULL

for(i in IDs)
{
  datai <- subset(data.long, id == i)

  sfiCZ <- survfit(CZ, newdata = datai[1,])
  ssfiCZ <- summary(sfiCZ,
                    times = datai$Tstart)
  KZti <- c(KZti, ssfiCZ$surv)
}

data.long$KZti <- KZti
```

Results for the toy dataset are illustrated in Table 1.2.

Step 3: Compute IPCW weights

The weight corresponding to subject j at time point t is computed as the inverse of the estimated probability of remaining uncensored until time t given covariates $\mathbf{Z}_j$, i.e. $W_j(t) = 1/\hat{K}_j^{\mathbf{Z}}(t)$. These weights can easily be calculated with the R-code below. Results for the toy dataset are shown in Table 1.2.

```
data.long$Weight <- 1/data.long$KZti
```

Step 4: Estimate corrected survival model

The estimated weights should be included in the time to event model, such that subjects who remain at risk are given extra weight in order to correct for the censored subjects. To include weights in the estimation, one could implement the adjusted Kaplan-Meier Estimator (1.8) in R. However, function `survfit` provides the argument `weights`, which enables to weight subjects in the dataset at each observed time point. Therefore, this function can be used to weight the contributions of all subjects in the analysis (R-code below), and there is no need to implement the adjusted Kaplan-Meier Estimator.

```
survNoIPCW <- survfit(Surv(Tstart,
                          Tstop,
                          status) ~ 1,
                      data=data.long)
```

```
survIPCW <- survfit(Surv(Tstart,
                         Tstop,
                         status) ~ 1,
                    data=data.long,
                    weights=Weight)
```

The survival probabilities for time to event at each observed time point estimated with and without IPCW are stored in objects `survIPCW` and `survNoIPCW`. Results for the toy data are given in Table 1.3. These probabilities can be used to plot the survival curves.

As illustrated, it is possible to use the `survival` package to implement all steps of the IPCW procedure. Once data is prepared in a long format, few lines of R-code are needed to apply IPCW on the data.

1.6.2 Simulation study

In this subsection, the Monte Carlo simulation study used to assess the performance of IPCW is outlined. First a short introduction to Monte Carlo

simulations is given. Then the algorithm developed to simulate data including dependent censoring is described. Finally, the IPCW results are compared with standard survival methods (Kaplan-Meier estimation) in several scenarios.

Introduction to Monte Carlo simulations

Mathematical theory assumes that information on all individuals in the population of interest is known. However, in reality only information on a sample of the population is known. Therefore, the method will not give the true parameter θ , but an estimate $\hat{\theta}$. When a different sample from the population is used, a different estimate for the parameter θ will be found. Both estimates are samples from the sampling distribution of $\hat{\theta}$.

Monte Carlo simulations can be used to investigate the sampling distribution of $\hat{\theta}$ (Mooney, 1997). Data is repeatedly drawn from an artificial population and for each sample $\hat{\theta}$ is estimated. The vector of estimates $\hat{\theta}_1, \dots, \hat{\theta}_M$, where M represents the number of Monte Carlo repeats, can be used to estimate the sampling distribution of $\hat{\theta}$. Usually a large number, such as 10,000 or 25,000, is chosen for the value of M .

Monte Carlo procedure:

Step 1: Specify the artificial population, i.e. develop an algorithm to generate data samples.

Step 2: Sample data from the artificial population of interest.

Step 3: Calculate the parameter estimate $\hat{\theta}$ of interest and store in vector $\hat{\theta}$.

Step 4: Repeat Step 2 and 3 M times, where M is the number of trials.

Step 5: Summarize the distribution of $\hat{\theta}_1, \dots, \hat{\theta}_M$ to get an estimate of the parameter of interest.

Data generation

For each subject in the simulated dataset, two covariates were generated, denoted by Z_1 and Z_2 . Covariates Z_1 and Z_2 were sampled respectively from a standard normal distribution, and a Bernoulli distribution with parameter equal to 0.5. For each subject j in the data sample, with $j = 1, \dots, n$, the observed pairs (t_j, δ_j) have to be generated, where $t_j = \min(x_j, c_j)$ and $\delta_j = \mathbb{1}_{\{t_j = x_j\}}$. To obtain the observed data (t_j, δ_j) , time to event x_j and time to censoring c_j for each patient should be generated first.

Exponential distributions were assumed to generate event and censoring times (Bender et al., 2005), i.e. $X \sim \text{Exp}(h_X(t|\mathbf{Z}))$ and $C \sim \text{Exp}(h_C(t|\mathbf{Z}))$, with

hazards depending on the covariates

$$h_X(t|\mathbf{Z}) = h_{0_X}(t) \exp(\boldsymbol{\beta}^t \mathbf{Z}), \quad (1.12)$$

$$h_C(t|\mathbf{Z}) = h_{0_C}(t) \exp(\boldsymbol{\phi}^t \mathbf{Z}). \quad (1.13)$$

Here, h_{0_X} and h_{0_C} are the baseline hazards for time to event and time to censoring, and $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ are the parameter vectors for time to event and time to censoring respectively. It is assumed that both baseline hazards are constant over time, i.e. for all $t \geq 0$, $h_{0_X}(t) = h_{0_X}$ and $h_{0_C}(t) = h_{0_C}$, which results in constant hazard rates h_X and h_C for each subject.

Aim of this simulation study is to assess the performance of IPCW compared to standard Kaplan-Meier in several scenarios. Different sample sizes, strengths of the censoring mechanism, and censoring percentages were used. The sample size is denoted by n . The strength of the censoring mechanism can be changed by taking different values of the parameter ϕ . However, if the parameter ϕ is changed, and all the others remain constant, the percentage of censored subject will change as well. Hence, in order to keep the censoring percentage constant, the baseline hazard h_{0_C} was adjusted for each choice of ϕ . The values of h_{0_C} were obtained with simulation. In Table 1.7 all parameter values used in the simulation study are shown.

	n	h_{0_X}	β	h_{0_C}	ϕ
Sample size n	100				
	250	0.1	(0.5, 1.5)	0.00557	(1.5, 5)
	500				
Strength of censoring model ϕ				0.00557	(1.5, 5)
	100	0.1	(0.5, 1.5)	0.041	(0.75, 2)
				0.102	(0, 0)
Censoring percentage				0.00557	
	100	0.1	(0.5, 1.5)	0.0173	(1.5, 5)
				0.0542	

Table 1.7: *Parameter values chosen in several scenarios for the simulation study.*

The observed data (t_j, δ_j) for patient j with covariates $\mathbf{Z}_j$ can be generated by using the algorithm described below.

For patient j :

Step 1: Generate $u_{j1}, u_{j2} \sim U(0, 1)$.

Step 2: Generate x_j and c_j with the theorem of probability integral transform:

$$\begin{aligned} x_j &= F_j^{-1}(u_{j1}) \\ c_j &= G_j^{-1}(u_{j2}), \end{aligned}$$

where $F_j(t) = 1 - \exp(h_X(t|\mathbf{Z}_j))$ and $G_j(t) = 1 - \exp(h_C(t|\mathbf{Z}_j))$, with $h_X(t|\mathbf{Z}_j)$ and $h_C(t|\mathbf{Z}_j)$ as in (1.12) and (1.13).

Step 3: Derive t_j and δ_j :

$$t_j = \min(x_j, c_j),$$

and

$$\delta_j = \begin{cases} 1 & \text{if } x_j \leq c_j \\ 0 & \text{if } x_j > c_j \end{cases}.$$

Estimating survival curves and model parameters

The aim of this simulations study is to investigate the performance of IPCW in case dependent censoring is present in the data. Results based on IPCW are compared with the standard Kaplan-Meier results by comparing the survival curves estimated by the two methods with the true survival curve.

Real survival curve

The survival model is fit to the generated event times $x_{m1}, \dots, x_{mn}$ in each simulated dataset m , with $m = 1, \dots, M$. The survival probabilities are estimated with the Kaplan-Meier estimator at predefined time points $\tau_1, \dots, \tau_p$, resulting in survival estimates $\hat{S}_m(\tau_1), \dots, \hat{S}_m(\tau_p)$. This approach for estimating the true survival curve represents the situation where underlying parameters are unknown and event times of all subjects in the dataset are observed. Hence, it gives accurate estimates of the true survival probabilities. The estimated survival probabilities for the real survival curve are denoted as $\hat{S}_m(\tau_k)$ for $m = 1, \dots, M$ and $k = 1, \dots, p$. Results based on the M simulations were summarized by taking the average survival probability at each time point, i.e.

$$\bar{\hat{S}}(\tau_k) = \sum_{m=1}^M \frac{\hat{S}_m(\tau_k)}{M}.$$

Standard Kaplan-Meier result

For each sample m the observed data (t, δ) for all patients in the sample is used to estimate the survival curve with standard Kaplan-Meier, which ignores dependent censoring. Results are denoted as $\tilde{S}_j(\tau_k)$, and the M curves were averaged at each time point τ_k , i.e.

$$\bar{\tilde{S}}(\tau_k) = \sum_{m=1}^M \frac{\tilde{S}_m(\tau_k)}{M}.$$

IPCW result

The generated observed data (t, δ) is now used to estimate the survival probabilities with IPCW. The estimated survival probabilities are denoted as $\check{S}_j(\tau_k)$. As before, the survival probability at each time point τ_k is computed as

$$\bar{\check{S}}(\tau_k) = \sum_{m=1}^M \frac{\check{S}_m(\tau_k)}{M}.$$

The three survival curves $\bar{\tilde{S}}(\tau_k)$, $\bar{\check{S}}(\tau_k)$, and $\bar{S}(\tau_k)$ can be illustrated together in one figure to compare the performance of IPCW with standard Kaplan-Meier. Figure 1.9 shows an example of the three survival curves. In this case, the estimated survival obtained with IPCW (KM_{IPCW}) is much more close to the real survival curve (*Real*) than the one obtained with traditional Kaplan-Meier (KM). Based on the results shown in Figure 1.9 we would conclude that IPCW gives a less biased result than Kaplan-Meier.

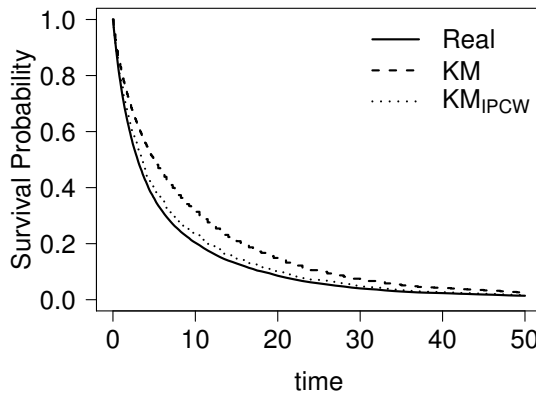


Figure 1.9: *Real survival curve (Real) and its Kaplan-Meier estimates with and without implementing IPCW (KM_{IPCW} and KM respectively).*

COMBINING OPTIMAL SCALING AND SURVIVAL ANALYSIS TECHNIQUES TO IDENTIFY POSSIBLE PREDICTORS FOR UNEMPLOYMENT DURATION

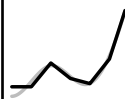
2

In this paper we propose a new approach in survival analysis in a non-traditional field of application; unemployment data. A common practice is to use factor analysis to first summarize survey data, and then fit a binomial logistic regression model to estimate the regression weights, which are used to identify factors associated with work status at a prespecified time point. In this paper, a combination of optimal scaling and survival analysis methods is proposed as an alternative to find possible predictors for unemployment duration. This combination of techniques is illustrated and compared to the traditional approach. Data from the Dutch Employee Insurance Agency are used to illustrate the method.

2.1 Introduction

Identifying possible predictors for unemployment duration can be useful to, for example, provide appropriate counseling or predict the costs for unemployment benefits. To identify these predictors, data on job seekers are, among others, collected via questionnaires and registries, and these are used to find associations between characteristics of the unemployed and their probability of finding a new job. A data preparation step is usually performed to summarize many survey items into fewer composite scores. These resulting scores can be used to model unemployment duration or employment status. An abundance of methods has been developed to summarize survey data and to model employment status. In this paper, the popular choice of combining factor analysis with logistic

This chapter is published as Willems, S. J. W., Fiocco, M., and Meulman, J. J. (2019) *Combining optimal scaling and survival techniques to identify possible predictors for unemployment duration*. International Journal of Statistics & Economics, 20(3), 1–22.



regression will be discussed and an alternative combination, optimal scaling principal component analysis with survival analysis, will be proposed.

The aim of a data preparation step is to summarize many items into a few composite scores. Factor analysis seems to be a popular choice for this data preparation step. This method is used to obtain a composite score for items with a common underlying factor. Each item within a factor is given a factor loading, which indicates the contribution of this item to the factor. These factor loadings are then used as weights to compute the weighted average of the factor items, the so-called factor score. Often, it is predefined which items have the same underlying factor, and that each item has just one underlying factor. Hence, each factor is a summary of a group of items, and each item belongs to one factor. This makes interpretation of the factor model straightforward, and therefore a popular choice. When factor analysis is used as a data preparation step, the resulting factor scores are used in a statistical model to identify possible predictors. Including only the factor scores instead of all individual items reduces the correlation between the predictors considerably. As a consequence, the problem of collinearity can be controlled. However, correlation between the variables is only reduced by factor analysis, but not eliminated completely since there might still be high correlation among the resulting factor scores. Furthermore, this method assumes the items are numeric, so it works optimally for numeric items. Ordinal category levels are coded by integers, which are often treated as numeric data by the model. In this way, a linear relation between the categorical item scores and the factor score is assured. However, since the category levels cannot be assumed to be equally spaced, a nonlinear relation possibly gives a better fit for this type of data.

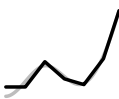
As an alternative to factor analysis, optimal scaling principal component analysis (OS-PCA) can be applied to summarize items into fewer summary scores. This technique can maintain the properties of ordinal categorical data, while finding composite scores that are completely uncorrelated. This method is also often referred to as nonlinear principal component analysis (NLPCA) (Linting et al., 2007; Meulman et al., 2004). The purpose of OS-PCA is to reduce the dimensionality of a dataset by summarizing the original variables into a smaller set of uncorrelated variables, called principal components. In this way, all collinearity between the composite scores is removed.

For each component, items get component loadings, which are used to calculate object scores for each component; the weighted average of the items. Only the most important components (the principal components) are included in the final model. The chosen number of components indicates the new dimensions of the data. Since all components are a weighted average of all items, interpretation of OS-PCA results is more challenging than the interpretation of a factor model in which each factor underlies a subset of the items.

Additionally to finding principal components, OS-PCA provides the option to optimally transform the category levels by giving them new values, called quantifications. The quantifications are chosen in an optimal way such that the nonlinear relationship is linearized. This option can be used to maintain the properties of ordinal data. While constructing principal components, the OS-PCA method aims to account for as much of the variance in the original dataset as possible. As a result, the principal components will also reveal the correlation structure between items, and thus provide a better understanding of how items in the survey are related. Since OS-PCA will remove the collinearity and will maintain the ordinal properties of the survey data, it might be a valuable alternative to factor analysis. The linear version of OS-PCA, PCA, has already been applied in the context of reemployment data (Wanberg et al., 2002).

Once the item data are summarized into composite scores, the next step is to identify variables associated with the probability of reemployment. For this step of the analysis, a binomial logistic regression analysis is a popular approach. This technique can be used to estimate the probability of reemployment within a prespecified time period, for example, to predict whether a person is reemployed within one year. This model is very useful if researchers are interested in the probability of an event within a specific time period. However, the binary outcome and prespecified time point might be too restricting if interest lies in the estimation of the reemployment probability over a time interval. In this case, techniques from the survival field can be used instead of a binomial logistic regression model. Survival analysis techniques estimate the time to an event of interest based on a set of variables. Hence, for reemployment data, survey results can be used to assess the probability of being reemployed over a range of time, instead of at one specific time point. Survival analysis techniques are often used in a medical setting to, for example, compare the effect of different treatments on the the survival time of patients. There are several instances where these techniques have been used in the field of reemployment prediction (Boršič and Kavkler, 2009; Kavkler et al., 2009; Tutkun and Karasoy, 2016; Wanberg et al., 2002), but it is not widely used. Logistic regression appears to be the default method.

In this paper, the two combinations of methods will be introduced in more detail, and they will be illustrated with an application on reemployment data from the Dutch Employee Insurance Agency (from hereon referred to by the Dutch abbreviation UWV). The paper is organized as follows. Details on the UWV dataset are provided in section 2.2, together with some details on the factor and logistic regression analysis results. The OS-PCA and survival analysis techniques are discussed in section 2.3. Results of the application of these techniques on the UWV data are shown in section 2.4. Advantages and disadvantages of the proposed methodology are discussed in section 2.5 where these methods are



compared to factor analysis combined with logistic regression.

2.2 Development of UWV's Work Profiler 1.0

In the Netherlands, job seekers who recently became unemployed can apply for unemployment benefits and counseling at the UWV agency. Since the budget of this institute is limited, it aims to reduce counseling service expenses. One of the strategies is to reduce face-to-face counseling by replacing it with computerized services. For this replacement, UWV has developed an online instrument, the Work Profiler (Wijnhoven and Havinga, 2014). The purpose of this instrument is to select those individuals who experience difficulties in finding a job and could therefore benefit from unemployment counseling. Once these individuals are selected, the Work Profiler should additionally provide a quick diagnosis of the main obstacles faced by these persons, such that appropriate services can be provided for each individual. To make the selection and to give the diagnosis, the Work Profiler makes use of an online questionnaire given to the unemployed persons. The replacement of the selection and diagnosis procedures by an online tool will greatly reduce the costs for face-to-face counseling.

The UWV Centre for Knowledge (Kenniscentrum UWV) and the School of Medical Sciences of the University Medical Centre Groningen (UMCG) in the Netherlands collaborated to develop the first version of the Work Profiler (Brouwer et al., 2015). Aim of this research was to identify possible baseline predictors for resuming work within 12 months after becoming unemployed. It consisted of three steps: a literature study, a cross-sectional study and a longitudinal study. First, many factors that possibly influence the probability of finding a new job were listed from literature. A 500-item questionnaire was created with items corresponding to these factors. Then, during the cross-sectional study the questionnaire's length was reduced to 155 items, each corresponding to one of the 70 remaining factors found to be most relevant. These 155 items were included in the survey used in the third step, the longitudinal study. Newly unemployed were asked to fill in this survey, and after 12 months their work status was registered. The mean scores of the items corresponding to each factor were included as variables in a logistic regression analysis. This analysis identified 10 predictive factors associated with work resumption within one year. An additional 11th factor was retained in the model at UWV's request. All items corresponding to these 11 predictors were included in the first version of the Work Profiler, resulting in a survey consisting of 20 items. This version has been used since 2013 as part of UWV's services.

Since the realization of Work Profiler 1.0, a new study has started to further develop this instrument. Work Profiler 1.0 was extended to include more factors

and used to collect data for the second longitudinal study. In the remaining of this section, the data collection procedure for this new study will be described in more detail and the first analyses on this data will be discussed.

2.2.1 Data collection

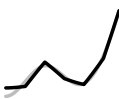
Participants

The data used for this research is from job seekers associated with one of the 11 participating UWV offices. The cohort of this study includes persons younger than 64 years who claimed unemployment benefits between March 1st 2014 and October 31st 2014. Individuals should be eligible for at least three months of benefits. Six to ten weeks after the start of their unemployment benefits, the extended version of Work Profiler 1.0 (in Dutch) was provided via UWV's online system. Hence, only persons who had access to the online system could participate (almost everyone eligible for unemployment benefits). Unemployed whose unemployment benefits ended within 10 weeks, e.g. due to being reemployed or due to other reasons, were excluded from the analysis. The data consists of 32,623 observations.

Factors and items

Since the aim was to further develop the Work Profiler 1.0, all 20 items of the 11 factors from this version were included in the new one. To find more predictors for unemployment duration, items corresponding to other factors were added as well. This resulted in a total of 32 factors (15 hard factors and 17 soft factors) measured by 55 items. Most hard factors, like age, gender, and education were already available from UWV's registries, while the majority of the soft factors, which indicate a person's psycho-social situation in relation to reemployment, was measured by the questionnaire. All hard factors are displayed in Table 2.1, and all soft factors and the number of corresponding items are given in Table 2.2. The classification of the items into factors is based on the study by Brouwer et al. (2015).

All soft factors are ordinal categorical variables, i.e. there is an ordering among the possible answer options. Most of the soft factors were measured on 5-point Likert scales. For example, to indicate to what extent newly unemployed agreed with a statement, the provided options were as follows: *strongly disagree*, *disagree*, *don't disagree/agree*, *agree*, or *strongly agree*. Most of the remaining items had similar or comparable scales, like a 5-point scale to indicate the importance of an aspect, or a 1–10 scale to grade certain aspects of life. Also, many hard factors were measured on an ordinal scale. Exceptions were factors like age, nationality, and industry.



Hard factor	Measurement level	# categories
Additional income - work	nominal	2
Additional income - benefits	nominal	2
Age group	ordinal	5
Duration last job (years)	numeric	-
Education level	nominal	11
Extend of unemployment (hours)	numeric	-
Former working hours	numeric	-
Resigned themselves (fraction)	numeric	-
Gender	nominal	2
Household position	nominal	6
Industry	nominal	12
Maximum duration of benefits (weeks)	numeric	-
Nationality	nominal	4
Profession level	ordinal	5
UWV office	nominal	11

Table 2.1: *Hard factor candidates for Work Profiler 2.0, with their measurement levels and number of categories (if applicable).*

The category frequencies for some of the measured covariates are given in Table 2.3.

Unemployment duration

The aim of further developing Work Profiler 1.0 is to identify possible predictors for unemployment duration and to make good predictions for the probability of being reemployed. Therefore, for each person in the dataset the starting dates of unemployment and reemployment (if applicable) were registered. Starting dates of unemployment were specified by the unemployed when they claimed their benefits. If applicable, the dates of reemployment were either provided by the persons claiming benefits, or retrieved by UWV. If a person was reemployed while being eligible for the unemployment benefits, these benefits were stopped, and hence the reemployment date was registered in UWV's administration on unemployment benefits. If a person did not find a job while being eligible for unemployment benefits, the date of reemployment was checked in the POLIS registry in which all gainful employment in the Netherlands is registered. In this way, for each person in the dataset it was determined whether he/she had found a job within one year. The duration of unemployment was determined from the starting dates of unemployment and reemployment.

Soft factor	# of items	# categories
Acceptance readiness - full time	1	5
Acceptance readiness - time	2	5, 5
Acceptance readiness - work	2	5, 5
External variable attribution	3	5, 5, 5
Childcare problems	1	5
Financial need/problems	2	3, 5
Hours per week capable of work	1	5
Job search attitude - advantageous / pleasant	2	5, 5
Job search attitude - utility / necessity	2	5, 5
Job search behavior - applications	1	4
Job search behavior - direct contact employers	3	4, 4, 4
Job search intention	3	5, 5, 5
Perceived health	5	5, 5, 5, 5, 10
Self-efficacy - preparation applications	3	5, 5, 5
Subjective norm	1	14
Balance pros and cons for not working	1	4
View return to work	3	5, 5, 5

Table 2.2: Ordinal soft factors candidates for Work Profiler 2.0, with the number of items used to assess them and the number of answer categories for each of these items.

Covariate and category levels	# observations
Age	
- ≤ 27	4,213
- > 27 and ≤ 39	9,669
- > 39 and ≤ 49	8,777
- > 49 and ≤ 54	4,084
- > 54	5,880
Gender	
- male	15,641
- female	16,982
Household position	
- Living alone	6,024
- Married / cohabiting, no children	8,440
- Married / cohabiting, youngest child ≥ 7 years	7,330
- Married / cohabiting, youngest child < 6 years	5,832
- Single parent	2,308
- Other	2,689
Profession level	
- Elementary	2,607
- Lower	9,290
- Middle	12,443
- Higher	7,005
- Academic	1,227
- <i>Missing</i>	51
Nationality	
- 1st Dutch, 2nd no or western	31,220
- 1st or 2nd Polish	406
- 1st or 2nd non-western	349
- 1st western (other countries)	619
- <i>Missing</i>	29

Table 2.3: *Number of observations in each category of some of the covariates in the dataset, $n = 32,623$.*

2.2.2 Statistical analysis to develop Work Profiler 2.0

The Netherlands Organisation for applied scientific research (from hereon referred to by the Dutch abbreviation TNO) has conducted an extensive analysis of the data collected by UWV. The results of this analysis were described in an internal report which is not yet published. In this section a short description of the analysis performed by TNO will be given.

Recall that Brouwer et al. (2015) used the mean of all items corresponding to a factor as the factor score. TNO extended the data preparation step by performing a factor analysis on the items corresponding to the soft factors. Restrictions were put on the factor model to enforce some assumptions about the model. For example, the factor classification of each item was prespecified according to the research by Brouwer et al. (2015) and correlations between the factors were allowed for in the factor model, since it was expected that some of the factors are related. The factor loadings resulting from this analysis were used as weights to compute the factor scores; the weighted means of the items corresponding to each factor.

A univariate logistic regression model was fit for each factor to study the association with work status after one year. All relevant factors were included in the multivariate logistic regression model, either as a numeric variable or as a categorical variable. Then, by using a combination of forward and backward selection, the variables were removed from the model to derive a parsimonious model with an easy interpretation. TNO's final model consists of 18 factors.

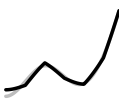
In order to get more insight in the probability of finding a new job during the first year, logistic regression models were fitted for work status at six and nine months in a similar way as for 12 months.

2.3 Alternative method to analyze reemployment data

As discussed in the introduction, a combination of OS-PCA and survival analysis to identify possible predictors for unemployment duration are introduced as an alternative to the combination of factor analysis and logistic regression.

2.3.1 Optimal scaling principal component analysis

The extended version of Work Profiler 1.0 consists of 55 items intended to measure 32 factors. Many of these factors are closely related, which implies their factor scores to be correlated. The correlation between scores is a common phenomenon in survey data and may lead to problems in the estimation of a statistical model due to the presence of collinearity. For example, TNO found two pairs of hard factors to be collinear. To prevent the problem of collinearity,



it is preferable to include only weakly correlated or uncorrelated variables in such models. Some data preparation procedures can help to summarize items in less correlated scores.

Principal component analysis (PCA) was developed to reduce the dimensionality of a dataset by summarizing numeric variables into a smaller set of uncorrelated summary variables, the principal components (Jolliffe, 2002). While constructing the principal components, the PCA method aims to account for as much variance of the original variables in the dataset as possible. So, the resulting components are uncorrelated, but still retain much of the correlation between the original item scores. In this way, the method reveals the correlation structure of the original variables.

Since the PCA algorithm is based on calculating correlations, it can only be applied to numeric data. To deal with categorical variables, optimal scaling principal component analysis (OS-PCA) was developed. This technique transforms categorical variables into numeric variables (quantifications) while simultaneously calculating the principal components. OS-PCA uses the quantifications to calculate correlations. Details of this method were described by Meulman et al. (2004), Linting et al. (2007), and Linting and van der Kooij (2012). OS-PCA is currently available in the Categories package (Van der Kooij and Meulman, 1999) of the statistical software SPSS (SPSS Inc., 2008).

Several restrictions can be put on the OS transformation. For example, to preserve the ordering of ordinal category levels, one would choose a monotone transformation. As output, OS-PCA provides the quantifications of the categorical variables. Furthermore, correlation between each item and each component are given by component loadings. These loadings indicate how well items are explained by the component, i.e. how much information about each item is included in the component. Items that correlate strongly with a component (high component loading), are represented well by this component. For each of the components, OS-PCA will also provide the variance accounted for (VAF), which indicates the total variance in the data explained by the component. Components are ordered according to their VAF. Hence, the first few components explain most of the variance in the data and are therefore the most important components, i.e. the *principal* components. The score of each object on each of the components (object scores) can be calculated from these results.

The items corresponding to the soft factors in the Work Profiler survey measured a person's psycho-social situation. These items were highly correlated. Therefore, OS-PCA was applied as a data preparation step to remove the correlation among these factors and to reveal their correlation structure. It was expected that for many of these covariates, the effect of the category levels was monotonically associated with the outcome. For example, it seems reasonable to expect that the more a person agrees with the statement *I am highly motivated*

to find work the coming month, the sooner this person will be reemployed. Therefore, it was preferred to keep the ordering of the category levels in the model. Furthermore, these effects were not expected to grow linearly with the category levels, so no linear restrictions were enforced. The components resulting from the OS-PCA preparation step can be used as input for a statistical model.

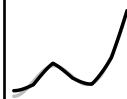
2.3.2 Survival analysis

Most of the previous research done to develop Work Profiler 1.0, and TNO's recent research to develop the second version was focused on predicting work status at 12 months. However, TNO extended their research and performed logistic regression for work status at six and nine months as well to get a better understanding of the change of reemployment probability over time. As an alternative, survival analysis can be used to analyze reemployment data if the time aspect is of interest.

Survival analysis techniques study the distribution of time to a certain event of interest (see for example Klein and Moeschberger (2003)). This could be any type of event, for example death (hence *survival* analysis), recovery, or reemployment. Typical situations in which survival analysis methods are used, are those where the time to the event of interest is not observed for some individuals. There could be many reasons why the event was not observed, for example subjects are lost to follow-up, or another event occurs which prevents the event of interest (competing risks (Putter et al., 2007)). If the event of an individual is not observed, this is a censored observation. For these observations, it is only known at which time (the *censoring time*) the event had not occurred yet. For these observations, the corresponding event time was longer than the censoring time. The censoring times provides some information about the distribution of the event times and are therefore included in the survival analysis.

There are several survival models which provide the possibility to include covariates to estimate the effect of covariates associated with event times. A popular model is the Cox proportional hazards model proposed (Cox, 1972), in which the hazard function is estimated and the effect of a covariate on the hazard is quantified by its hazard ratio (HR). The HR is the ratio of survival probabilities between subjects with different values for a particular covariate. A HR close to 1 indicates no effect from this specific covariate on the hazard. A subject with HR larger than 1 will experience the event faster than someone in the reference group, while an individual with a HR smaller than 1 will need more time than those in the reference group. Model based estimated survival probabilities can be plotted over time for different individuals.

To develop the Work Profiler, both Brouwer et al. (2015) and TNO applied



logistic regression to find possible predictors for work status at one year, i.e. a binary outcome. Additionally, TNO fitted logistic regression models at six and nine months. Survival analysis techniques can be applied to the UWV dataset to investigate reemployment probabilities during the whole first year. In this analysis, the exact unemployment duration is used as the outcome variable instead of the status at a predetermined time point. To illustrate the use of survival analysis for the duration of unemployment, a Cox proportional hazards model was fitted on the UWV data. In the model, the hard factors, the principal components that summarize the soft factors, and some interactions to find covariates associated with unemployment duration were included. In this context, a HR close to 1 indicates no effect for the corresponding risk factor on the probability of reemployment. Compared to the reference group, a HR smaller than 1 predicts a longer unemployment duration, while a HR larger than 1 indicates a shorter unemployment duration. Estimated probabilities of remaining unemployed can be plotted over time for different types of individuals.

2.4 Statistical analysis

2.4.1 Optimal scaling principal component analysis

OS-PCA was applied to the UWV dataset to investigate the correlation structure among the items of the soft factors and to reduce the observed variables to a smaller number of uncorrelated principal components. All soft factors were included in the analysis except for the items of the factors *Hours capable of work*, *Acceptance readiness - Full Time* and *Child care hindrance*, since they are strongly correlated with the hard factor *Household position*.

The choice of the number of principal components for the OS-PCA analysis was based on the combination of the VAF by each of the components and their interpretability. The items *Financial Hindrance* and *Job Search Behavior - Applications* were removed from the model, because the total VAF by these items in the final model was smaller than 0.25 (the suggested minimum by Linting and van der Kooij (2012)), which means that these items were poorly explained by the OS-PCA result. In the final OS-PCA model, 31 items were analyzed on an ordinal analysis level and summarized into seven components. In total, this model accounted for 60.9% of the variance in the original dataset. The contribution of each component to the total VAF is shown in Figure 2.1. This plot shows that the first component explained a large part of the total VAF. It accounted for around 22% of the total variance, while the other components explained 8.3% or less.

The component loadings for all items on all components are given in Table 2.4.

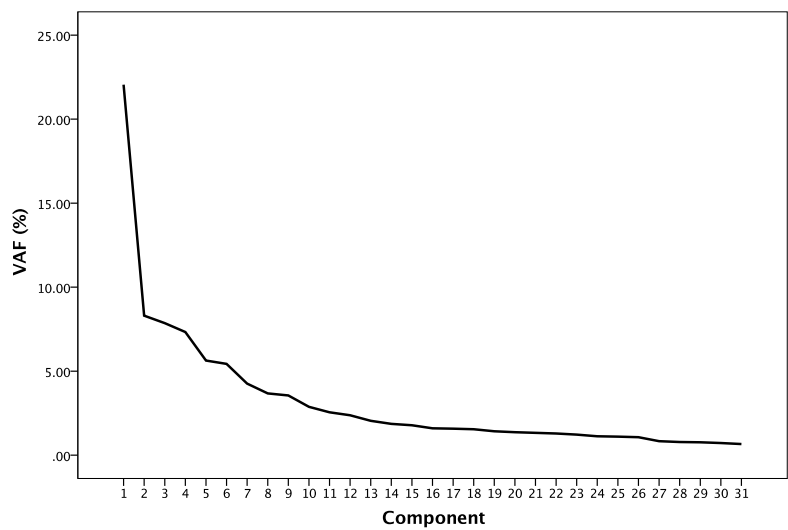


Figure 2.1: *Variance accounted for (VAF) by each component in the OS-PCA model of 7 dimensions on the 31 items corresponding to soft factors.*

The factors whose items were highly correlated with the components are the ones with a high absolute component loading. Linting and van der Kooij (2012) used 16% VAF (absolute value component loading > 0.4, printed in bold gray in Table 2.4) as a cut-off value. However, 20% VAF (absolute value component loading > 0.447, printed in bold black in Table 2.4) may be a more reasonable choice. The first principal component is associated to persons who feel healthy, have high expectations, are highly motivated, and know how to find a job. These factors appeared to be highly correlated, and most of the variance in the dataset was explained by differences in these factors. There is a striking resemblance between the factors that correlated strongly with the first components, and the Theory of Planned Behavior (Ajzen, 1991) and the Motivation Model (Vroom, 1964). According to the Theory of Planned Behavior, high scores on factors like subjective norm, job search attitude, and job search self-efficacy indicate a strong job search intention, which will usually result in job search behavior. As shown in Table 2.4, items on these factors were strongly correlated with the first component, and with each other. This result supports the Theory of Planned Behavior. Furthermore, the high component loadings on the items on *view return to work* showed that unemployed who have high expectations are more motivated to find a job, also showed more job search behavior. This is in accordance with the Motivation Model.

Table 2.4 suggests that the second component is related to health perception; individuals who score high on the five perceived health items, will also get a

high score on the second component.

Short interpretations of the other components are as follows. Component 3 and 6 contain mainly variables about a person's acceptance readiness for different type of work or different working hours compared to a previous job. Component 4 contrasts persons who score high on external variable attribution and those who are not ready to accept different working hours. Component 5 is about job search behavior, while component 7 indicates a person's job search attitude. The objects scores for each person for each principal component was calculated using this OS-PCA model.

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
Acceptance - time 1	0.330	-0.166	0.434	-0.414	0.166	0.339	0.151
Acceptance - time 2	0.322	-0.170	0.410	-0.429	0.209	0.315	0.147
Acceptance - work 1	0.122	-0.261	0.436	-0.253	-0.030	0.434	0.121
Acceptance - work 2	0.228	-0.183	0.303	-0.184	-0.026	0.529	0.112
External variable 1	-0.254	-0.319	0.435	0.412	0.195	-0.244	0.101
External variable 2	-0.039	-0.310	0.336	0.473	0.258	-0.183	0.108
External variable 3	-0.103	-0.225	0.333	0.432	0.242	-0.123	0.083
Financial need	0.363	-0.360	0.224	-0.232	-0.027	-0.264	-0.276
Search attitude - a/p 1	0.286	-0.266	-0.127	-0.154	-0.333	-0.309	0.635
Search attitude - a/p 2	0.310	-0.298	-0.111	-0.161	-0.397	-0.312	0.570
Search attitude - u/n 1	0.569	-0.353	0.131	-0.080	-0.284	-0.231	-0.246
Search attitude - u/n 2	0.598	-0.315	0.060	-0.093	-0.316	-0.234	-0.168
Search beh - contact 1	0.387	-0.040	-0.241	-0.205	0.553	-0.174	0.045
Search beh - contact 2	0.322	-0.134	-0.161	-0.189	0.580	-0.243	0.062
Search beh - contact 3	0.280	-0.103	-0.153	-0.246	0.620	-0.254	0.107
Search intention 1	0.640	-0.260	0.011	0.265	-0.032	0.071	-0.181
Search intention 2	0.592	-0.177	-0.279	0.279	0.016	0.212	-0.047
Search intention 3	0.689	-0.241	-0.046	0.267	-0.036	0.089	-0.155
Perceived health 1	0.615	0.423	0.253	0.134	-0.022	-0.038	0.082
Perceived health 2	0.652	0.490	0.277	0.125	0.009	-0.082	0.072
Perceived health 3	0.639	0.521	0.287	0.146	0.004	-0.087	0.097
Perceived health 4	0.578	0.458	0.269	0.184	-0.034	-0.055	0.048
Perceived health 5	0.663	0.482	0.304	0.118	0.004	-0.112	0.075
Self-effic - prep appl 1	0.484	-0.221	-0.307	0.365	0.005	0.288	0.107
Self-effic - prep appl 2	0.484	-0.214	-0.329	0.381	0.087	0.259	0.123
Self-effic - prep appl 3	0.523	-0.172	-0.335	0.377	0.050	0.269	0.084
Subjective norm	0.448	-0.287	0.189	-0.119	-0.076	-0.209	-0.276
Balance pros cons	-0.397	0.157	-0.164	0.156	0.046	0.091	0.320
View return to work 1	0.524	0.212	-0.346	-0.325	0.022	0.017	-0.042
View return to work 2	0.521	0.220	-0.327	-0.223	-0.071	0.063	-0.089
View return to work 3	0.537	0.091	-0.326	-0.213	0.079	-0.020	-0.013

Table 2.4: Component loadings for the final OS-PCA model. All loadings with absolute value more than $\sqrt{0.16} = 0.4$, i.e. $> 16\%$ VAF, are printed bold and in gray. The ones accounting for more than 20% variance are printed bold and in black. See Table 2.2 for the variables' full names.

2.4.2 Survival analysis

A Cox proportional hazards model was initially fitted on all hard factors (see Table 2.1) and the seven principal components found by the OS-PCA as described in subsection 2.4.1. Since one of the aims of UWV's research is to remove items that do not predict unemployment duration, a parsimonious model was fitted by removing some of the variables from the model containing all factors. The hard factor on education level was removed because this is closely related to someone's profession level, and it probably loses its importance when work experience increases. Items on former working hours and on the extend of unemployment were also removed since they were closely related to the number of days per week a person is able to work. Variables that had no significant effect on unemployment duration, and significant variables whose hazard ratio was close to 1, i.e. which had only a very small effect on unemployment duration, were also removed. Interactions between gender and household position, and between gender and the number of days per week able to work were included in the model.

Variables associated with unemployment duration, according to the final model, along with their corresponding HRs are given in Figure 2.2. Age is one of the most important indicators for unemployment duration. The older an individual is, the more time it will take to find a new job. Figure 2.3 shows survival curves which describe the model based probability of remaining unemployed as a function of time for different individuals. The covariate and principal component values used to make the curves shown in Figure 2.3 are displayed in Table 2.5. Note that all survival curves are equal to 1 from the beginning of the unemployment status, until around 70 days. This is due to how the data are collected. Only individuals who received unemployment benefits for at least ten weeks were included in the analysis (see subsubsection 2.2.1). Therefore, for this group, nobody was reemployed within 70 days, leading of a 100% survival (all people are unemployed) for this time period.

In Figure 2.3 survival curve estimates of several types of unemployed persons are shown. In Table 2.5 all values of the covariates used to plot the survival curves are provided.

First we compare the probabilities of remaining unemployed for young Dutch men and women who have little work experience and who live alone were studied. In Figure 2.3a the probability of remaining unemployed for this type of person with a high job search intention (value 1 for the first principal component) is shown. Figure 2.3b shows a person with the same characteristics, but with a low job search intention (value -1 on the first principal component). These plots show that there is almost no gender difference. However, the effect of the

first principal component is large; the probability of finding a job within one year is around 80% for individuals with a high intention to find a job, while this probability is much lower, around 55%, for those with a low score on the first principal component. This means that this component is strongly associated with unemployment duration, which was already indicated by the HR corresponding to this component.

The effect of household position on Dutch men and women in their 40s who are highly motivated to find a job, but can only work for four days per week (see Table 2.5) was studied. Figure 2.3c and Figure 2.3d show the probabilities of remaining unemployed for these men and women respectively. These plots show, for example, that finding a job takes more time for married or cohabiting mothers with young children. This phenomena is quantified by the HR of the interaction of gender and household position ($HR < 1$).

Figure 2.3e and Figure 2.3f show the probability of reemployment for men and women in their 40s at different profession levels, who are married or cohabiting and have a young child of at most six years, with an average level of intention to find a new job (see Table 2.5 for exact covariate values). The plots suggest that the probability of finding a new job is approximately similar for mothers at all profession levels, while for fathers, profession seems to be associated to the reemployment probability. The plots show that overall the probability of begin reemployed is smaller for mothers than for fathers.



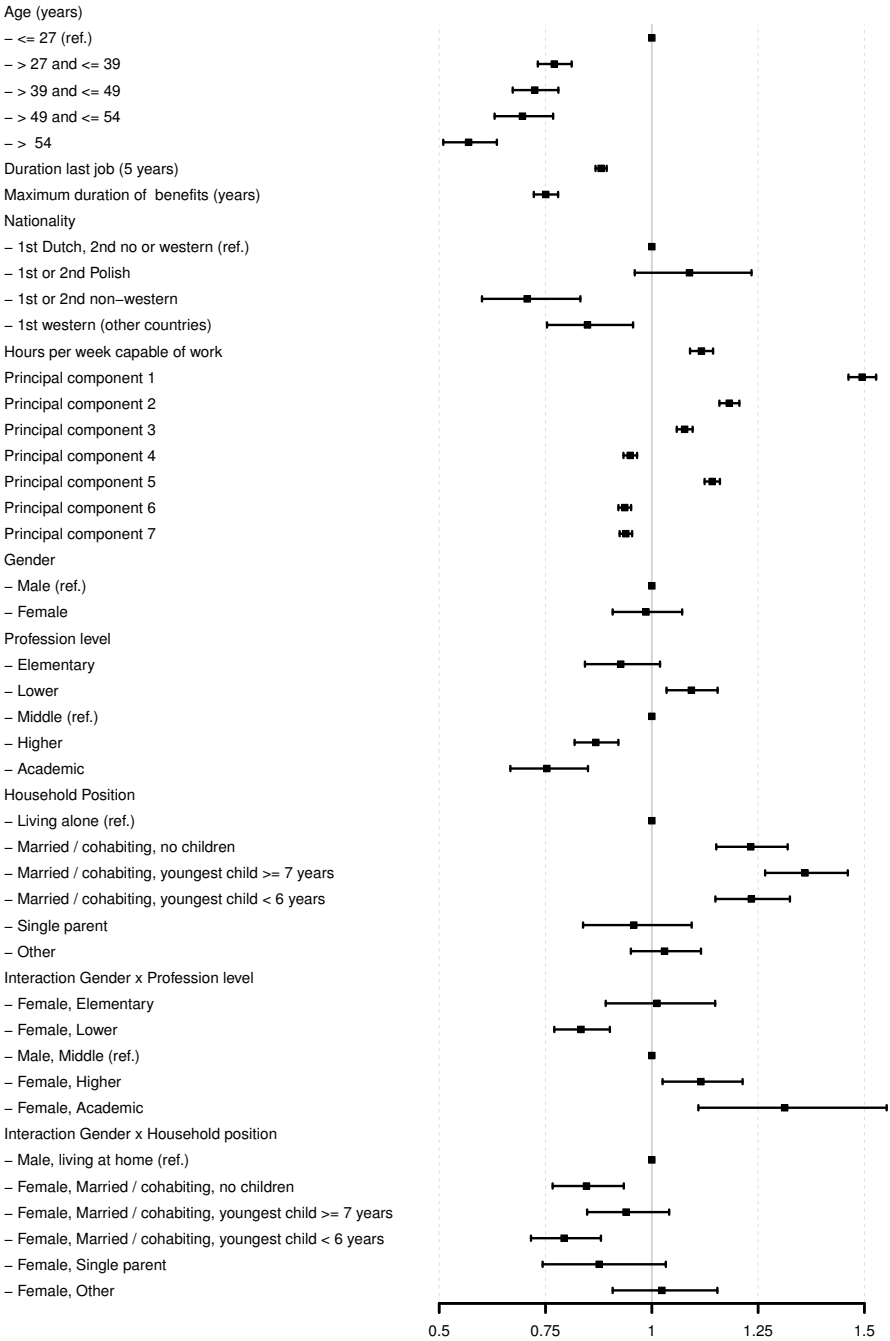


Figure 2.2: *Estimated hazard ratios and their 95% confidence intervals for the covariates in the final Cox model fitted on the hard factors and principal components. The reference category is indicated for each categorical variable.*

	Age (years)	Duration last job (years)	Gender	Household position	Maximum duration of benefits (years)	Nationality	Profession level	Days per week capable of work	PC 1	PC 2-7
Figure 2.3a	≤ 27	2.5	<i>All categories are shown</i>	Living alone	0.75	1st: Dutch 2nd: no/western	Middle	5	1	0
Figure 2.3b	≤ 27	2.5	<i>All categories are shown</i>	Living alone	0.75	1st: Dutch 2nd: no/western	Middle	5	-1	0
Figure 2.3c	40 – 49	10	Male	<i>All categories are shown</i>	2	1st: Dutch 2nd: no/western	Middle	4	1	0
Figure 2.3d	40 – 49	10	Female	<i>All categories are shown</i>	2	1st: Dutch 2nd: no/western	Middle	4	0	0
Figure 2.3e	40 – 49	10	Male	Married/ cohabiting, youngest child < 6 years	2	1st: Dutch 2nd: no/western	<i>All categories are shown</i>	4	0	0
Figure 2.3f	40 – 49	10	Female	Married/ cohabiting, youngest child < 6 years	2	1st: Dutch 2nd: no/western	<i>All categories are shown</i>	4	0	0

Table 2.5: Covariate values used to estimate the survival curves shown in Figure 2.3. The only difference between the individuals in Figure 2.3a and Figure 2.3b is the value of Principal Component 1 (printed in bold). The differences between the individuals in Figure 2.3c and Figure 2.3d, and the individuals in Figure 2.3e and Figure 2.3f is gender (printed in bold).

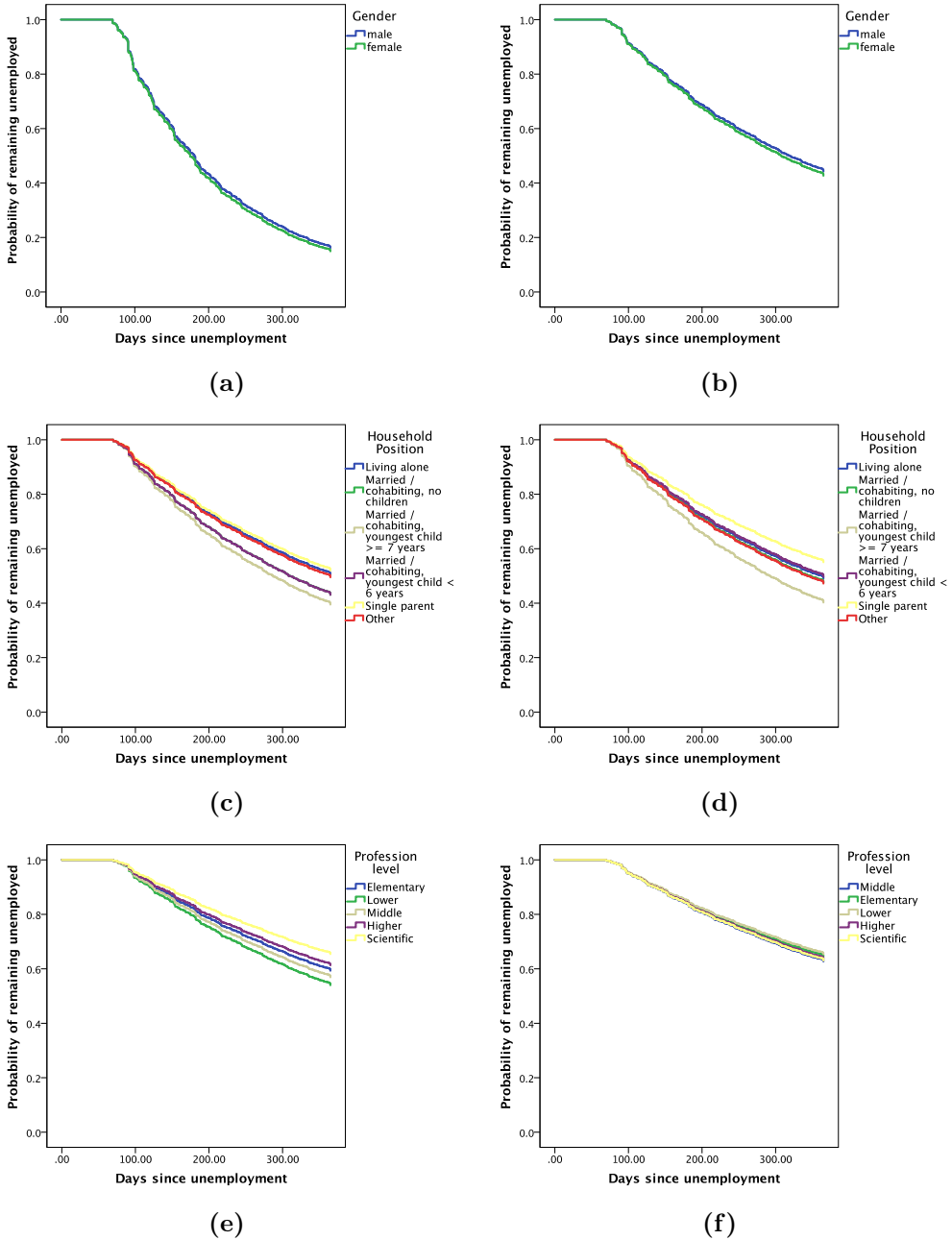


Figure 2.3: Estimated probabilities of remaining unemployed for the six individuals with covariate values as given in Table 2.5. The upper panel shows the effect of the first principal component for young persons with little work experience and who live alone. The middle and lower panel respectively show the survival curves corresponding to different household positions and profession levels for both genders.

2.4.3 Conclusions

The two-step analysis performed on the UWV data provided an interesting insight of the characteristics of the unemployed. The OS-PCA analysis revealed the correlation structure of the survey items from the extended version of the Work Profiler. The survival analysis provided possible predictors for unemployment duration.

The OS-PCA resulted in a model with seven components. The first component seemed to have the strongest association with unemployment duration. This component shows strong similarities with two theories on job search behavior: the Theory of Planned Behavior and the Motivation Model. Soft factors related to unemployment duration provided by the methodology discussed in this paper are the same factors suggested by the two theories. Additionally, health perception seemed to be important indicator for job search behavior.

Survival analysis results indicate age as one of the best indicators for unemployment duration. The model shows that older unemployed have the smallest probabilities of finding a new job. Furthermore, gender seems to have no significant effect on unemployment duration. However, some household positions are more disadvantageous for women when it comes to finding a new job, and the association of profession level with unemployment duration is different among the two genders.

This analysis was based on a dataset that only contained unemployed who claimed unemployment benefits and remained unemployed for at least ten weeks. Hence, it did not take into account that some characteristics may cause a very short (< 10 weeks) unemployment duration.

The analysis proposed in this paper aimed to illustrate how OS-PCA can be used to assess the correlation structure between survey items and how survival methodology can be applied to investigate covariates associated with reemployment duration. This is a preliminary analysis and is not meant for policy making. The analysis could be extended, for example by considering different settings in the OS-PCA, or by including some of the hard factors in the data preparation step as well. Depending on the research question, the analysis could be extended to also compare, for example, the probability of reemployment for different sectors or different regions.

2.5 Discussion

In this paper, two alternative methods for a two-step analysis of reemployment data were discussed. First, OS-PCA was proposed as an alternative to factor analysis to summarize item scores in survey data. Next, survival analysis was proposed as an alternative to logistic regression to analyze the probability of the



occurrence of an event when the time duration is of interest. In this section the advantages and disadvantages of these methods are compared

Usually many survey items are very closely related, which results in highly correlated data. Correlated data may cause problems when fitting regression models. Therefore, the item scores are often summarized in several composite scores which are not or only a little correlated. Using these composite scores as input in the model reduces the number of variables in the model. This usually simplifies the model interpretation.

Several methods have been developed to summarize items into summary scores. In this paper, two methods were illustrated: factor analysis and OS-PCA. Factor analysis summarizes related items into composite scores, resulting in factor scores. Each item corresponds to only one factor and the factors may be correlated. OS-PCA estimates several principal components. These components are a weighted average of all items and are uncorrelated. Component loadings indicate the importance of each item in each component.

Since factor scores are a weighted average of a subset of the items, they are usually easier to interpret than principal components in which all items are included in the composite score. As a result, the estimated factor model is easy to interpret. Furthermore, if researchers aim to reduce the number of items in a survey, factor analysis will allow them to remove factors and their corresponding items one by one. Since all items only belong to one factor, removing all items corresponding to one factor will not influence the scores of the other factors. On the other hand, in the OS-PCA setting, removing one item will influence all component scores, since these are weighted averages of all items. Hence, reducing the length of a survey can more easily be done using factor analysis. However, a disadvantage of factor analysis is that factor scores may still be correlated with each other. Therefore, the collinearity problem may still occur when factor scores are included in a statistical model. Principal components analysis overcomes this problem since the estimated components are uncorrelated.

A strong advantage of OS-PCA is its ability to transform categorical item scores nonlinearly, as opposed to the numerical interpretation in factor analysis. The nonlinearity transformation allows for the evaluation of categorical data without losing data properties. In case unequal distances are expected between the categorical levels, one may consider using the OS-PCA method for dimension reduction.

Based on the differences between the two techniques, the choice between them mainly depends on the aims of a study. If the model should be easy to interpret and allow for removal of items, factor analysis is probably the best analysis method. If interpretation is less important, but keeping the properties of the ordinal categorical data is preferable, OS-PCA might be a solution for

dimension reduction.

Once survey data has been summarized, the summary scores can be used to predict the outcome of interest. In the context of reemployment data this could be the probability of being reemployed after a prespecified time point, or the actual duration of unemployment. The choice between logistic regression and survival analysis depends mostly on the outcome of interest. If interest lies in the reemployment probability at a specific time point, logistic regression is the appropriate model. However, if a binary outcome at a specific time point seems too strict, survival analysis is the proper methodology to be applied, since this method estimates the distribution of the unemployment duration. Although the research question should be leading in choosing the analysis method, other model properties might play a role as well.

Prediction error estimation may be one of these properties. If researchers want an easily interpretable prediction error, logistic regression may be the better choice. Since prediction is performed at a specific time point, one can check whether the prediction was correct at that time point. For example, if the probability of being reemployed within the set time was estimated to be $> 50\%$ for a person with particular characteristics, and this person had actually found a job before that time, one could say the prediction was correct. In this way, the ratio of correct predictions among all predictions gives an indication of the prediction error. A ratio of 0.5 indicates a bad performance, comparable to flipping a coin. A score of 1 indicates perfect prediction. Instead of 50%, other cut-off values could be chosen as well. The optimal cut-off value can be determined by minimizing the prediction error over all possible cut-off values.

In survival analysis, estimation of the prediction error is slightly more complicated. Since the actual time duration is included in the model, the prediction error is evaluated at a grid of time points. The Brier score (Graf et al., 1999), for example, computes at each time point the mean squared difference between the estimated survival probabilities and the actual outcomes (still alive/unemployed = 0, dead/reemployed = 1). Hence, a low Brier score is preferable. Although the Brier Score will give a good indication of the prediction error over the time grid, it is not possible to interpret it as the ratio of correctly predicted outcomes. The C-index (Harrell et al., 1996) might be used to quantify the ratio of concordant pairs of observations, i.e. the ratio of pairs of observations whose events were predicted to be in the same order as they actually occurred. This index is useful, but harder to interpret compared to the ratio of correct predictions used in logistic regression.

Another aspect that may play a role in the model choice is the possible presence of missing data. Since observing an event takes time, the events of some subjects may not be observed. In the context of reemployment data for



example, some reasons for not observing the time for reemployment might be emigration to another country, retirement, or death. In a logistic regression analysis, the outcome would be missing for these subjects, which makes the analysis more complicated. Survival analysis methods, however, were designed to include this type of subjects in the analysis. When time to event is not observed, the time until last contact would be included in the model. Depending on the type of incomplete information, the observation will be used as censored (as emigration) or as competing event (as retirement and death). It will include all the known information in the model as the censoring time, i.e. the date of emigration, retirement, or death, and will estimate the model while using this information.

In this paper, the advantages and disadvantages of the proposed methods to analyze reemployment data were discussed and the research questions that can be answered by each method were characterized. Aspects like interpretability, collinearity of variables, prediction error estimation, and missing data must be considered when presenting the method. Although this paper's focus is on the combinations of factor analysis with logistic regression and OS-PCA with survival analysis, researchers are not limited to only these two combinations.

Acknowledgements

The Dutch Employee Insurance Agency (UWV) is gratefully acknowledged for making the data available for this chapter. In particular, Harriët Havinga and Gijsbert van Lomwel (Kenniscentrum UWV) are acknowledged for their help on the interpretation of the covariates and model results.

OPTIMAL SCALING FOR SURVIVAL ANALYSIS WITH ORDINAL DATA

3

Medical and psychological studies often involve the collection and analysis of categorical data with nominal or ordinal category levels. Nominal categories have no ordering property, like gender, while ordinal category levels do have an ordering, for example when subjects are classified according to their education level, often categorized as low, medium, or high education. Currently two methods can be chosen to include ordinal covariates in the Cox proportional hazards model in survival analysis. Dummy covariates can be used to indicate category memberships, as is done for nominal covariates. Then the estimated parameters for each category indicate the risk of experiencing the event of interest relative to the reference category. Since these parameters are estimated independently from each other, the ordering property of the categories is lost in the process. To keep the ordinal property, integer values can be given to the category levels (e.g. low = 0, medium = 1, high = 2), and the variable can be included in the model as a numeric covariate. However, the ordinal data are now interpreted as numeric data, so the property of equal distances between consecutive categories is introduced. This assumption may be too strict for ordinal data. In this paper a method is described to include ordinal data in the Cox model. The method implements optimal scaling to find quantifications for the ordinal category levels. These quantifications are chosen such that they preserve the categories' ordering and do not force equal distances between consecutive category levels. A simulation study is carried out to compare the performance of optimal scaling, and dummy and integer coding. Results

This chapter is published as Willems, S. J. W., Fiocco, M., and Meulman, J. J. (2017) *Optimal scaling for survival analysis with ordinal data*. Computational Statistics & Data Analysis, 115, 155–171.

show that the optimal scaling method increases the model fit if ordinal covariates are included in the model.

3.1 Introduction

In medical and psychological studies a lot of data about patients are collected, for example their gender, age, education level, weight, and socio-economic status. These characteristics can have different measurement levels, namely numeric or categorical. Numeric variables are those variable that are measured on a continuous scale, like age and blood pressure. Categorical variables are not measured on a continuous scale, but instead subjects are assigned to one of the pre-defined category levels. There are two types of categorical data, nominal and ordinal. Category levels of nominal variables are unordered, while the categories of ordinal variables are ordered. Nominal variables seen in medical studies are, for example, gender, treatment group, and ethnicity. Gender has the two unordered categories, male and female. Treatment groups may be defined as treatment A, B and C, or treatment vs. placebo, and these are usually unordered. Ethnicity can have several category levels, depending on the ethnicities of interest, but there is no ordering involved. Examples of ordinal categorical variables are education level, and scales like pain severity scales, Likert scales, or the modified Rankin Scale (mRS). Schools and diplomas may be categorized into *low*, *medium* and *high* education levels, which clearly have an ordering. Pain severity scales are used to get an indication of the intensity of a patient's pain. Likert scales are used to measure how strongly people agree or disagree with a statement, e.g. with response options *strongly disagree*, *disagree*, *I don't know*, *agree*, and *strongly agree*. The mRS is used to measure the degree of disability or dependence in daily activities of patients who suffer from neurological disabilities, e.g. caused by a stroke (van Swieten et al., 1988). A property of the ordered category levels in ordinal data is that the distances between consecutive category levels do not necessarily represent an equal degree of difference. For example, the mRS score ranges from 0 to 5 where 0 indicates no symptoms and 5 severe disability. There is a slight difference between scores 0 and 1; from no symptoms (0) to no significant disability (1). However, the difference between scores 2 and 3 is large, since it indicates the transition from being functionally independent (2) to being functionally dependent (3).

Researchers may choose between analyzing a specific variable according to its measurement level, or to adjust the scale for analysis. For example, the measurement level of age may be numeric (exact ages of patients are known), but researchers may decide to discretized the covariate and include the resulting age groups in the statistical models instead of the exact ages. Due to this

discretization the analysis level is ordinal, while the measurement level was numerical.

In many statistical models a linear combination of predictor variables is used to predict an outcome or response variable. Examples of these types of models are the standard linear model, where the outcome is predicted directly from the linear combination of predictors; generalized linear models, in which the outcome is predicted from the linear model through a link function; and the Cox model in survival analysis, where the linear predictor is included in the hazard function. Models with linear predictors are directly applicable for variables that are analysed on either a numeric or nominal level. Numeric variables are included in the model, where the coefficients indicate the increase or decrease in risk for every unit increase. For nominal data, $C_k - 1$ dummy variables are introduced, where C_k represents the number of categories for variable k . The corresponding $C_k - 1$ estimated model parameters indicate the difference in risk between a category level relative to the reference level.

Complications arise for ordinal categorical data. In most literature on models with linear predictors, no methods on how to fit these models for ordinal data are discussed. Researchers usually use either the nominal or numeric approach. In the nominal approach, dummy variables are introduced and the model is fitted in the same way as for nominal data. However, this method ignores the ordering property of the ordinal category levels, since it assumes unordered (nominal) category levels. Therefore, it is not guaranteed that the linear predictor increases (or decreases) with each increase of category level. To keep the monotonicity, one can analyze the ordinal data using a numeric approach. In this case, each category is given an integer value (e.g. 0, 1, 2, etc.), and the variable is then included in the model as a numeric variable. By using the integer coding, equal distances between consecutive categories is assumed, although the distances are not necessarily equal in the data. Hence, unfortunately, neither of these two approaches respect the ordinal categorical data characteristics and are therefore not suitable for analyzing this data type.

To analyze ordinal data, optimal scaling techniques have been developed (Gifi, 1990). In regression analysis, this method provides an optimal nonlinear transformation of the category levels such that the relation between the response and the predictors is optimal. In this way, the optimal scaling method turns qualitative data (ordered category levels) into quantitative data (numeric values). The resulting optimal quantifications can be treated as numeric data in the model. The nonlinear optimal quantifications are found by fitting a nonlinear monotone transformation on the original category values. The monotonicity restriction of the transformation guarantees that the ordering of the category levels is maintained and the nonlinearity enables unequal distances between consecutive category levels.



The optimal scaling method was first developed for simple linear models, but was extended to more complicated models that include a linear combination of predictors. Actually, optimal scaling can easily be included in any model that is fitted with a least squares algorithm, as the regression (Meulman et al., 2019) and the principal components model (Linting et al., 2007; Meulman et al., 2004). Including the optimal scaling step results in an alternating least squares algorithm in which the loss function is iteratively minimized over the model parameters and the optimal scaling quantifications.

The inclusion of optimal scaling is more complicated for models that are fitted with a maximum likelihood approach. This complexity may be the reason why optimal scaling is not yet used to analyze variables on an ordinal level in the Cox proportional hazards model in survival analysis, a model that is fitted by the maximum likelihood method. Currently, researchers include ordinal variables in the model by analyzing them on a nominal or numeric level, and in this way lose the ordering property or introduce equal distances between consecutive categories.

Our research focuses on optimal scaling in survival analysis, and in this paper we show how the optimal scaling method can be incorporated in the Cox model. In section 3.2 we will first describe how ordinal data are currently included in a Cox model, and how optimal scaling is currently used for simple linear regression. In section 3.3, a least squares approach to find the maximum likelihood estimator for the Cox model is described, and optimal scaling is incorporated in this algorithm. In section 3.4, the performances of different approaches to fit the Cox model for ordinal data (nominal, numeric and optimal scaling) are compared in a simulation study. The simulation results show that the optimal scaling approach gives the most accurate model fit.

3.2 Current practice

In this section we will first describe in more detail the methods currently used to incorporate ordinal data in the Cox proportional hazard model. Then, we will discuss the basic principles of the optimal scaling method by showing an application to the simple linear model.

3.2.1 Ordinal data in survival analysis

The aim of survival analysis is to estimate the time to an event of interest, measured from a specific origin. For example, survival models can be used in a medical setting to determine whether a certain treatment prolongs the life time of patients since start of treatment. Since survival times may differ between patients with different characteristics, patient information is collected and incorporated

in survival models. Survival data for individual i are represented as a triplet $(t_i, \delta_i, \mathbf{z}_i)$, (for $i = 1, \dots, n$). If subject i 's event is observed ($\delta_i = 1$), t_i represents the survival time x_i . If the event is not observed ($\delta_i = 0$), t_i represents the censoring time c_i , i.e. $t_i = \min(x_i, c_i)$. Observed covariate values are denoted by $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})$, with p the number of measured covariates.

The relation of covariates with event times is often modeled with the Cox proportional hazards model. In this model, the hazard rate at time t is as follows

$$\begin{aligned} h(t|\mathbf{Z}) &= h_0(t) \exp(\mathbf{Z}\boldsymbol{\beta}) \\ &= h_0(t) \exp \left[\sum_{k=1}^p \beta_k Z_k \right], \end{aligned} \quad (3.1)$$

where $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_p)$ are the model parameters and $Z_1, \dots, Z_p$ the covariates. The survival probabilities of individuals with covariate values $\mathbf{z}_i$ and $\mathbf{z}_j$ can be compared by looking at the proportion of their hazards, i.e.

$$\begin{aligned} \frac{h(t|\mathbf{z}_i)}{h(t|\mathbf{z}_j)} &= \frac{h_0(t) \exp \left[\sum_{k=1}^p \beta_k z_{ik} \right]}{h_0(t) \exp \left[\sum_{k=1}^p \beta_k z_{jk} \right]} \\ &= \exp \left[\sum_{k=1}^p \beta_k (z_{ik} - z_{jk}) \right], \end{aligned} \quad (3.2)$$

which is a constant. Ratio (3.2) is called the *hazard ratio* and represents the *relative risk* of an individual with risk factor $\mathbf{z}_i$ experiencing the event as compared to an individual with risk factor $\mathbf{z}_j$. Regression coefficient β_k , for $k = 1, \dots, p$, in the model indicates the change in the relative risk for different values of covariate Z_k .

The way in which a covariate is incorporated in the model depends on its analysis level. Covariates with a numeric analysis level can be included directly. In this case, the regression coefficient β_k indicates the change in the relative risk when the covariate value is increased by one unit. For nominal covariates a dummy coding will be introduced, and fitted regression coefficients will indicate the relative risk between category levels. For details see the book by Klein and Moeschberger (2003).

If a covariate Z_k has C_k category levels, $C_k - 1$ dummies are required. For example, to code categories *low*, *medium*, and *high* two dummy covariates D_1 and D_2 can be defined as

$$\begin{aligned} D_{1_i} &= 1 \text{ if subject } i \text{ is in category } \textit{medium}, & 0 \text{ otherwise,} \\ D_{2_i} &= 1 \text{ if subject } i \text{ is in category } \textit{high}, & 0 \text{ otherwise.} \end{aligned}$$

The resulting dummy coding for each category is presented in Table 3.1.

	Dummy Coding		Integer Coding	Optimal Scaling
	D_1	D_2	Z_{int}	Z_{os}
Low	0	0	0	0
Medium	1	0	1	1.2
High	0	1	2	1.8

Table 3.1: Three categories' codings used in for dummy coding, integer coding and the optimal scaling method.

	Dummy Coding	Integer Coding	Optimal Scaling
$h(t Low)$	$h_{0D}(t)$	$h_{0int}(t)$	$h_{0os}(t)$
$h(t Medium)$	$h_{0D}(t) \exp(\beta_{D_1})$	$h_{0int}(t) \exp(1 \beta_{int})$	$h_{0os}(t) \exp(1.2 \beta_{os})$
$h(t High)$	$h_{0D}(t) \exp(\beta_{D_2})$	$h_{0int} \exp(2 \beta_{int})$	$h_{0os} \exp(1.8 \beta_{os})$

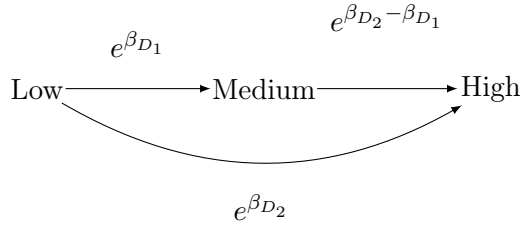
Table 3.2: Three categories' hazard functions in dummy coding, integer coding and the optimal scaling method. Indices " D_1 " and " D_2 " indicate dummies 1 and 2, index " int " indicates integer coding, and index " os " indicates optimal scaling.

For each dummy a model parameter is estimated. This results in the hazard rates shown in Table 3.2.

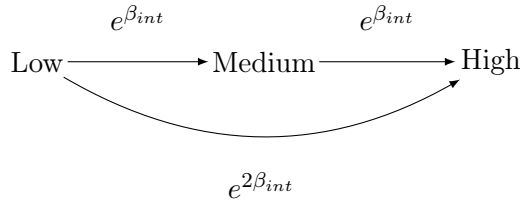
The relative risks between each category are shown in Figure 3.1a. If $\beta_{D_1}, \beta_{D_2} > 0$ or $\beta_{D_1}, \beta_{D_2} < 0$, and $|\beta_{D_2}| > |\beta_{D_1}|$ the relative risk between category levels *low* and *high* will be larger than the relative risk between levels *low* and *medium*, and the ordering of the category levels will be maintained. In all other cases, the relative risks do not correspond with the ordering of the category levels.

If there are a priori reasons to expect the relative risks to have the same order as the category levels, an integer coding system can be used instead of dummy coding to preserve the ordering of the category levels. In this coding system, integer values are given to each category level such that the ordering of the integers corresponds to the ordering of the categories. For example, categories *low*, *medium*, and *high* could be coded as 0, 1, and 2 respectively, see Table 3.1. The covariate is now included in the model as a numerical covariate, and only a single parameter β_{int} will be estimated. For the integer coding the hazard rates for subjects in the three categories will be as in Table 3.2. The relative risks between the categories are shown in Figure 3.1b. Since $|\beta_{int}| < |2\beta_{int}|$ holds for any β_{int} , the correct ordering of the category levels is always ensured by using this integer coding system. Due to the choice of codings, the relative risks

between *low* and *medium* and *medium* and *high* are equal, namely $\exp(\beta_{int})$. Therefore, this coding system forced equal relative risks between the consecutive categories. Since the distances between category levels of ordinal data are not necessarily equal, assuming that the relative risks is equal is inappropriate for this type of data.



(a) *Dummy coding method.*



(b) *Integer coding method.*

Figure 3.1: Relative risks between subjects in categories low, medium, and High for the dummy and integer coding methods.

The two currently used coding systems, dummy and integer coding, are not appropriate for ordinal data. Dummy coding does not ensure the preservation of the ordering of category levels, and integer coding will keep the ordering of the category levels, but will force equal relative risks between consecutive category levels. In this paper we describe a method to find numerical values (quantifications) for each category level which will preserve the ordering of the categories, but will not force equal distances. An example concerning quantifications for the category levels *low*, *medium*, and *high* are shown in Table 3.1 and corresponding hazard functions are given in Table 3.2. Figure 3.2 shows that the quantifications are in the same order as the category levels, but the distances between category levels are not the same (nonlinear), i.e. the distance between levels *low* and *medium* is larger than the distance between *medium* and *high*.

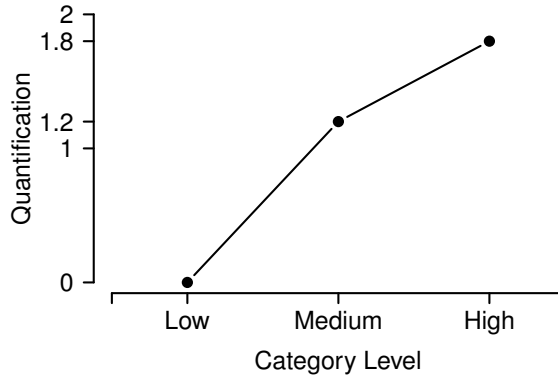


Figure 3.2: *Example of quantifications for a categorical variable with three levels.*

Quantifications should be chosen in an optimal manner, such that the relation between the covariates and the outcome of interest is maximized. Optimization can be obtained by maximizing the likelihood function for the Cox model. The method used to find quantifications is called *optimal scaling*. For simple regression models, quantifications are estimated by maximizing the relation between the outcome and covariates by minimizing the sum of squared residuals. The optimal scaling procedure for regression will be discussed in subsection 3.2.2.

3.2.2 Optimal scaling in simple linear regression

The aim of optimal scaling is to find a transformation that assigns numerical values (quantifications) to each category level of a covariate in such a way that the relation between subjects' covariates and the model outcome is maximized, while respecting the measurement characteristics of the data, e.g. ordering of category levels. Maximizing this relation can be done, for example, by minimizing the loss function or maximizing the likelihood. Restrictions are placed on the transformation to preserve the characteristics of the data. Nominal quantifications preserve only class membership information, i.e. if individuals i and j are in the same category, they should be assigned the same numerical value. For ordinal data, the order of the category levels should be preserved as well. If for example individual i is in a lower level than individual j , then the quantification for individual i should be smaller (or equal) to the quantification for individual j . In the latter case, the category levels and quantifications are related by a monotonic function. This monotonic function can take different forms, for example a step function or a spline function. The monotonic regression approach proposed by Kruskal (1964) is used if the number of category levels is

low. Spline transformations are often used to keep the fine grid when there are many category levels. The method developed by Ramsay (1988) and implemented by Meulman et al. (2019) can be used to fit spline transformations. This paper concentrates on ordinal categorical variables with only a few category levels, so focus is on nonmonotone step functions.

Many statistical models aim to predict an outcome from a set of predictor values. For linear regression, the outcome is usually denoted as Y and the p predictor values as $X_1, \dots, X_p$. To avoid confusion with the notation used in the survival model, we will denote the set of predictor values by $Z_1, \dots, Z_p$ for the linear model. The outcome will be denoted as Y .

The model is fitted on observed data from n subjects. Let $\mathbf{y}$ be the vector of length n that consists of all n observed outcomes y_i , with $i = 1, \dots, n$. Denote by $\mathbf{Z}$ the matrix with dimensions $n \times p$ that contains the observed covariate values for all subjects, i.e. if z_{ik} is the observed value of covariate k for subject i , then

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad \mathbf{Z}_{n \times p} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1k} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2k} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots \\ z_{i1} & z_{i2} & \cdots & z_{ik} & \cdots & z_{ip} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{nk} & \cdots & z_{np} \end{pmatrix}.$$

Let i be the row index of matrix $\mathbf{Z}$ with $i = 1, \dots, n$. Denote by $\mathbf{Z}_{i*}$ the vector of length p that contains the p observed covariates corresponding to subject i , i.e. $\mathbf{Z}_{i*} = (z_{i1}, \dots, z_{ip})$. Let k be the column index of $\mathbf{Z}$, with $k = 1, \dots, p$; then the vector $\mathbf{Z}_{*k}$ contains the n observed values for the specific covariate k , i.e. $\mathbf{Z}_{*k} = (z_{1k}, \dots, z_{nk})^T$. In the linear model, the response y_i corresponding to subject i is modeled as

$$y_i = \mathbf{Z}_{i*}\boldsymbol{\beta} + \epsilon_i, \quad (3.3)$$

with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ the vector of regression coefficients and ϵ_i the error term. The parameters $\boldsymbol{\beta}$ are estimated by minimizing the loss function

$$L(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{Z}_{i*}\boldsymbol{\beta})^2 = \|\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}\|^2 = \left\| \mathbf{y} - \sum_{k=1}^p \mathbf{Z}_{*k}\beta_k \right\|^2. \quad (3.4)$$

The loss function is minimized by the ordinary least squares solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y}. \quad (3.5)$$

In case of categorical data, the linear model can be extended to include transformations φ_k for each variable Z_k (with $k = 1, \dots, p$). The transformed values give numerical representations, quantifications, of the category values. The vector $\varphi_k(\mathbf{Z}_{*k}) = (\varphi_k(z_{1k}), \dots, \varphi_k(z_{nk}))^T$ contains the quantifications for all observed categories of covariate Z_k . For example, if for Z_k we have observed category levels *low*, *medium*, *low*, and *high* for four individuals, φ_k will represent the four quantifications corresponding to these levels. Using the example quantifications given in Figure 3.2 in subsection 3.2.1, the resulting vector $\varphi_k(\mathbf{Z}_{*k})$ for the four individuals is

$$\mathbf{Z}_{*k} = \begin{pmatrix} \text{Low} \\ \text{Medium} \\ \text{Low} \\ \text{High} \end{pmatrix} \rightarrow \varphi_k(\mathbf{Z}_{*k}) = \begin{pmatrix} 0 \\ 1.2 \\ 0 \\ 1.8 \end{pmatrix}.$$

In the new model, the covariate values $\mathbf{Z}_{*k}$ are replaced by their quantifications $\varphi_k(\mathbf{Z}_{*k})$ and are interpreted as numeric values, i.e. outcome y_i is modeled as

$$y_i = \sum_{k=1}^p \beta_k \varphi_k(z_{ik}) + \epsilon_i.$$

This results in the following loss function

$$L(\boldsymbol{\beta}, \boldsymbol{\varphi}) = \left\| \mathbf{y} - \sum_{k=1}^p \beta_k \varphi_k(\mathbf{Z}_{*k}) \right\|^2. \quad (3.6)$$

To find the optimal fit for this model, the loss function (3.6) should be minimized over both $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$. This minimization is done for one covariate at the time. In each step, covariate k and its regression parameter are separated from the other covariates, i.e.

$$L(\boldsymbol{\beta}, \boldsymbol{\varphi}) = \left\| \mathbf{y} - \sum_{l \neq k} \beta_l \varphi_l(\mathbf{Z}_{*l}) - \beta_k \varphi_k(\mathbf{Z}_{*k}) \right\|^2. \quad (3.7)$$

Parameters β_l and φ_l with $l \neq k$ are assumed to be fixed and optimization is performed over covariate β_k and φ_k . Therefore these terms can be merged with $\mathbf{y}$ to form a single fixed term $\mathbf{u}_k$ in the loss function, i.e.

$$\mathbf{u}_k = \mathbf{y} - \sum_{l \neq k} \beta_l \varphi_l(\mathbf{Z}_{*l}),$$

which yields

$$L(\beta_k, \varphi_k) = \|\mathbf{u}_k - \beta_k \varphi_k(\mathbf{Z}_{*k})\|^2. \quad (3.8)$$

For each categorical covariate Z_k with C_k category levels, let $\mathbf{G}_k$ be the indicator matrix with dimensions $n \times C_k$ which indicates the category levels for each of the n subjects. Row $\mathbf{G}_{ik}$ contains only zeros except in the column that refers to the category of subject i . Furthermore, define $\mathbf{v}_k$ as the vector of dimensions $C_k \times 1$, with quantifications for all category levels of Z_k . Matrix $\mathbf{G}_k$ and vector $\mathbf{v}_k$ for the example data introduced above are given as

$$\mathbf{G}_k = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{v}_k = \begin{pmatrix} 0 \\ 1.2 \\ 1.8 \end{pmatrix}.$$

From the definitions of $\mathbf{G}_k$ and $\mathbf{v}_k$, it follows that $\mathbf{G}_k \mathbf{v}_k = \varphi_k(\mathbf{Z}_{*k})$. This is shown below for the example.

$$\mathbf{G}_k \mathbf{v}_k = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1.2 \\ 1.8 \end{pmatrix} = \begin{pmatrix} 0 \\ 1.2 \\ 0 \\ 1.8 \end{pmatrix} = \varphi_k(\mathbf{Z}_{*k}).$$

Using the new notation, the loss function (3.8) can be rewritten as

$$L(\beta_k, \mathbf{v}_k) = \|\mathbf{u}_k - \beta_k \mathbf{G}_k \mathbf{v}_k\|^2. \quad (3.9)$$

The loss function (3.9) should be minimized over both β_k and $\mathbf{v}_k$. Infinite combinations of β_k and $\mathbf{v}_k$ will minimize this function. Therefore, $\mathbf{v}_k$ is standardized such that the method is restricted to finding a unique combination. Then, β_k is assumed to be fixed in order to estimate $\mathbf{v}_k$. The unrestricted quantifications $\check{\mathbf{v}}_k$ for covariate k is the least squares solution for a simple linear regression model (see subsection 3.6.1 for details)

$$\check{\mathbf{v}}_k = \beta_k^{-1} \mathbf{D}_k^{-1} \mathbf{G}_k^T \mathbf{u}_k, \quad (3.10)$$

with $\mathbf{D}_k = \mathbf{G}_k^T \mathbf{G}_k = \text{diag}(n_{k1}, \dots, n_{kC_k})$ the diagonal matrix containing the number of subjects n_{kc} in each category c for the specific covariate k , with $c = 1, \dots, C_k$. For the example illustrated before, the matrix $\mathbf{D}_k$ is as follows

$$\mathbf{D}_k = \mathbf{G}_k^T \mathbf{G}_k = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

As explained above, the unrestricted quantifications $\check{\mathbf{v}}_k$ are standardized to provide a unique solution, i.e.

$$\bar{\mathbf{v}}_k = n^{1/2} \check{\mathbf{v}}_k (\check{\mathbf{v}}_k^T \mathbf{D}_k \check{\mathbf{v}}_k)^{-1/2}. \quad (3.11)$$

The unrestricted quantifications are the parameter estimates for nominal data which do not necessarily preserve the ordering of categories. In case of ordinal data, the ordering of the categories should be preserved, i.e. the values of $\bar{\mathbf{v}}_k$ should be adjusted such that they are in the same order as their underlying category levels. In this paper, weighted monotonic regression (Kruskal, 1964) is used to find a monotonic step function which preserves the category orderings. This method uses a weighted average of the unrestricted quantifications if these are in the wrong order. The resulting restricted version of $\bar{\mathbf{v}}_k$ is denoted as $\hat{\mathbf{v}}_k$. As an example, consider the example data given before in which categories of four individuals are observed. Assume that the fit for the nominal case results in the transformation as shown in Figure 3.3a. This is not a monotone transformation, since the quantification of categories *low* and *medium* are in the wrong order. In the weighted monotone regression algorithm, a monotone transformation is made by replacing the quantifications of both these categories by their weighted average. The weighted average is calculated as in Table 3.3. The resulting monotone transformation is shown in Figure 3.3b.

	Low	Medium	High
Nominal Quantifications	2	1	3
Ordinal Quantifications	$\frac{5}{3}$	$\frac{5}{3}$	3

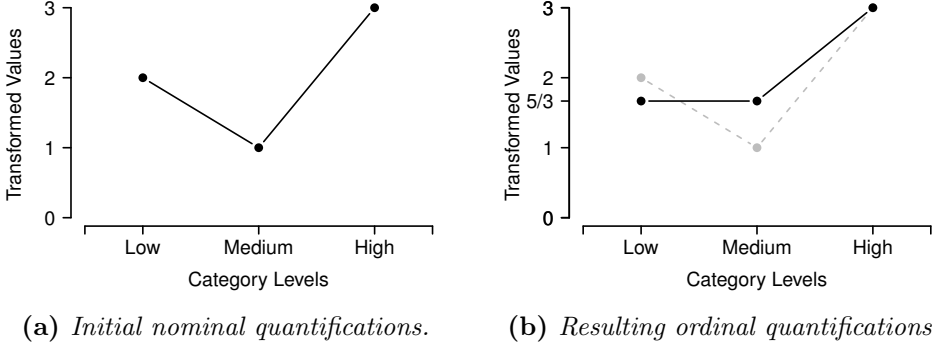
Table 3.3: *Weighted monotone regression algorithm on small data example in which the observed categories were low, low, medium, and high.*

With $\hat{\mathbf{v}}_k$ being the result of the monotone regression algorithm, i.e. the restricted version of $\bar{\mathbf{v}}_k$, it follows that $\hat{\varphi}_k(\mathbf{Z}_{*k}) = \mathbf{G}_k \hat{\mathbf{v}}_k$ is the vector of quantifications for covariate k corresponding to the n subjects.

Once the loss function (3.9) has been minimized over φ_k , the next step is to minimize this loss function over β_k . The least squares solution for β_k is derived from the ordinary least squares solution for (3.9) (see subsection 3.6.1 for details), and is estimated by

$$\hat{\beta}_k = \mathbf{u}_k^T \hat{\varphi}_k(\mathbf{Z}_{*k}). \quad (3.12)$$

Now that both quantifications $\hat{\varphi}_k(\mathbf{Z}_{*k})$ and model parameters $\hat{\beta}_k$ have been updated, the algorithm continues minimizing the loss function (3.6) step by step



(a) Initial nominal quantifications. (b) Resulting ordinal quantifications.

Figure 3.3: Initial nominal quantifications and transformed ordinal quantifications for example data, calculated with weighted monotonic regression.

minimizing over the remaining covariates, until all p covariates are updated. The process of updating the β_k 's and $\mathbf{v}_k$'s for all covariates k is repeated until convergence criteria are satisfied. The algorithm can be summarized as follows.

Optimal scaling regression algorithm:

Step 1: Initialize β_k and $\tilde{\mathbf{v}}_k$ for $k = 1, \dots, p$.

Step 2: For $k = 1, \dots, p$, do:

Step 2a: Calculate $\tilde{\mathbf{u}}_k = \mathbf{y} - \sum_{l \neq k} \tilde{\beta}_l \mathbf{G}_l \tilde{\mathbf{v}}_l$.

Step 2b: Find $\tilde{\mathbf{v}}_k$ minimizing

$$\left\| \tilde{\mathbf{u}}_k - \tilde{\beta}_k \mathbf{G}_k \tilde{\mathbf{v}}_k \right\|^2.$$

Standardize $\tilde{\mathbf{v}}_k$, and denote the standardized version by $\bar{\mathbf{v}}_k$. If covariate Z_k is ordinal, apply ordinal restrictions on $\bar{\mathbf{v}}_k$, resulting in restricted quantifications $\hat{\mathbf{v}}_k$. Set $\tilde{\mathbf{v}}_k = \hat{\mathbf{v}}_k$, and $\tilde{\varphi}_k(\mathbf{Z}_{*k}) = \mathbf{G}_k \tilde{\mathbf{v}}_k$.

Step 2c: Find $\hat{\beta}_k$ minimizing

$$\left\| \mathbf{u}_k - \beta_k \tilde{\varphi}_k(\mathbf{Z}_{*k}) \right\|^2.$$

Set $\tilde{\beta}_k = \hat{\beta}_k$.

Step 3: Repeat step 2 until convergence of $\tilde{\mathbf{v}}_k$ and $\tilde{\beta}_k$.

The algorithm is called *Alternating Least Squares*, since it alternates between minimizing the quadratic loss $\left\| \mathbf{u}_k - \beta_k \varphi_k(\mathbf{Z}_{*k}) \right\|^2$ over quantifications $\varphi_k(\mathbf{Z}_{*k})$ and model parameters β_k while keeping all other parameters constant. Note that by keeping all terms fixed except the one that is optimized, and by separating the fixed part $\mathbf{u}_k$ from the variable part $\tilde{\beta}_k \mathbf{G}_k \tilde{\mathbf{v}}_k$ of the loss function (3.9), this

method becomes an ordinary least squares problem. Merging all fixed parts and separating them from the variable part is a crucial step, because it reduces the optimization problem from the multivariate to the univariate case. The merging and separation steps described can easily be implemented in any least squares problem. Therefore, if a model is fit by using a least squares approach, then the merging and separation steps can be used to fit the model in case of categorical covariates. As a consequence, optimal scaling can be easily implemented in all models that are fitted by a least squares approach by implementing the alternating least squares algorithm.

3.3 Optimal scaling in survival analysis

As mentioned above, the optimal scaling procedure can easily be implemented for models that are fitted using a least squares algorithm. For the Cox proportional hazards regression model used in survival analysis the parameters are not fitted with a least squares approach, but by maximizing the partial likelihood. Therefore, the optimal scaling procedure as described for the simple linear regression model cannot be implemented directly.

In this paper we propose a least squares approach to fit the Cox model that includes the optimal scaling procedure. Simon et al. (2011) developed a method similar to the standard Newton-Raphson algorithm to transform the maximum likelihood approach for the Cox model into a reweighted least squares approach in order to penalize the model parameters β . In this section, we will first discuss the reweighted least squares approach for the standard Cox model setting (i.e. without penalization of the model parameters). Then, we will show how the optimal scaling approach can be included in this algorithm to find optimal quantifications for ordinal covariates in the Cox model.

3.3.1 A reweighted least squares approach to fitting the Cox model

Recall from subsection 3.2.1 the notation for survival data. Survival data for subject i are represented by the triplet $(t_i, \delta_i, \mathbf{z}_i)$, with $i = 1, \dots, n$, where n is the number of subjects. Variable t_i represents the observed time point, either the event time x_i or the censoring time c_i , i.e. $t_i = \min(x_i, c_i)$. The indicator δ_i shows whether t_i is an event ($\delta_i = 1$) or censoring time ($\delta_i = 0$). The Cox

proportional hazards model is defined as

$$\begin{aligned} h(t|\mathbf{Z}) &= h_0(t) \exp(\mathbf{Z}\boldsymbol{\beta}) \\ &= h_0(t) \exp \left[\sum_{k=1}^p \beta_k Z_k \right], \end{aligned} \quad (3.13)$$

with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$.

In this paper, we will allow for tied event times and for weighted observations. Let $t_1 < t_2 < \dots < t_D$ denote the D distinct and ordered event times and let D_m be the set of all individuals with an event at time t_m , for $m = 1, \dots, D$. If w_i denotes the weight of subject i , then let d_m be the sum of the weights of subjects who experience an event at time t_m , i.e. $d_m = \sum_{j \in D_m} w_j$. Define R_{t_m} to be the set of individuals at risk just prior to t_m . Let $\mathbf{Z}$ be the matrix of dimensions $n \times p$ of observed covariate values, as defined in subsection 3.2.2. Vector $\mathbf{Z}_{i*} = (z_{i1}, \dots, z_{ip})$ is row i of matrix $\mathbf{Z}$ and contains the p observed covariate values for individual i . Column k is defined as $\mathbf{Z}_{*k} = (z_{1k}, \dots, z_{nk})^T$ and contains the n observed values for covariate Z_k , with $k = 1, \dots, p$. Let $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta}$ be the vector of length n with elements $\eta_i = \mathbf{Z}_{i*}\boldsymbol{\beta} = z_{i1}\beta_1 + \dots + z_{ip}\beta_p$.

To fit the Cox model with ties and weighted observations, the Breslow approximation of the partial likelihood (Breslow, 1972) is used,

$$Lik(\boldsymbol{\eta}) = \prod_{m=1}^D \frac{\exp(\sum_{j \in D_m} w_j \eta_j)}{\left(\sum_{r \in R_{t_m}} w_r \exp(\eta_r) \right)^{d_m}}. \quad (3.14)$$

Maximizing this likelihood is equivalent to maximizing the log of the partial likelihood,

$$\ell(\boldsymbol{\eta}) = \sum_{m=1}^D \sum_{j \in D_m} w_j \eta_j - \sum_{m=1}^D d_m \log \left(\sum_{r \in R_{t_m}} w_r \exp(\eta_r) \right). \quad (3.15)$$

Simon et al. (2011) proposed a Newton-Raphson approach to assess the maximum of (3.15). This procedure results in a reweighted least squares problem with associated loss function

$$L(\boldsymbol{\eta}) = - \sum_{i=1}^n \omega_i(\boldsymbol{\eta}) (\zeta_i(\boldsymbol{\eta}) - \mathbf{Z}_{i*}\boldsymbol{\beta})^2, \quad (3.16)$$

where $\omega_i(\boldsymbol{\eta})$ is the i -th diagonal entry of $\ell''(\boldsymbol{\eta})$, the second partial derivative of $\ell(\boldsymbol{\eta})$ with respect to η_i , and $\zeta_i(\boldsymbol{\eta}) = \boldsymbol{\eta}_i - (\ell''(\boldsymbol{\eta})_{i,i})^{-1} \ell'(\boldsymbol{\eta})_i$. Details on how to

derive loss function (3.16) are given in subsection 3.6.2. The loss function (3.16) can be rewritten as follows

$$L(\boldsymbol{\beta}) = \|\boldsymbol{\zeta}(\boldsymbol{\eta}) - \mathbf{Z}\boldsymbol{\beta}\|_{\boldsymbol{\Omega}(\boldsymbol{\eta})}^2, \quad (3.17)$$

where $\boldsymbol{\zeta}(\boldsymbol{\eta})$ is the vector of length n with elements $\zeta_i(\boldsymbol{\eta})$, and $\boldsymbol{\Omega}(\boldsymbol{\eta})$ is the diagonal matrix with elements $(-\omega_1(\boldsymbol{\eta}), \dots, -\omega_n(\boldsymbol{\eta}))$. To calculate the $\omega_i(\boldsymbol{\eta})$'s and $\zeta_i(\boldsymbol{\eta})$'s for $i = 1, \dots, n$, we need the first and second partial derivatives of log likelihood $\ell(\boldsymbol{\eta})$ with respect to η_i . Details about the derivatives are given in subsection 3.6.3. The first partial derivative of $\ell(\boldsymbol{\eta})$ with respect to η_i is

$$\ell'(\boldsymbol{\eta})_i = \delta_i w_i - \sum_{s \in S_i} \frac{d_s w_i \exp(\eta_i)}{\sum_{r \in R_{t_s}} w_r \exp(\eta_r)}, \quad (3.18)$$

where S_i is the set of all individuals s that experience the event before the observed time point of person i , i.e. $\delta_s = 1$ and $t_s \leq t_i$. Second partial derivative of $\ell(\boldsymbol{\eta})$ is

$$\ell''(\boldsymbol{\eta})_{i,i} = - \sum_{s \in S_i} d_s \frac{w_i \exp(\eta_i) \sum_{r \in R_{t_s}} w_r \exp(\eta_r) - (w_i \exp(\eta_i))^2}{(\sum_{r \in R_{t_s}} w_r \exp(\eta_r))^2}. \quad (3.19)$$

The first and second derivatives, (3.18) and (3.19), can be used to find explicit formulas for $\omega_i(\boldsymbol{\eta})$ and $\zeta_i(\boldsymbol{\eta})$, yielding

$$\begin{aligned} \omega_i(\boldsymbol{\eta}) &= \ell''(\boldsymbol{\eta})_{i,i} \\ &= - \sum_{s \in S_i} d_s \frac{w_i \exp(\eta_i) \sum_{r \in R_{t_s}} w_r \exp(\eta_r) - (w_i \exp(\eta_i))^2}{(\sum_{r \in R_{t_s}} w_r \exp(\eta_r))^2} \end{aligned} \quad (3.20)$$

and

$$\begin{aligned} \zeta_i(\boldsymbol{\eta}) &= \eta_i - \frac{\ell'(\boldsymbol{\eta})_i}{\ell''(\boldsymbol{\eta})_{i,i}} \\ &= \eta_i - \frac{1}{\omega_i(\boldsymbol{\eta})} \left(\delta_i w_i - \sum_{s \in S_i} \frac{d_s w_i \exp(\eta_i)}{\sum_{r \in R_{t_s}} w_r \exp(\eta_r)} \right). \end{aligned} \quad (3.21)$$

Therefore, to maximize the likelihood (3.14), the loss function (3.17) should be minimized over the regression coefficients $\boldsymbol{\beta}$.

3.3.2 Including optimal scaling in the reweighted least squares algorithm for the Cox model

To find optimal quantifications for category levels of the p covariates, we can include the optimal scaling procedure into the reweighted least squares algorithm described in subsection 3.3.1. The first step is to replace the covariates $\mathbf{Z}$ by

quantifications $\boldsymbol{\varphi}(\mathbf{Z}) = (\varphi_1(\mathbf{Z}_{*1}), \dots, \varphi_p(\mathbf{Z}_{*p}))$. Hence, the Cox proportional hazards model with quantifications is now defined as

$$h(t|\mathbf{Z}) = h_0(t) \exp(\boldsymbol{\varphi}(\mathbf{Z})\boldsymbol{\beta}). \quad (3.22)$$

By defining $\boldsymbol{\eta}^* = \boldsymbol{\varphi}(\mathbf{Z})\boldsymbol{\beta}$ as the vector of length n with elements $\eta_i^* = \sum_{k=1}^p \varphi_k(z_{ik})\beta_k$, the partial likelihood can easily be extended to the case with quantified variables. This results in

$$Lik(\boldsymbol{\eta}^*) = \prod_{m=1}^D \frac{\exp(\sum_{j \in D_m} w_j \eta_j^*)}{\left(\sum_{r \in R_{t_m}} w_r \exp(\eta_r^*)\right)^{d_m}}. \quad (3.23)$$

This likelihood can be maximized by maximizing its log,

$$\ell(\boldsymbol{\eta}^*) = \sum_{m=1}^D \sum_{j \in D_m} w_j \eta_j^* - \sum_{m=1}^D d_m \log \left(\sum_{r \in R_{t_m}} w_r \exp(\eta_r^*) \right),$$

which can then be translated in a reweighted least squares problem with associated loss function

$$L(\boldsymbol{\eta}^*) = - \sum_{i=1}^n \omega_i(\boldsymbol{\eta}^*) (\zeta_i(\boldsymbol{\eta}^*) - \boldsymbol{\varphi}(\mathbf{Z}_{i*})\boldsymbol{\beta})^2, \quad (3.24)$$

where

$$\begin{aligned} \omega_i(\boldsymbol{\eta}^*) &= \ell''(\boldsymbol{\eta}^*)_{i,i} \\ &= - \sum_{s \in S_i} d_s \frac{w_i \exp(\eta_i^*) \sum_{r \in R_{t_s}} w_r \exp(\eta_r^*) - (w_i \exp(\eta_i^*))^2}{\left(\sum_{r \in R_{t_s}} w_r \exp(\eta_r^*)\right)^2}, \end{aligned} \quad (3.25)$$

and

$$\begin{aligned} \zeta_i(\boldsymbol{\eta}^*) &= \eta_i^* - \frac{\ell'(\boldsymbol{\eta}^*)_i}{\ell''(\boldsymbol{\eta}^*)_{i,i}} \\ &= \eta_i^* - \frac{1}{\omega_i(\boldsymbol{\eta}^*)} \left(\delta_i w_i - \sum_{s \in S_i} \frac{d_s w_i \exp(\eta_i^*)}{\sum_{r \in R_{t_s}} w_r \exp(\eta_r^*)} \right). \end{aligned} \quad (3.26)$$

The loss function (3.24) which includes the quantifications can be rewritten as

$$L(\boldsymbol{\beta}, \boldsymbol{\varphi}) = \|\boldsymbol{\zeta}(\boldsymbol{\eta}^*) - \boldsymbol{\varphi}(\mathbf{Z})\boldsymbol{\beta}\|_{\boldsymbol{\Omega}(\boldsymbol{\eta}^*)}^2, \quad (3.27)$$

where $\boldsymbol{\zeta}(\boldsymbol{\eta}^*)$ is the vector of length n with elements $\zeta_i(\boldsymbol{\eta}^*)$, and $\boldsymbol{\Omega}(\boldsymbol{\eta}^*)$ is the diagonal matrix with elements $(-\omega_1(\boldsymbol{\eta}^*), \dots, -\omega_n(\boldsymbol{\eta}^*))$.

In this optimal scaling setting for the Cox model, the loss function (3.27) has to be minimized over both the regression coefficients $\boldsymbol{\beta}$ and the set of

transformations φ . Since the problem has been transformed into a reweighted least squares problem, almost the same methodology can be applied to optimize the loss function over both β and φ as used for OS regression. Again, the loss function is optimized over one covariate Z_k at the time, and alternating between optimizing $\varphi_k(\mathbf{Z}_{*k})$ and optimizing β_k while assuming all other terms fixed. Similar to OS in regression, in each step, the quantifications $\varphi_k(\mathbf{Z}_{*k})$ and regression parameter β_k are separated from the linear combination of the other predictors $\sum_{l \neq k} \beta_l \varphi_l(\mathbf{Z}_{*l})$. Therefore, (3.27) can be rewritten as

$$L(\beta, \varphi) = \left\| \zeta(\eta^*) - \sum_{l \neq k} \beta_l \varphi_l(\mathbf{Z}_{*l}) - \beta_k \varphi_k(\mathbf{Z}_{*k}) \right\|_{\Omega(\eta^*)}^2. \quad (3.28)$$

By assuming all terms, except β_k and φ_k , fixed and merging all these fixed terms into one term defined as $\mathbf{u}_k = \zeta(\eta^*) - \sum_{l \neq k} \beta_l \varphi_l(\mathbf{Z}_{*l})$, (3.28) becomes

$$L(\beta_k, \varphi_k) = \|\mathbf{u}_k - \beta_k \varphi_k(\mathbf{Z}_{*k})\|_{\Omega(\eta^*)}^2. \quad (3.29)$$

By introducing the indicator matrix $\mathbf{G}_k$ to show the category levels of covariate Z_k for all individuals and $\mathbf{v}_k$ the vector of quantifications as defined for the simple linear regression case, (3.29) can be rewritten as

$$L(\beta_k, \varphi_k) = \|\mathbf{u}_k - \beta_k \mathbf{G}_k \mathbf{v}_k\|_{\Omega(\eta^*)}^2. \quad (3.30)$$

As in OS regression, infinite combinations of β_k and $\mathbf{v}_k$ minimize this loss function. Hence, $\mathbf{v}_k$ is standardized in order to find a unique solution.

The first step is to find the unrestricted quantifications $\check{\mathbf{v}}_k$ that minimize the loss function (3.30) while keeping β_k constant. This estimate is given by the univariate weighted least square solution, i.e.

$$\check{\mathbf{v}}_k = \beta_k^{-1} \mathbf{D}_k^{-1} \mathbf{G}_k^T \Omega(\eta^*) \mathbf{u}_k.$$

Let $\bar{\mathbf{v}}_k$ be the standardized version of $\check{\mathbf{v}}_k$, as defined in (3.11). Using the same methods as for OS in regression, the restricted version of $\bar{\mathbf{v}}_k$ is determined, and is denoted by $\hat{\mathbf{v}}_k$. Then $\hat{\varphi}_k(\mathbf{Z}_{*k}) = \mathbf{G}_k \hat{\mathbf{v}}_k$ contains the current estimated quantifications for covariate Z_k .

In the next step, loss function (3.30) is minimized over β_k while keeping $\mathbf{v}_k$ constant. This parameter is estimated by using the univariate weighted least squares solution

$$\hat{\beta}_k = (\varphi_k(\mathbf{Z}_{*k})^T \Omega(\eta^*)^T \varphi_k(\mathbf{Z}_{*k}))^{-1} \varphi_k(\mathbf{Z}_{*k})^T \Omega(\eta^*) \mathbf{u}_k. \quad (3.31)$$

Once both $\widehat{\mathbf{v}}_k$ and $\widehat{\beta}_k$ have been updated, the algorithm moves to the next covariate. After updating the parameters for all covariates, convergence is checked by using the stopping rule

$$\max_k A_k^*(\widetilde{\beta}_k^{old} - \widetilde{\beta}_k^{new}) < \epsilon^2, \quad (3.32)$$

with

$$A_k^* = \sum_{m=1}^D \frac{1}{4n} \left(\max_{r \in R_{tm}} (\varphi_k(Z_{k_r})) - \min_{r \in R_{tm}} (\varphi_k(Z_{k_r})) \right)^2,$$

and ϵ a convergence parameter defined by the user (Yang and Zou, 2013). The optimal scaling algorithm for the Cox proportional hazards model is summarized below.

Optimal scaling algorithm for Cox' proportional hazards model:

Step 1: Initialize $\widetilde{\beta}_k$ and $\widetilde{\mathbf{v}}_k$ for $k = 1, \dots, p$.

Step 2: For $k = 1, \dots, p$, do:

Step 2a: Compute $\ell'(\widetilde{\boldsymbol{\eta}}^*)$ and $\ell''(\widetilde{\boldsymbol{\eta}}^*)$, and use these quantities to derive $\omega(\widetilde{\boldsymbol{\eta}}^*)$ and $\zeta(\widetilde{\boldsymbol{\eta}}^*)$.

Step 2b: Calculate $\mathbf{u}_k = \zeta(\boldsymbol{\eta}^*) - \sum_{l \neq k} \beta_l \varphi_l(\mathbf{Z}_{*l})$.

Step 2c: Find $\widetilde{\mathbf{v}}_k$ minimizing

$$\left\| \mathbf{u}_k - \widetilde{\beta}_k \mathbf{G}_k \mathbf{v}_k \right\|_{\Omega(\boldsymbol{\eta}^*)}^2.$$

Standardize $\widetilde{\mathbf{v}}_k$, and denote the standardized version by $\bar{\mathbf{v}}_k$. If covariate Z_k is ordinal, apply ordinal restrictions on $\bar{\mathbf{v}}_k$, resulting in restricted quantifications $\widehat{\mathbf{v}}_k$. Set $\widetilde{\mathbf{v}}_k = \widehat{\mathbf{v}}_k$, and $\widetilde{\varphi}_k(\mathbf{Z}_{*k}) = \mathbf{G}_k \widetilde{\mathbf{v}}_k$.

Step 2d: Find $\widetilde{\beta}_k$ minimizing

$$\left\| \mathbf{u}_k - \beta_k \widetilde{\varphi}_k(\mathbf{Z}_{*k}) \right\|_{\Omega(\boldsymbol{\eta}^*)}^2.$$

Set $\widetilde{\beta}_k = \widehat{\beta}_k$.

Step 3: Repeat step 2 until convergence criterium (3.32) is met.

3.4 Simulation study

A large simulation study was done to investigate the performance of the optimal scaling method for survival analysis proposed in this paper. The new method is compared with the two currently used methods: dummy and integer coding.

Z	Z_0	Z_1	$\cdots$	Z_{C-1}
0	1	0	$\cdots$	0
1	0	1	$\cdots$	0
$\vdots$			$\vdots$	
$C-1$	0	0	$\cdots$	1

Table 3.4: Coding corresponding to the C categories to generate data.

To investigate the performance of the different methods, several scenarios were simulated. We investigated the effect of a nonlinear monotone increasing set of model parameters $\beta_{Z_0}, \beta_{Z_1}, \dots, \beta_{Z_{C-1}}$, different sample sizes, and different percentages of censored subjects. In this section, first an overview of the set up of the simulation study is given, and then results coming from several simulation scenarios are discussed.

3.4.1 Set up of simulation study

In this section we will illustrate how the survival data were generated, how the three models were fitted, and how results were compared.

Generating the data

For this simulation study, we generated n subjects with a single categorical covariate Z with C category levels $0, 1, \dots, C-1$. For each subject i (with $i = 1, \dots, n$), one category level was sampled and denoted as z_i , i.e. $z_i \in \{0, 1, \dots, C-1\}$. Event times X and censoring times C were sampled from an exponential distributed with constant hazards $h_{X|Z}$ and h_C respectively, i.e. $X \sim \exp(h_{X|Z})$ and $C \sim \exp(h_C)$. The hazard $h_{X|Z}$ is related to the covariate as defined in the Cox proportional hazards model

$$h_{X|Z} = h_{0_X} \exp(\beta Z), \quad (3.33)$$

where the baseline hazard h_{0_X} is assumed to be constant over time. It is also assumed that h_C in the censoring model is constant over time. However, this parameter is independent from the covariate Z , i.e. it is equal to the baseline hazard of being censored,

$$h_C = h_{0_C}. \quad (3.34)$$

The coding used to generate the data sets is shown in Table 3.4. This coding system results in the following hazards for time to event X for each category level

Z	D_1	D_2	$\dots$	D_{C-1}
0	0	0	$\dots$	0
1	1	0	$\dots$	0
$\vdots$			$\vdots$	
$C-1$	0	0	$\dots$	1

Table 3.5: *Dummy coding of C simulated categories.*

$$\begin{aligned}
h(t|Z=0) &= h_0(t) \exp(1\beta_{Z_0} + 0\beta_{Z_1} + \dots 0\beta_{Z_{C-1}}) = h_0(t) \exp(\beta_{Z_0}) \\
h(t|Z=1) &= h_0(t) \exp(0\beta_{Z_0} + 1\beta_{Z_1} + \dots 0\beta_{Z_{C-1}}) = h_0(t) \exp(\beta_{Z_1}) \\
&\vdots \\
h(t|Z=C-1) &= h_0(t) \exp(0\beta_{Z_0} + 0\beta_{Z_1} + \dots 1\beta_{Z_{C-1}}) = h_0(t) \exp(\beta_{Z_{C-1}}).
\end{aligned}$$

To make Z an ordinal categorical variable for which the effect on the hazard rate is increasing with the category levels, we can choose the parameters for each category level in an increasing way, i.e. such that $\beta_{Z_0} \leq \beta_{Z_1} \leq \dots \leq \beta_{Z_{C-1}}$. Another option to generate an ordinal covariate is to choose the parameters such that the effect always decreases with category levels, i.e. $\beta_{Z_0} \geq \beta_{Z_1} \geq \dots \geq \beta_{Z_{C-1}}$. In this way the effects of the category levels are still ordered, and the hazard rate decreases with category levels. Hence, any monotone increasing or decreasing function can be chosen to simulate an ordinal covariate.

For each observation i , the event time x_i and censoring time c_i are generated from exponential distributions with parameter $h_{X|z_i} = h(t|Z = z_i)$ and h_{0C} respectively. Actually, only the first of these time points is observed. Hence, the observed time point can be calculated as $t_i = \min(x_i, c_i)$, and the status indicator $\delta_i = \mathbf{1}_{\{t_i = x_i\}}$ is used to indicate whether the observed time point is an event ($\delta_i = 1$) or censoring time ($\delta_i = 0$).

Hence, for each observation i we have generated category level z_i , event time x_i , censoring time c_i , observed time point t_i , and status indicator δ_i . The triple (t_i, δ_i, z_i) will be used to fit the survival models.

Fitting the models

The generated survival data is used to fit the Cox proportional hazards model with three different methods to incorporate the ordinal categorical covariate. When applying the dummy coding method, $C-1$ dummy covariates are generated (see Table 3.5), and the model parameters $\beta_{D_1}, \dots, \beta_{D_{C-1}}$ are estimated for each of these dummies. In total $C-1$ parameters are estimated for the linear predictor in this model. The estimated hazards for each of the category levels are

$$\begin{aligned}
\hat{h}(t|Z=0) &= \hat{h}_{0_D}(t) \exp(0\hat{\beta}_{D_1} + 0\hat{\beta}_{D_2} + \dots 0\hat{\beta}_{D_{C-1}}) = \hat{h}_{0_D}(t) \\
\hat{h}(t|Z=1) &= \hat{h}_{0_D}(t) \exp(0\hat{\beta}_{D_1} + 1\hat{\beta}_{D_2} + \dots 0\hat{\beta}_{D_{C-1}}) = \hat{h}_{0_D}(t) \exp(\hat{\beta}_{D_1}) \\
&\vdots \\
\hat{h}(t|Z=C-1) &= \hat{h}_{0_D}(t) \exp(0\hat{\beta}_{D_1} + 0\hat{\beta}_{D_2} + \dots 1\hat{\beta}_{D_{C-1}}) = \hat{h}_{0_D}(t) \exp(\hat{\beta}_{D_{C-1}}).
\end{aligned}$$

For each simulation, the model is fitted with the standard procedures currently used for nominal data. Therefore, when fitting the model there are no ordering restrictions on the model parameters, i.e. $\hat{\beta}_{D_1} \leq \hat{\beta}_{D_2} \leq \dots \leq \hat{\beta}_{D_{C-1}}$ is not required.

For the method of integer coding, the category levels are given integer values $(0, \dots, C-1)$ which are interpreted as numeric values. There is only one parameter in the linear predictor, namely β_{int} . The estimated hazards are

$$\begin{aligned}
\hat{h}(t|Z=0) &= \hat{h}_{0_{int}}(t) \exp(0\hat{\beta}_{int}) = \hat{h}_{0_{int}}(t) \\
\hat{h}(t|Z=1) &= \hat{h}_{0_{int}}(t) \exp(1\hat{\beta}_{int}) = \hat{h}_{0_{int}}(t) \exp(1\hat{\beta}_{int}) \\
&\vdots \\
\hat{h}(t|Z=C-1) &= \hat{h}_{0_{int}}(t) \exp((C-1)\hat{\beta}_{int}) = \hat{h}_{0_{int}}(t) \exp((C-1)\hat{\beta}_{int}).
\end{aligned}$$

The parameter β_{int} is estimated with the standard Cox procedures used for survival analysis with numeric covariates.

The hazards estimated in the optimal scaling method contain the parameter β_{os} and C quantifications $(\varphi(0), \varphi(1), \dots, \varphi(C-1))$, one for each category level. These quantifications are interpreted as numeric values, so the estimated hazards for this method are

$$\begin{aligned}
\hat{h}(t|Z=0) &= \hat{h}_{0_{os}}(t) \exp(\hat{\varphi}(0)\hat{\beta}_{os}) \\
\hat{h}(t|Z=1) &= \hat{h}_{0_{os}}(t) \exp(\hat{\varphi}(1)\hat{\beta}_{os}) \\
&\vdots \\
\hat{h}(t|Z=C-1) &= \hat{h}_{0_{os}}(t) \exp(\hat{\varphi}(C-1)\hat{\beta}_{os}).
\end{aligned}$$

The parameters are estimated with the alternating least squares procedure described in subsection 3.3.2. The quantifications will be estimated such that restriction $\hat{\varphi}(0) \leq \hat{\varphi}(1) \leq \dots \leq \hat{\varphi}(C-1)$ holds.

Comparing performance

Direct comparison of the model parameters estimated by the three different methods is not possible, since the parameters do not have the same interpretation in each of the three models (see Table 3.6). We will therefore instead compare the estimated hazard ratios for category c vs category $c-1$, for $c = 1, \dots, C-1$, from dummy coding, integer coding, and optimal scaling with the hazard ratios from the true model underlying the data. The hazard ratio between category

Z	Dummy Coding	Integer Coding	Optimal Scaling
0	$\hat{h}_{0_D}(t)$	$\hat{h}_{0_{int}}(t)$	$\hat{h}_{0_{os}}(t) \exp(\hat{\varphi}(0) \hat{\beta}_{os})$
1	$\hat{h}_{0_D}(t) \exp(\hat{\beta}_{D_1})$	$\hat{h}_{0_{int}}(t) \exp(1 \hat{\beta}_{int})$	$\hat{h}_{0_{os}}(t) \exp(\hat{\varphi}(1) \hat{\beta}_{os})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C-1$	$\hat{h}_{0_D}(t) \exp(\hat{\beta}_{D_{C-1}})$	$\hat{h}_{0_{int}}(t) \exp((C-1) \hat{\beta}_{int})$	$\hat{h}_{0_{os}}(t) \exp(\hat{\varphi}(C-1) \hat{\beta}_{os})$

Table 3.6: Hazards estimated for each category level by each of the three methods, dummy coding, integer coding and optimal scaling.

c vs $c-1$	$\widehat{HR}_{D(c \text{ vs } c-1)}$	$\widehat{HR}_{int(c \text{ vs } c-1)}$	$\widehat{HR}_{os(c \text{ vs } c-1)}$
1 vs 0	$\exp(\hat{\beta}_{D_1})$	$\exp(\hat{\beta}_{int})$	$\exp((\hat{\varphi}(1) - \hat{\varphi}(0)) \hat{\beta}_{os})$
2 vs 1	$\exp(\hat{\beta}_{D_2} - \hat{\beta}_{D_1})$	$\exp(\hat{\beta}_{int})$	$\exp((\hat{\varphi}(2) - \hat{\varphi}(1)) \hat{\beta}_{os})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$C-1$ vs $C-2$	$\exp(\hat{\beta}_{D_{C-1}} - \hat{\beta}_{D_{C-2}})$	$\exp(\hat{\beta}_{int})$	$\exp((\hat{\varphi}(C-1) - \hat{\varphi}(C-2)) \hat{\beta}_{os})$

Table 3.7: Hazard ratios between categories c and $c-1$, for $c = 1, \dots, C-1$ estimated with the three methods, dummy coding ($\widehat{HR}_{D(c \text{ vs } c-1)}$), integer coding ($\widehat{HR}_{int(c \text{ vs } c-1)}$), and optimal scaling ($\widehat{HR}_{os(c \text{ vs } c-1)}$).

level c and $c-1$ for the true underlying model from which the data is generated is

$$HR_{(c \text{ vs } c-1)} = \frac{h_0(t) \exp(\beta_{Z_c})}{h_0(t) \exp(\beta_{Z_{c-1}})} = \exp(\beta_{Z_c} - \beta_{Z_{c-1}}). \quad (3.35)$$

The hazard ratios estimated by the three methods are given in Table 3.7 and can be compared to the hazard ratio in (3.35).

3.4.2 Results

To investigate the performance of the three methods, several scenarios were chosen to simulate the data. First, we chose a nonlinear monotone increasing function for the parameter set $\beta_{Z_0}, \beta_{Z_1}, \dots, \beta_{Z_{C-1}}$ and looked at the performance of the three methods in this scenario. Then, we increased the sample size from 100 to 500 to study the effect of sample size on the model fit. Finally, we increased the percentage of censored subjects in the dataset from 35% to 60% to investigate the effects of missing information on the model fit. In each scenario, $M = 10,000$ datasets are simulated and the three methods are applied to each dataset. Parameter settings for each of the simulation scenarios are given in

	n	β 's	h_{0_X}	h_{0_C}
Scenario 1	100	as in Figure 3.4	0.015	0.02
Scenario 2	500	as in Figure 3.4	0.015	0.02
Scenario 3	100	as in Figure 3.4	0.015	0.07

Table 3.8: Chosen parameters in each scenario of the simulation study. Bold values indicate the changes compared to baseline scenario 1.

Table 3.8. As discussed in subsubsection 3.4.1, the hazard ratios estimated by each method will be used to compare their performance. The results of all M datasets are summarized by box plots of the estimated log hazard ratios.

Monotonically increasing model parameters

The parameters $\beta_{Z_0}, \beta_{Z_1}, \dots, \beta_{Z_{C-1}}$ determine the strength of the relation between the category levels $0, 1, \dots, C-1$ and event time X . As explained in subsubsection 3.4.1, a monotone increasing or decreasing function should be chosen for these parameters to simulate ordinal category levels. We restricted this simulation study to the nonlinear monotone increasing transformation as shown in Figure 3.4, for 7 categories ($C = 7$). In the first scenario, we set the simulation parameters to the values as given in Table 3.8, which resulted in approximately 35% censoring in each data set. Box plots of the log hazard ratios

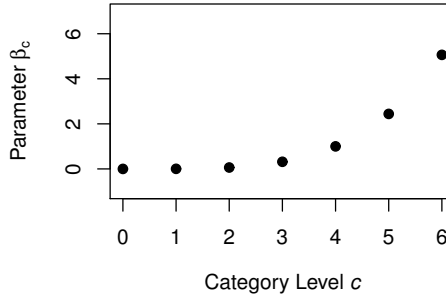


Figure 3.4: Simulation parameters $\beta_{Z_0}, \beta_{Z_1}, \dots, \beta_{Z_6}$.

estimated by the three methods in the M simulated data sets are shown in Figure 3.5. Results for this first scenario are shown by the upper set of box plots in each figure. These box plots show that for the first category levels, the integer coding system overestimates the log hazard ratio between consecutive categories, while for the highest levels, it underestimates this ratio. This can be explained

by considering the restrictions put on the category levels when assuming the integer values of the categories to be numeric. Since numeric restrictions are assumed, the integer coding method will approximate the true β_{Z_c} values by a linear function, i.e. the method assumes that the difference between consecutive β_{Z_c} 's is constant. Therefore, it will use the categories' average difference in the data as estimate in the model. In the true data the difference is very small for the low categories and increases with the category levels, so the integer coding method overestimates $\beta_{Z_c} - \beta_{Z_{c-1}}$ for low c 's, and underestimates the difference for large c 's. This is supported by the box plots in Figure 3.5. Summarizing, the integer coding method gives biased results, i.e. the numeric restriction on the integer values is too strict in case the differences $\beta_{Z_c} - \beta_{Z_{c-1}}$ are nonlinearly increasing with category level c .

The box plots corresponding to the dummy coding methods show that the average log hazard ratios are quite close to the true values. The box plots also show that the log hazard ratios are regularly estimated to be negative by the dummy coding method, especially for the first four category pairs. A negative log hazard ratio indicates that $\hat{\beta}_{D_{c-1}} \geq \hat{\beta}_{D_c}$, which means that the ordering of the model parameters is wrong. This happens because the dummy coding does not apply the restriction of $\hat{\beta}_{D_1} \leq \hat{\beta}_{D_2} \leq \dots \leq \hat{\beta}_{D_6}$. A logical explanation why the dummy coding method regularly gives the wrong ordering for the first four categories, but the correct ordering for the last three categories, is that the differences between hazard ratios increases with the category levels. For the first category levels the difference between the parameter values are very small, and therefore the correct ordering corresponding to these category levels is more difficult to detect. For the last category levels the difference is large, so the dummy coding system can easily detect the correct ordering.

The results based on the optimal scaling method show that the average log hazard ratios estimated by the optimal scaling method are quite close to the true log hazard ratio. All log hazard ratios estimated with the optimal scaling method are nonnegative because of the ordering restriction, since $\beta_{Z_c} \leq \beta_{Z_{c-1}}$ for $c = 1, \dots, C - 1$. This is also clear from the box plots truncated at 0. The ordinal restrictions on the β_{Z_c} 's also results in less variation for the category levels where the difference $\beta_{Z_c} - \beta_{Z_{c-1}}$ is small. For large differences $\beta_{Z_c} - \beta_{Z_{c-1}}$, the optimal scaling and the dummy coding provide equal results. This is because optimal scaling starts with the dummy coding result, and in case this result satisfies the ordering restriction (which it does for the higher category levels), it does not change this result (see in Figure 3.5 the box plots for category levels higher than 5).



Increasing sample size

To study the effect of increasing sample size n , we also considered n equal to 500. All the other parameters remained constant in this second scenario, see Table 3.8. The results of n equal to 100 and 500 (Scenarios 1 and 2) are shown in the upper and middle box plots in each subfigure in Figure 3.5.

When the sample size of a dataset is large, there is more observed information for each of the category levels, and hence the underlying model can be estimated more precisely. Therefore, the variation of the estimated log hazard ratios decreases with sample size for all three methods. This is confirmed by the more compact box plots for $n = 500$ compared to for $n = 100$ (Scenario 2 versus Scenario 1 in Figure 3.5). The dummy coding method shows a better performance for the large sample size, since there are more observations in each category, which makes it easier to detect the correct ordering of consecutive category levels. Simulation results suggest that increasing the sample size reduces the variation of the estimated log hazard ratios. This has a positive effect on the results from the dummy coding and optimal scaling method, since all estimated log hazard ratios are closer to their true values. However, it does not improve the results from the integer coding method. For this method, there is also less variation in the log hazard ratios, but the estimated values are still biased.

Increasing censoring percentage

A typical characteristic of survival data is that some subjects in the dataset are censored over time. This prevents us from observing the event time. Since censoring is a characteristic of survival data, we have studied the effect of an increasing censoring percentage on the estimated parameters. We choose censoring percentages equal to 35% and 60%. All other covariates were chosen as in the first scenario, see Table 3.8. Results corresponding to 35% and 60% censoring are shown by the upper and lower box plots (Scenarios 1 and 3) in Figure 3.5.

When subjects are censored, we have less information, and this will affect the precision of the estimated parameters. Hence, the higher the percentage of censored subjects, the worse the estimated parameters. This is confirmed by the increased variation in estimated log hazard ratios for the three methods, shown in Figure 3.5. Results based on this simulation study indicate as before that integer coding shows the worst performance. The advantages of optimal scaling compared to dummy coding are again more visible for the category levels whose parameters are close to the parameters of their neighbouring category levels.

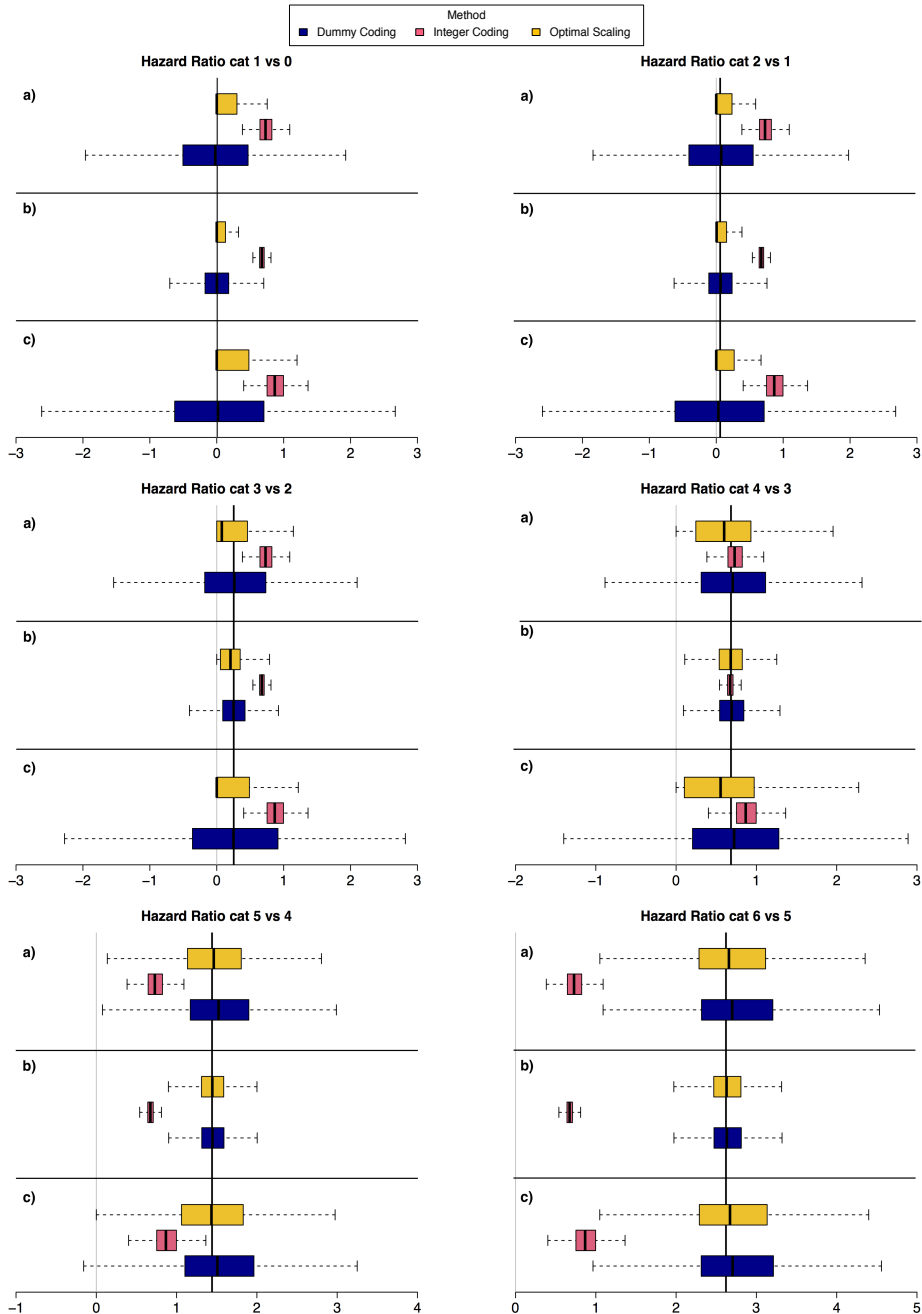


Figure 3.5: Box plots of estimated log hazard ratios between consecutive categories based on the three methods (see Table 3.7) for each of the three scenarios; a) Scenario 1, b) Scenario 2, and c) Scenario 3. Simulation parameters for each scenario were chosen as in Table 3.8. The black vertical lines indicate the true log hazard ratio, derived from the model parameters given in (3.35).

3.5 Discussion

In many studies, categorical variables are collected and used as predictors to model an outcome of interest. Often, the category levels of the data have an ordering. In medical research many scales are used to assess the severity of a disease. Pain intensity, quality of life, and modified Rankin scales are just three among a broad range of scales. For many of these scales it is expected that they have a monotone relation with the time to an event of interest. The modified Ranking Scale (mRS) is an example of this type of scales. It indicates the degree of disability or dependence in daily activities of patients who suffer from neurological disabilities. This scale is a good indicator for the medical rehabilitation needs of a patient. Patients with few disability complaints score low on this scale, and are expected to have a short rehabilitation process. Patients who are disabled severely (highest possible score on the scale) have a long rehabilitation process ahead of them. This indicates that mRS scores and medical rehabilitation time are monotonically related.

The two currently used methods to implement ordinal categorical data in the Cox proportional hazards model, dummy and integer coding, do not preserve the characteristics of this type of data. Dummy coding will not guarantee the correct ordering of the category levels. Integer coding will, but it will also force equal distances between consecutive category levels.

In this paper we have described the method of optimal scaling, in which numerical representations (quantifications) are estimated for each category level of the data. These quantifications can then be used as numerical input for the model. Restrictions can be put on the quantifications such that the data characteristics are preserved by their numerical representations. Optimal scaling is already used in several regression models that are fitted with a least squares approach. In this paper, we have described how the maximum likelihood approach to fit the Cox model can be transformed into a reweighted least squares approach in which the optimal scaling steps can be implemented.

A simulation study was carried out in order to assess the performance of the optimal scaling method in case event times are dependent on an ordinal categorical covariate. Data were generated for different scenarios by increasing sample size and censoring percentage. Simulation results suggest that the integer coding method will provide biased results when the parameters are not linearly increasing with category levels. It will estimate the parameters to be close to the linear regression line for the true parameters, and will therefore not find the nonlinearity. Dummy coding gives results quite close to the optimal scaling method in case the difference between consecutive parameter values is large. However, in case this difference is small, the dummy coding method may fail to detect the correct ordering. Since optimal scaling puts restriction on the

estimated model parameters, it will always find the correct ordering. For large sample sizes, both the dummy coding and optimal scaling method provide more accurate results. However, if the censoring percentage is increased, there will be more variation in the estimated log hazard ratios, due to the higher rate of missing information.

To get more insight into the performance of the three methods, the simulations study can be further extended. In this simulation study, we have only looked at a specific set of parameters. One could investigate the effect of other nonlinear monotone relations, or could decrease or increase the number of category levels. Furthermore, we have restricted the simulation study to a Cox model with only a single covariate associated to the occurrence of the event of interest. The model could also be extended to include more ordinal, nominal and/or numerical covariates. To assess the performance of the methods, one could also compare other outcomes than the log hazard ratios, for example the prediction error.

We think that currently in survival analysis, too little attention is given to ordinal categorical data. The two currently used methods to implement this type of data into the Cox model do not guarantee that data characteristics are preserved. Researchers should consider using the optimal scaling method discussed here to implement ordinal data in the Cox proportional hazards model correctly when they expect a monotone relation between category levels and the event times.

Note that application of optimal scaling in survival analysis is a new concept and this paper is the first step of our research. We plan to implement more techniques that use optimal scaling for the Cox proportional hazards model. For example, we would like to apply optimal scaling to reduce the dimensionality of high dimensional survival data with ordinal covariates. Scales are often used in the medical field to assess a patient's status according to p different characteristics, denoted as $Z_1, \dots, Z_p$. The dimension is then reduced by summarizing the category choices of the scales into one or more composite scores. These composite scores can then be used in the statistical model to predict outcome Y of a regression model (see Figure 3.6 for a representation of reduction to one dimension (one composite score) for a regression model).

Principal component analysis (PCA) can be used to reduce the dimensionality of the data. The resulting principal components are used as predictors in the regression model. Categorical PCA (Linting et al., 2007; Meulman et al., 2004), the PCA method that includes an optimal scaling step, can be used to find optimal scores that preserve the characteristics of ordinal covariates, i.e. it allows for nonlinear quantifications for the category levels of the scales. The next goal in our research is to extend the optimal scaling procedure described in this paper to analyze high dimensional survival data as described above. We aim



to extend the categorical optimal scaling technique to the Cox model to find optimal quantifications of the scales to better predict the event time X .

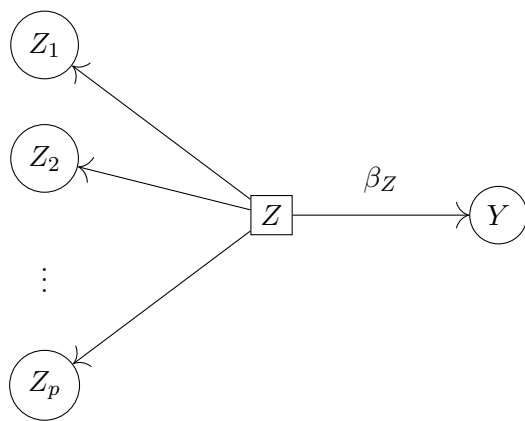


Figure 3.6: Representation of dimension reduction into one dimension for the regression model.

3.6 Supplementary material

3.6.1 Optimal scaling in ordinary linear regression

Recall the loss function

$$L(\beta_k, \mathbf{v}_k) = \|\mathbf{u}_k - \beta_k \mathbf{G}_k \mathbf{v}_k\|^2 \quad (3.36)$$

that corresponds to the ordinary linear regression model with optimal scaling. This loss function is minimized over $\mathbf{v}_k$ while keeping β_k fixed by using the ordinary least squares solution:

$$\begin{aligned} \mathbf{v}_k &= ((\beta_k \mathbf{G}_k)^T \beta_k \mathbf{G}_k)^{-1} (\beta_k \mathbf{G}_k)^T \mathbf{u}_k \\ &= (\mathbf{G}_k^T \beta_k \beta_k \mathbf{G}_k)^{-1} \mathbf{G}_k^T \beta_k \mathbf{u}_k \\ &= (\mathbf{G}_k^T \mathbf{G}_k \beta_k \beta_k)^{-1} \mathbf{G}_k^T \beta_k \mathbf{u}_k \\ &= \beta_k^{-2} (\mathbf{G}_k^T \mathbf{G}_k)^{-1} \mathbf{G}_k^T \beta_k \mathbf{u}_k \\ &= \beta_k^{-1} (\mathbf{G}_k^T \mathbf{G}_k)^{-1} \mathbf{G}_k^T \mathbf{u}_k \\ &= \beta_k^{-1} D_k^{-1} \mathbf{G}_k^T \mathbf{u}_k, \end{aligned}$$

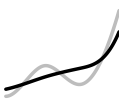
with $D_k = \mathbf{G}_k^T \mathbf{G}_k$, the diagonal matrix that gives the number of objects in each category. Similarly, the ordinary least squares solution can be used to minimize loss function (3.36) over β_k while keeping $\mathbf{v}_k$ fixed. i.e.

$$\begin{aligned} \hat{\beta}_k &= (\hat{\varphi}_k(\mathbf{x}_k)^T \hat{\varphi}_k(\mathbf{x}_k))^{-1} \hat{\varphi}_k(\mathbf{x}_k)^T \mathbf{u}_k \\ &= (\hat{\varphi}_k(x_{k1})^2 + \dots + \hat{\varphi}_k(x_{kp})^2)^{-1} \hat{\varphi}_k(\mathbf{x}_k)^T \mathbf{u}_k \\ &= (\|\hat{\varphi}_k(\mathbf{x}_k)\|^2)^{-1} \hat{\varphi}_k(\mathbf{x}_k)^T \mathbf{u}_k \\ &= 1^{-1} \hat{\varphi}_k(\mathbf{x}_k)^T \mathbf{u}_k \\ &= \mathbf{u}_k^T \hat{\varphi}_k(\mathbf{x}_k). \end{aligned} \quad (3.37)$$

3.6.2 From maximum likelihood to least squares

To transform the maximum partial likelihood approach of the Cox model into an iterated reweighted least squares framework, Simon et al. (2011) used a method similar to the Newton-Raphson method.

The second order Taylor expansion for the log-partial likelihood $\ell(\boldsymbol{\beta})$ centered at current estimate $\tilde{\boldsymbol{\beta}}$ has the form



$$\begin{aligned}
\ell(\beta) &\approx \ell(\tilde{\beta}) + (\beta - \tilde{\beta})^T \dot{\ell}(\tilde{\beta}) + \frac{1}{2}(\beta - \tilde{\beta})^T \ddot{\ell}(\tilde{\beta})(\beta - \tilde{\beta}) \\
&= \ell(\tilde{\eta}) + (\beta - \tilde{\beta})^T \ell'(\tilde{\eta}) \frac{\partial \eta}{\partial \beta} + \frac{1}{2}(\beta - \tilde{\beta})^T \frac{\partial \tilde{\eta}}{\partial \beta} \ell''(\tilde{\eta}) \frac{\partial \tilde{\eta}}{\partial \beta} (\beta - \tilde{\beta}) \\
&= \ell(\tilde{\eta}) + (\beta - \tilde{\beta})^T \ell'(\tilde{\eta}) \mathbf{Z} + \frac{1}{2}(\beta - \tilde{\beta})^T \mathbf{Z} \ell''(\tilde{\eta}) \mathbf{Z} (\beta - \tilde{\beta}) \\
&= \ell(\tilde{\eta}) + (\mathbf{Z}\beta - \mathbf{Z}\tilde{\beta})^T \ell'(\tilde{\eta}) + \frac{1}{2}(\mathbf{Z}\beta - \mathbf{Z}\tilde{\beta})^T \ell''(\tilde{\eta})(\mathbf{Z}\beta - \mathbf{Z}\tilde{\beta}) \\
&= \ell(\tilde{\eta}) + (\mathbf{Z}\beta - \tilde{\eta})^T \ell'(\tilde{\eta}) + \frac{1}{2}(\mathbf{Z}\beta - \tilde{\eta})^T \ell''(\tilde{\eta})(\mathbf{Z}\beta - \tilde{\eta}) \\
&= (\mathbf{Z}\beta - \tilde{\eta})^T \ell'(\tilde{\eta}) + \frac{1}{2}(\mathbf{Z}\beta - \tilde{\eta})^T \ell''(\tilde{\eta})(\mathbf{Z}\beta - \tilde{\eta}) + C(\tilde{\beta}, \tilde{\eta}),
\end{aligned}$$

where $C(\tilde{\beta}, \tilde{\eta})$ does not depend on β . As we will show below, the β that maximizes this log likelihood approximation is the same β that maximizes the function

$$\frac{1}{2}(\zeta(\tilde{\eta}) - \mathbf{Z}\beta)^T \ell''(\tilde{\eta})(\zeta(\tilde{\eta}) - \mathbf{Z}\beta),$$

with $\zeta(\tilde{\eta}) = \tilde{\eta} - \ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta})$. This can be shown as follows.

$$\begin{aligned}
&\frac{1}{2}(\zeta(\tilde{\eta}) - \mathbf{Z}\beta)^T \ell''(\tilde{\eta})(\zeta(\tilde{\eta}) - \mathbf{Z}\beta) \\
&= \frac{1}{2}(\tilde{\eta} - \ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}) - \mathbf{Z}\beta)^T \ell''(\tilde{\eta})(\tilde{\eta} - \ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}) - \mathbf{Z}\beta) \\
&= \frac{1}{2}(\tilde{\eta} - \mathbf{Z}\beta - \ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}))^T \ell''(\tilde{\eta})(\tilde{\eta} - \mathbf{Z}\beta - \ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta})) \\
&= \frac{1}{2} [(\tilde{\eta} - \mathbf{Z}\beta)^T \ell''(\tilde{\eta})(\tilde{\eta} - \mathbf{Z}\beta) - (\tilde{\eta} - \mathbf{Z}\beta)^T \ell''(\tilde{\eta})(\ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta})) \\
&\quad - (\ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}))^T \ell''(\tilde{\eta})(\tilde{\eta} - \mathbf{Z}\beta) + (\ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}))^T \ell''(\tilde{\eta})(\ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}))] \\
&= \frac{1}{2} [(\tilde{\eta} - \mathbf{Z}\beta)^T \ell''(\tilde{\eta})(\tilde{\eta} - \mathbf{Z}\beta) - (\tilde{\eta} - \mathbf{Z}\beta)^T \ell'(\tilde{\eta}) \\
&\quad - \ell'(\tilde{\eta})^T (\tilde{\eta} - \mathbf{Z}\beta) + (\ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}))^T \ell'(\tilde{\eta})] \\
&= \frac{1}{2} [(\tilde{\eta} - \mathbf{Z}\beta)^T \ell''(\tilde{\eta})(\tilde{\eta} - \mathbf{Z}\beta) - 2(\tilde{\eta} - \mathbf{Z}\beta)^T \ell'(\tilde{\eta}) + (\ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}))^T \ell'(\tilde{\eta})] \\
&= \frac{1}{2} [(\mathbf{Z}\beta - \tilde{\eta})^T \ell''(\tilde{\eta})(\mathbf{Z}\beta - \tilde{\eta}) + 2(\mathbf{Z}\beta - \tilde{\eta})^T \ell'(\tilde{\eta}) + (\ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}))^T \ell'(\tilde{\eta})] \\
&= \frac{1}{2} (\mathbf{Z}\beta - \tilde{\eta})^T \ell''(\tilde{\eta})(\mathbf{Z}\beta - \tilde{\eta}) + (\mathbf{Z}\beta - \tilde{\eta})^T \ell'(\tilde{\eta}) + (\ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}))^T \ell'(\tilde{\eta}) \\
&= (\mathbf{Z}\beta - \tilde{\eta})^T \ell'(\tilde{\eta}) + \frac{1}{2} (\mathbf{Z}\beta - \tilde{\eta})^T \ell''(\tilde{\eta})(\mathbf{Z}\beta - \tilde{\eta}) + (\ell''(\tilde{\eta})^{-1} \ell'(\tilde{\eta}))^T \ell'(\tilde{\eta}) \\
&= (\mathbf{Z}\beta - \tilde{\eta})^T \ell'(\tilde{\eta}) + \frac{1}{2} (\mathbf{Z}\beta - \tilde{\eta})^T \ell''(\tilde{\eta})(\mathbf{Z}\beta - \tilde{\eta}) + C(\tilde{\beta}, \tilde{\eta}).
\end{aligned}$$

This is equal to the Taylor approximation of the log likelihood $\ell(\boldsymbol{\beta})$ as shown before. Therefore, maximizing the approximation of the log likelihood over $\boldsymbol{\beta}$ will give the solution as maximizing

$$(\zeta(\tilde{\boldsymbol{\eta}}) - \mathbf{Z}\boldsymbol{\beta})^T \ell''(\tilde{\boldsymbol{\eta}})(\zeta(\tilde{\boldsymbol{\eta}}) - \mathbf{Z}\boldsymbol{\beta}), \quad (3.38)$$

with $\zeta(\tilde{\boldsymbol{\eta}}) = \tilde{\boldsymbol{\eta}} - \ell''(\tilde{\boldsymbol{\eta}})^{-1} \ell'(\tilde{\boldsymbol{\eta}})$, over over $\boldsymbol{\beta}$. Since calculation of $\ell''(\tilde{\boldsymbol{\eta}})$ would require a lot of computations, Simon et al. (2011) proposed to replace it by a diagonal matrix with the diagonal entries of $\ell''(\tilde{\boldsymbol{\eta}})$, denoted as $\omega_i(\tilde{\boldsymbol{\eta}})$. Then, maximizing (3.38) comes down to minimizing

$$L(\tilde{\boldsymbol{\beta}}) = - \sum_{i=1}^n \omega_i(\tilde{\boldsymbol{\eta}}) (\zeta_i(\tilde{\boldsymbol{\eta}}) - \varphi(\mathbf{Z}_{i*})\boldsymbol{\beta})^2. \quad (3.39)$$

In this way the maximum likelihood approach has been recasted into a weighted least squares framework, where the observations are weighted by their second derivatives at the current estimate $\omega_i(\tilde{\boldsymbol{\eta}})$. An iteration procedure is applied to estimate the parameters. In each step, the loss function (3.39) is minimized over $\boldsymbol{\beta}$. The term $\tilde{\boldsymbol{\beta}}$ is then replaced by the estimates $\hat{\boldsymbol{\beta}}$. This procedure is repeated until convergence. This process is called Iteratively Reweighted Least Squares (IRLS) (Green, 1984).

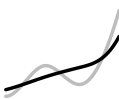
3.6.3 Derivatives log likelihood Cox model

Let $t_1 < t_2 < \dots < t_D$ denote the D distinct and ordered event times. Denote by D_m the set of all individuals who die at time t_m . Let d_m be the sum of the weights for subjects who experience an event at time t_m , i.e. $d_m = \sum_{j \in D_m} w_j$, and let R_{t_m} be the set of individuals r at risk just prior to t_m , with $m = 1, \dots, D$. For random covariates $\mathbf{Z} = (Z_1, \dots, Z_p)$ let $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta}$ be the $(n \times 1)$ -vector with elements $\eta_i = \mathbf{Z}_{i*}\boldsymbol{\beta} = z_{i1}\beta_1 + \dots + z_{ip}\beta_p$, with $i = 1, \dots, n$. The Breslow approximation of the partial likelihood for ties (Breslow, 1972), extended to weighted subjects is

$$L(\boldsymbol{\eta}) = \prod_{m=1}^D \frac{\exp(\sum_{j \in D_m} w_j \eta_j)}{\left(\sum_{r \in R_{t_m}} w_r \exp(\eta_r) \right)^{d_m}}. \quad (3.40)$$

The log likelihood is

$$\begin{aligned} \ell(\boldsymbol{\eta}) &= \sum_{m=1}^D \log \left(\frac{\exp(\sum_{j \in D_m} w_j \eta_j)}{\left(\sum_{r \in R_{t_m}} w_r \exp(\eta_r) \right)^{d_m}} \right) \\ &= \sum_{m=1}^D \left(\sum_{j \in D_m} w_j \eta_j \right) - \sum_{m=1}^D \left(d_m \log \left(\sum_{r \in R_{t_m}} w_r \exp(\eta_r) \right) \right). \end{aligned}$$



The first partial derivative of $\ell(\boldsymbol{\eta})$ with respect to η_i is derived as follows:

$$\begin{aligned}
\ell'(\boldsymbol{\eta})_i &= \frac{\partial \ell(\boldsymbol{\eta})}{\partial \eta_i} \\
&= \left[\sum_{m=1}^D \left(\sum_{j \in D_m} w_j \eta_j \right) \right]' - \left[\sum_{m=1}^D \left(d_m \log \left(\sum_{r \in R_{t_m}} w_r \exp(\eta_r) \right) \right) \right]' \\
&= \left[\sum_{j \in D_m} w_j \eta_j \right]' - \sum_{m=1}^D d_m \frac{1}{\sum_{r \in R_{t_m}} w_r \exp(\eta_r)} \left[\sum_{r \in R_{t_m}} w_r \exp(\eta_r) \right]' \\
&= \delta_i w_i - \sum_{m=1}^D d_m \frac{1}{\sum_{r \in R_{t_m}} w_r \exp(\eta_r)} \mathbf{1}_{\{i \in R_{t_m}\}} w_i \exp(\eta_i) \\
&= \delta_i w_i - \sum_{m=1}^D d_m \frac{1}{\sum_{r \in R_{t_m}} w_r \exp(\eta_r)} \mathbf{1}_{\{t_i \geq t_m\}} w_i \exp(\eta_i) \\
&= \delta_i w_i - \sum_{s \in S_i} \frac{d_s w_i \exp(\eta_i)}{\sum_{r \in R_{t_s}} w_r \exp(\eta_r)},
\end{aligned}$$

where S_i is the set of all individuals s that experience the event before person i 's observed time point, i.e. $\delta_s = 1$ and $t_s \leq t_i$. The second partial derivative of $\ell(\boldsymbol{\eta})$ with respect to η_i is derived as follows:

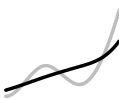
$$\begin{aligned}
\ell''(\boldsymbol{\eta})_{i,i} &= \frac{\partial^2 \ell(\boldsymbol{\eta})}{\partial \eta_i^2} \\
&= [\delta_i w_i]' - \left[\sum_{s \in S_i} \frac{d_s w_i \exp(\eta_i)}{\sum_{r \in R_{t_s}} w_r \exp(\eta_r)} \right]' \\
&= 0 - \sum_{s \in S_i} \frac{[d_s w_i \exp(\eta_i)]' \sum_{r \in R_{t_s}} w_r \exp(\eta_r) - d_s w_i \exp(\eta_i) [\sum_{r \in R_{t_s}} w_r \exp(\eta_r)]'}{(\sum_{r \in R_{t_s}} w_r \exp(\eta_r))^2} \\
&= - \sum_{s \in S_i} \frac{d_s w_i \exp(\eta_i) \sum_{r \in R_{t_s}} w_r \exp(\eta_r) - d_s w_i \exp(\eta_i) w_i \exp(\eta_i)}{(\sum_{r \in R_{t_s}} w_r \exp(\eta_r))^2} \\
&= - \sum_{s \in S_i} d_s \frac{w_i \exp(\eta_i) \sum_{r \in R_{t_s}} w_r \exp(\eta_r) - (w_i \exp(\eta_i))^2}{(\sum_{r \in R_{t_s}} w_r \exp(\eta_r))^2}.
\end{aligned}$$

OPTIMAL SCALING TRANSFORMATIONS TO MODEL NONLINEAR RELATIONS IN GLMs WITH ORDERED AND UNORDERED PREDICTORS

4

In Generalized Linear Models (GLMs) it is assumed that there is a linear effect of the predictor variables on the outcome. However, this assumption is often too strict, because in many applications predictors have a nonlinear relation with the outcome. Optimal Scaling (OS) transformations combined with GLMs can deal with this type of relations. Transformations of the predictors have been integrated in GLMs before, e.g. in Generalized Additive Models. However, the OS methodology several benefits. For example, the levels of categorical predictors are quantified directly, such that they can be included in the model without defining dummy variables. This approach enhances the interpretation and visualization of the effect of different levels on the outcome. Furthermore, monotonicity restrictions can be applied to the OS transformations such that the original ordering of the category values is preserved. This improves the interpretation of the effect and may prevent overfitting. The scaling level can be chosen for each individual predictor such that models can include mixed scaling levels. In this way, a suitable transformation can be found for each predictor in the model. The implementation of OS in logistic regression is demonstrated using three datasets that contain a binary outcome variable and a set of categorical and/or continuous predictor variables.

This chapter is submitted as: Willems, S. J. W., Van der Kooij, A. J., Fiocco, M., and Meulman, J. J. *Optimal Scaling transformations to model nonlinear relations in GLMs with ordered and unordered predictors.*



4.1 Introduction

Linear models are often used to model relations between a numeric outcome variable and a set of predictor variables. The ordinary least squares regression model (OLS) assumes normally distributed errors and linearity in the predictors. Due to these assumptions, the application to real data is sometimes limited. For example, consider a medical application in which the relation between the binary outcome of getting a particular disease and the predictor variable age is modeled. First of all, the binary outcome cannot be modeled with the standard linear regression model due to the assumption of normally distributed errors. Furthermore, due to their weaker immune systems, it may be expected that both young children and elderly people are more susceptible to the disease than people of intermediate ages. In such situations, the relation between the probability of getting the disease will have an inverted-u-shape and thus the linearity assumption is too strict. Hence, for these types of situations, the ordinary linear model is not appropriate.

To increase the applicability of the linear model, several extensions have been developed.

One extension is to allow for a nonlinear relation between the linear combination of the predictor variables and the outcome via a link function. This type of models are known as Generalized Linear Models (GLMs, McCullagh and Nelder (1989)). GLMs do not assume normally distributed errors and are therefore applicable if errors are distributed differently. A frequently used GLM for binary outcomes is the logistic regression model, which uses the logit link function to transform the linear predictor into the unit interval to model probabilities.

A second extension is by transforming the variables. This is done in, for example, additive models (Friedman and Stuetzle (1981); Hastie and Tibshirani (1990); Winsberg and Ramsay (1980)) and Optimal Scaling regression (OS-regression) (Gifi, 1990; Van der Kooij and Meulman, 1999; Young et al., 1976). The predictor variables are transformed using either a parametric or a nonparametric function.

In this paper, we will integrate two extensions of ordinary linear models by combining GLMs with optimal scaling techniques. As a result, a nonlinear link function (as in a GLM) is used to model the relation between the response variable and a linear combination of *transformed* predictor variables (as in the OS approach). Hence, the important difference between a regular GLM and a GLM with OS lies in the transformation of the predictor variables.

Initially, OS was developed to transform nominal or ordinal categorical variables into quantitative data by finding optimal numeric values for the category values. This process was referred to as *quantifying qualitative data* by Young

(1981) and the resulting transformations are called *quantifications*, denoted as $\varphi_k(\mathbf{x}_k)$ for variable k .

The quantifications can also be written in matrix form as $\varphi_k(\mathbf{x}_k) = \mathbf{G}_k \mathbf{v}_k$. Here, $\mathbf{v}_k$ is a vector with the quantifications for each category level of variable k , and $\mathbf{G}_k$ is an indicator matrix that represents the observed category values in $\mathbf{x}_k$. Namely, the number of columns in this matrix is equal to the number of categories and each row contains only zero's and a single one where the one is placed in the column that corresponds with i 's observed category level.

Although the OS methodology was originally developed for categorical data, it can also be applied to non linearly transform numeric data. In this case, all unique observations of the numeric variable are interpreted as an individual category level and they are modeled in the same way as for categorical predictors. Hence, if all objects have unique values, $\mathbf{G}_k$ is a permuted identity matrix.

In OS-regression, the response y_i of observation i is modeled as a linear combination of the quantifications of the p observed predictors. Hence, the model is as follows

$$y_i = \sum_{k=1}^p \beta_k \varphi_k(x_{ik}) + \epsilon_i,$$

where ϵ_i is the error term. After explaining the OS algorithm for linear models, we will show how it can be integrated in the Newton-Raphson method to fit a GLM model with OS transformations.

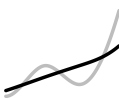
The type of transformation (also called *scaling level* in the categorical data analysis context) is chosen for each individual variable and may thus differ among predictors. The combination of coefficients and transformations calculated by the algorithm optimally describe the relation between the response and the predictors under the restrictions set by the chosen scaling levels.

Several types of scaling levels can be chosen, depending on the expected, or imposed, relation between the predictor and the outcome.

Usually a step-function is chosen for categorical predictors with few category levels which can either be monotone or nonmonotone, depending on whether the ordering of the category levels should be preserved. Kruskal (1964) described one of the first algorithms to find monotonic step transformations in multidimensional scaling and a similar technique is applied in OS.

If the predictor has many category levels (e.g. for a numeric variable), some smoothing may be appropriate to avoid overfitting and to improve interpretation. In these cases, either a monotone or nonmonotone spline function can be fit, again depending on whether the ordering of the categories should be preserved. I-splines (as described by Ramsay (1988)) are used to fit the (non)monotonic spline function.

In case a linear relation may actually be suitable for a predictor, a linear



(numeric) scaling level may also be chosen. If a numeric scaling level is chosen for all predictors, the GLM-OS will give the same output as a ordinary GLM.

In this paper, we will describe how OS can be integrated in the Newton-Raphson algorithm to find optimal quantifications for the predictor variables in a GLM. This combination of methods is defined as the Generalized Linear Model with Optimal Scaling (GLM-OS) or the generalized version of Optimal Scaling regression (GOS-regression). Since the OS method nonlinearly transforms the data, the term *linear predictor* that is used for the linear combination of transformed predictors in the GLM literature may be confusing. Therefore, it is referred to in this paper as the *weighted sum* or *linear combination* of transformed predictors.

Although applicable to more GLMs, we will focus on logistic regression with OS transformations and apply this model to three datasets. Each of these datasets has different types of predictor variables, which allows us to illustrate the benefits of OS with respect to visualization, interpretation, and predictability.

4.2 Optimal scaling in linear regression

In this section, we will explain how optimal scaling transformations are integrated in linear regression. To keep this explanation concise, we only show the basics and leave out the details and extensions. For more details about the OS-regression algorithm, including some adjustments to optimize calculation time, we refer to Van der Kooij (2007) and Meulman et al. (2019).

4.2.1 Model and notation

Let $\mathbf{X}$ be the data matrix of dimension $n \times p$ where n and p are the number of objects and predictors respectively. The n observed values of the response variable are collected in the vector $\mathbf{y}$.

In ordinary least squares regression (OLS), the outcome is modeled as a linear combination of the predictors, i.e. $y_i = \sum_{k=1}^p \beta_k x_{ik} + \epsilon_i$, where ϵ_i is the error term. In the optimal scaling setting, the original observed values $\mathbf{x}_k$ of each predictor variable k , for $k = 1, \dots, p$, are transformed and replaced by their quantifications that is denoted as $\varphi_k(\mathbf{x}_k)$. The outcome $\mathbf{y}$ is assumed to be centered and therefore no intercept is required. Hence, the OS-regression model is $\mathbf{y} = \sum_{k=1}^p \beta_k \varphi_k(\mathbf{x}_k) + \epsilon$.

The quantifications for all n observations can be written in matrix form. Let C_k be the number of unique observed values for predictor k , and denote by $\mathbf{G}_k$ the indicator matrix of dimensions $n \times C_k$. Each i th row of $\mathbf{G}_k$ consists of $C_k - 1$ zero's and a single one, placed in the column which corresponds to the value x_{ik} .

Furthermore, let $\mathbf{v}_k$ be the $C_k \times 1$ vector that contains the C_k quantifications of predictor k . Then, $\mathbf{G}_k \mathbf{v}_k$ is the $n \times 1$ vector of the transformed value for each object, i.e. $\varphi_k(\mathbf{x}_k) = \mathbf{G}_k \mathbf{v}_k$. Using this notation, the linear regression model with optimal scaling quantifications in matrix form can be written as

$$\mathbf{y} = \sum_{k=1}^p \beta_k \varphi_k(\mathbf{x}_k) + \epsilon = \sum_{k=1}^p \beta_k \mathbf{G}_k \mathbf{v}_k + \epsilon. \quad (4.1)$$

The matrices $\mathbf{G}_1, \dots, \mathbf{G}_p$ are derived from the data, and coefficients $\beta_1, \dots, \beta_p$ and quantifications $\mathbf{v}_1, \dots, \mathbf{v}_p$ need to be estimated.

4.2.2 Model estimation

The loss function corresponding to the OS-regression model in (4.1) is written as

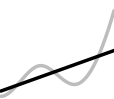
$$L(\mathbf{v}_1, \dots, \mathbf{v}_p; \beta_1, \dots, \beta_p) = \left\| \mathbf{y} - \sum_{k=1}^p \beta_k \mathbf{G}_k \mathbf{v}_k \right\|^2. \quad (4.2)$$

To fit the model, the loss function should be minimized over both the model coefficients $\beta_1, \dots, \beta_p$, and the quantifications $\mathbf{v}_1, \dots, \mathbf{v}_p$ simultaneously, where the quantifications are restricted according to their scaling level, as described above these are nominal and ordinal step or spline functions. As an infinite number of combinations of model coefficients and quantifications will optimize this function, the latter are standardized to ensure a unique solution.

Since no closed-form solution is available to minimize loss function (4.2) over all parameters simultaneously, the quantifications and coefficients are optimized for one variable at the time, and this process is iterated until convergence. This type of algorithm is referred to as *alternating least squares* in the psychometric literature (Gifi, 1990; Young et al., 1976), since the least squares solution is calculated by alternating the estimation of optimal quantifications and model coefficients for one variable at the time. In the statistical literature it is called *backfitting* and has been extensively used to fit Additive Models and GAMs (Friedman and Stuetzle, 1981; Hastie and Tibshirani, 1990). A variety of other terms is present in the literature, like *component-wise update* and *block relaxation*, but it is currently usually referred to as *coordinate descent*.

In the initialization step, standardized values of the observed variables are used as starting values for the quantifications $\mathbf{v}_1, \dots, \mathbf{v}_p$, and the corresponding ordinary least squares solution based on these standardized quantifications are used as starting values for $\beta_1, \dots, \beta_p$. If a numeric scaling level is chosen for variable k , standardizing $\mathbf{x}_k$ already gives a solution that satisfies the restrictions. Hence, quantifications of a numeric scaling level do not require any adjustments except for standardization.

After initialization, the parameters are updated for a single variable at the time. At each iteration, all regression coefficients and variables are assumed to



be fixed, except for the variable k that is currently (conditionally) optimized. All fixed terms are merged into a single vector denoted by $\mathbf{u}_k$ and variable k is then separated from this fixed part, i.e.

$$L(\mathbf{v}_k, \beta_k) = \left\| \mathbf{y} - \sum_{l \neq k} \beta_l \mathbf{G}_l \mathbf{v}_l - \beta_k \mathbf{G}_k \mathbf{v}_k \right\|^2 = \left\| \mathbf{u}_k - \beta_k \mathbf{G}_k \mathbf{v}_k \right\|^2. \quad (4.3)$$

If variable k 's scaling level is not numeric, quantifications $\mathbf{v}_k$ need to be updated. While updating $\mathbf{v}_k$, it is assumed that β_k is fixed, which enables us to calculate the unrestricted solution for $\mathbf{v}_k$ as the ordinary least squared solution for (4.3) with respect to $\mathbf{v}_k$. Hence, if $\tilde{\mathbf{v}}_k$ is the current estimate of $\mathbf{v}_k$, then it is updated as

$$\begin{aligned} \tilde{\mathbf{v}}_k^+ &= \left\{ (\tilde{\beta}_k \mathbf{G}_k)^T \tilde{\beta}_k \mathbf{G}_k \right\}^{-1} (\tilde{\beta}_k \mathbf{G}_k)^T \mathbf{u}_k \\ &= \left\{ \tilde{\beta}_k^2 \mathbf{G}_k^T \mathbf{G}_k \right\}^{-1} \mathbf{G}_k^T \tilde{\beta}_k \mathbf{u}_k \\ &= \tilde{\beta}_k^{-1} \mathbf{D}_k^{-1} \mathbf{G}_k^T \mathbf{u}_k, \end{aligned} \quad (4.4)$$

where $\tilde{\beta}_k$ is the current estimate of β_k and $\mathbf{D}_k = \mathbf{G}_k^T \mathbf{G}_k$. Actually, since $\tilde{\mathbf{v}}_k$ will be standardized later, $\tilde{\beta}_k^{-1}$ in (4.4) can be replaced by $\text{sign}(\tilde{\beta}_k)$.

This unrestricted solution is actually the solution to the optimal scaling problem for a nominal variable. For the other scaling levels, restrictions have to be applied to $\tilde{\mathbf{v}}_k^+$. For the ordinal scaling level, weighted monotonic regression (Kruskal (1964)) is applied, resulting in a monotonic step function. For the nonmonotone and monotone spline restrictions (with a specified number of knots and degree of the polynomial functions, Ramsay (1988)) are fitted to the unrestricted solution. After the appropriate restrictions have been applied, the result is standardized to ensure a unique solution. This restricted and standardized solution is then the current estimate $\tilde{\mathbf{v}}_k$ of $\mathbf{v}_k$.

Once the quantifications of the k th variable have been updated, model parameter β_k is estimated by again using the ordinary least squares solution for loss function (4.3) in which $\mathbf{G}_k \mathbf{v}_k$ is now fixed. Hence, the updated value for β_k is calculated as

$$\begin{aligned} \tilde{\beta}_k^+ &= \left\{ (\mathbf{G}_k \tilde{\mathbf{v}}_k)^T \mathbf{G}_k \tilde{\mathbf{v}}_k \right\}^{-1} (\mathbf{G}_k \tilde{\mathbf{v}}_k)^T \mathbf{u}_k \\ &= \left\{ \tilde{\mathbf{v}}_k^T \mathbf{D}_k \tilde{\mathbf{v}}_k \right\}^{-1} \tilde{\mathbf{v}}_k^T \mathbf{G}_k^T \mathbf{u}_k. \end{aligned} \quad (4.5)$$

The algorithm continues updating the quantifications and model parameters for the other variables. This process continues until the loss measured by (4.2) does not change anymore.

The final estimates of the model coefficients and quantifications (denoted as $\hat{\beta}_1, \dots, \hat{\beta}_p$ and $\hat{\mathbf{v}}_1, \dots, \hat{\mathbf{v}}_p$) are the updates from the last iteration. Usually the

final estimates of the quantifications ($\hat{\mathbf{v}}_k$) are plotted against the original values of variable k to visualize the transformations.

Note that $\mathbf{G}_k$ and $\mathbf{D}_k$ are sparse matrices which make the above calculations very inefficient. Therefore, in the implementation of the algorithm, methods are applied to perform these calculations without using these matrices.

OS-regression algorithm:

Initialization: Create $\mathbf{G}_1, \dots, \mathbf{G}_p$ based on the data, and initialize the model parameters $\tilde{\beta}_1, \dots, \tilde{\beta}_p$ and $\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_p$.

Cycle: For $k = 1, \dots, p$, do:

Step 1: Calculate $\mathbf{u}_k = \mathbf{y} - \sum_{l \neq k} \beta_l \mathbf{G}_l \mathbf{v}_l$.

Step 2: If the scaling level of variable k is nonnumeric, calculate the unrestricted estimates of the quantifications of k as

$$\tilde{\mathbf{v}}_k^+ = \tilde{\beta}_k^{-1} \mathbf{D}_k^{-1} \mathbf{G}_k^T \mathbf{u}_k.$$

Apply appropriate scaling restrictions to $\tilde{\mathbf{v}}_k^+$ and standardize the result.

Step 3: Update the estimate for model coefficient β_k as

$$\tilde{\beta}_k^+ = \{\tilde{\mathbf{v}}_k^T \mathbf{D}_k \tilde{\mathbf{v}}_k\}^{-1} \tilde{\mathbf{v}}_k^T \mathbf{G}_k^T \mathbf{u}_k.$$

Convergence: Repeat the cycle until convergence criteria are met.

4.3 Optimal scaling in generalized linear models

In this section we will explain how the OS procedure can be integrated in the Newton-Raphson algorithm used to fit GLMs. After describing the Newton-Raphson algorithm as it is used to fit regular GLMs, we will show how it can be modified to include OS transformations. Then we will show the specific example of how optimal scaling transformations can be calculated for the logistic regression model. This model will also be used for the data illustrations in the next section.

4.3.1 GLM-OS model and notation

For GLM-OS we use notation that is similar to the notation used for OS-regression. Hence, let $\mathbf{X}$ and $\mathbf{y}$ again be the data matrix and the vector with the outcome. In a GLM, the outcome is not centered and thus these models include an intercept. In the GLM-OS setting, we therefore assume that the intercept is

represented by the regression coefficient β_0 multiplied by a vector of ones, which is denoted as $\mathbf{x}_0$ and included in the data matrix $\mathbf{X}$.

A GLM consists of three components, namely

- 1) a random component that specifies the distribution of the response variable given the predictors;
- 2) a linear combination of the predictor variables, denoted as $\boldsymbol{\eta} = \beta_0\mathbf{x}_0 + \beta_1\mathbf{x}_1 + \dots + \beta_p\mathbf{x}_p$;
- 3) an invertible link function g which models the relation between the linear combination of object i and i 's response y_i , i.e. $g(\mu_i) = \eta_i = \beta_0x_{i0} + \dots + \beta_px_{ip}$ where $\mu_i = E(Y_i)$.

To extend GLMs to include optimal scaling transformation the linear combination of predictors is replaced by a linear combination of the quantifications, so

$$\boldsymbol{\eta} = \sum_{k=0}^p \beta_k \varphi_k(\mathbf{x}_k) = \sum_{k=0}^p \beta_k \mathbf{G}_k \mathbf{v}_k. \quad (4.6)$$

To fit the GLM-OS, coefficients $\beta_0, \dots, \beta_p$ and quantifications $\mathbf{v}_1, \dots, \mathbf{v}_p$ need to be estimated. Note that, to represent the intercept, $\varphi_0(\mathbf{x}_0) = \mathbf{1}_n$, and consequently $\mathbf{G}_0 = \mathbf{1}_n$ and $\mathbf{v}_0 = \{1\}$, are fixed, and hence these terms do not have to be estimated in each iteration.

4.3.2 Model estimation

The maximum likelihood approach is used to estimate GLMs. The exact form of the likelihood function depends on the random component of the GLM and the link function. The log-likelihood is a function of the linear combination of predictors and is denoted as $l(\boldsymbol{\eta})$. In a GLM $\boldsymbol{\eta}$ only depends on parameters $\beta_0, \dots, \beta_p$, while in GLM-OS it depends on both $\beta_0, \dots, \beta_p$ and $\mathbf{v}_1, \dots, \mathbf{v}_p$.

There is no closed-form solution to maximize the (log-)likelihood functions, hence a numerical method is required to find the maximum likelihood estimator (MLE). For GLMs, usually the Newton-Raphson method is used.

Newton-Raphson method for GLMs

The GLM fitting algorithm aims to find the roots of the gradient by using the Newton-Raphson algorithm. This method iteratively improves the initial starting values via the first-order Taylor approximation of the gradient $\nabla_l(\boldsymbol{\beta})$ of the log-likelihood around the current guess $\tilde{\boldsymbol{\beta}}$. Hence, the solutions are found as

follows

$$\begin{aligned}\mathbf{0} &= \nabla_l(\boldsymbol{\beta}) \approx \nabla_l(\tilde{\boldsymbol{\beta}}) + \mathbf{H}_l(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) \\ -\mathbf{H}_l(\tilde{\boldsymbol{\beta}}) \boldsymbol{\beta} &\approx \nabla_l(\tilde{\boldsymbol{\beta}}) - \mathbf{H}_l(\tilde{\boldsymbol{\beta}}) \tilde{\boldsymbol{\beta}} \\ \boldsymbol{\beta} &\approx \tilde{\boldsymbol{\beta}} - \mathbf{H}_l^{-1}(\tilde{\boldsymbol{\beta}}) \nabla_l(\tilde{\boldsymbol{\beta}}),\end{aligned}$$

where

- $\tilde{\boldsymbol{\beta}}$ is the current estimate of $\boldsymbol{\beta}$;
- $\mathbf{H}_l(\tilde{\boldsymbol{\beta}})$ is the Hessian matrix containing all the second-order partial derivatives of $l(\boldsymbol{\eta})$ w.r.t. $\boldsymbol{\beta}$ evaluated at $\tilde{\boldsymbol{\beta}}$;
- $\nabla_l(\tilde{\boldsymbol{\beta}})$ is the gradient vector that contains the first-order partial derivatives of $l(\boldsymbol{\eta})$ w.r.t. $\boldsymbol{\beta}$ evaluated at $\tilde{\boldsymbol{\beta}}$.

Then, the current estimate $\tilde{\boldsymbol{\beta}}$ is updated in each iteration as

$$\tilde{\boldsymbol{\beta}}^+ = \tilde{\boldsymbol{\beta}} - \mathbf{H}_l^{-1}(\tilde{\boldsymbol{\beta}}) \nabla_l(\tilde{\boldsymbol{\beta}}). \quad (4.7)$$

Each update $\tilde{\boldsymbol{\beta}}^+$ should be a better approximation of the root than the previous estimate $\tilde{\boldsymbol{\beta}}$ and the algorithm repeatedly updates these estimates until the convergence criteria are met.

In some applications an approximation of $\mathbf{H}_l^{-1}(\tilde{\boldsymbol{\beta}})$ is used to simplify the calculations. For example, in Fisher's Scoring method, the Hessian is replaced by its expectation. In some cases, the Hessian and its expectation are identical, in which case Newton-Raphson and Fisher's scoring method are equivalent. If $\mathbf{H}_l(\tilde{\boldsymbol{\eta}})$ is not a diagonal matrix, it can be approximated by a diagonal matrix to reduce calculation time. For example, Simon et al. (2011) and Willems et al. (2017) fitted Cox' Proportional Hazards model in the context of regularization and OS transformations respectively, and approximated the full Hessian matrix by its diagonal.

If it is easier to do calculations with the negative log-likelihood, the algorithm is modified such that it finds the minimum of $-l(\boldsymbol{\eta})$. In this case the algorithm does essentially not change except that it now uses the the gradient and Hessian of the negative log-likelihood $-l(\boldsymbol{\eta})$ to repeatedly update $\tilde{\boldsymbol{\beta}}$ to find $\hat{\boldsymbol{\beta}}$.

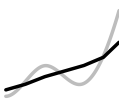
Modification of the Newton-Raphson method to fit GLM-OS

To estimate the GLM-OS model the coefficients $\beta_1, \dots, \beta_p$ and quantifications $\mathbf{v}_1, \dots, \mathbf{v}_p$ that maximize the log-likelihood function need to be computed. Hence, the following equations need to be solved,

$$\nabla_l(\mathbf{v}_1) = \dots = \nabla_l(\mathbf{v}_p) = \mathbf{0},$$

and

$$\nabla_l(\beta_0) = \dots = \nabla_l(\beta_p) = 0.$$



Since there is no closed-form solution to derive the parameters simultaneously, they will be calculated iteratively for one variable k at the time, as was done in the OS-regression algorithm. After initialization, the algorithm iterates over all $k = 0, \dots, p$ predictors and updates first the quantifications $\mathbf{v}_k$ (unless $k = 0$) and then model coefficient β_k .

As when updating β in ordinary GLMs, we set the first-order Taylor approximation of $\nabla_l(\mathbf{v}_k)$ around the current estimate $\tilde{\mathbf{v}}_k$ to zero and derive the update for $\mathbf{v}_k$ from that equation, i.e. via

$$\begin{aligned} \mathbf{0} &= \nabla_l(\mathbf{v}_k) \approx \nabla_l(\tilde{\mathbf{v}}_k) + \mathbf{H}_l(\tilde{\mathbf{v}}_k) (\mathbf{v}_k - \tilde{\mathbf{v}}_k) \\ \mathbf{v}_k &\approx \tilde{\mathbf{v}}_k - \mathbf{H}_l^{-1}(\tilde{\mathbf{v}}_k) \nabla_l(\tilde{\mathbf{v}}_k), \end{aligned}$$

where

- $\tilde{\mathbf{v}}_k$ is the current estimate of $\mathbf{v}_k$;
- $\mathbf{H}_l(\tilde{\mathbf{v}}_k)$ is the Hessian matrix containing all the second-order partial derivatives of $l(\boldsymbol{\eta})$ w.r.t. $\mathbf{v}_k$ evaluated at $\tilde{\mathbf{v}}_k$ (or some approximation thereof); and
- $\nabla_l(\tilde{\mathbf{v}}_k)$ is the gradient vector containing the first-order partial derivatives of $l(\boldsymbol{\eta})$ w.r.t. $\mathbf{v}_k$ evaluated at $\tilde{\mathbf{v}}_k$.

Since $\boldsymbol{\eta}$ is the weighted sum of transformed predictors, i.e. $\boldsymbol{\eta} = \sum_{k=0}^p \beta_k \mathbf{G}_k \mathbf{v}_k$, the gradient vector of $\boldsymbol{\eta}$ w.r.t. $\mathbf{v}_k$ is $\beta_k \mathbf{G}_k$. Hence, from the chain rule

$$\nabla_l(\tilde{\mathbf{v}}_k) = (\tilde{\beta}_k \mathbf{G}_k)^T \nabla_l(\tilde{\boldsymbol{\eta}})$$

and

$$\mathbf{H}_l(\tilde{\mathbf{v}}_k) = (\tilde{\beta}_k \mathbf{G}_k)^T \mathbf{H}_l(\tilde{\boldsymbol{\eta}}) \tilde{\beta}_k \mathbf{G}_k,$$

with $\nabla_l(\tilde{\boldsymbol{\eta}})$ the gradient vector and $\mathbf{H}_l(\tilde{\boldsymbol{\eta}})$ the Hessian matrix of $l(\boldsymbol{\eta})$ w.r.t. $\boldsymbol{\eta}$ evaluated at current estimate $\tilde{\boldsymbol{\eta}}$. Hence, the quantifications for all predictors with a nonnumeric scaling level are updated as

$$\begin{aligned} \tilde{\mathbf{v}}_k^+ &= \tilde{\mathbf{v}}_k - \mathbf{H}_l^{-1}(\tilde{\mathbf{v}}_k) \nabla_l(\tilde{\mathbf{v}}_k) \\ &= \tilde{\mathbf{v}}_k - \left\{ (\tilde{\beta}_k \mathbf{G}_k)^T \mathbf{H}_l(\tilde{\boldsymbol{\eta}}) \tilde{\beta}_k \mathbf{G}_k \right\}^{-1} (\tilde{\beta}_k \mathbf{G}_k)^T \nabla_l(\tilde{\boldsymbol{\eta}}). \end{aligned} \quad (4.8)$$

These updates are the unrestricted estimates of the quantifications and are the optimal solution for a nominal scaling level. For the other scaling levels, restrictions have to be applied to $\tilde{\mathbf{v}}_k^+$ by fitting a nonmonotone or monotone step or spline function, as is done in OS-regression. Then the quantifications are standardized to ensure a unique solution.

After updating the quantifications $\mathbf{v}_k$ for predictor k , β_k needs to be updated accordingly. Again, updates can be derived via the first-order Taylor approximation of $\nabla_l(\beta_k)$, which results in

$$\tilde{\beta}_k^+ = \tilde{\beta}_k - \mathbf{H}_l^{-1}(\tilde{\beta}_k) \nabla_l(\tilde{\beta}_k), \quad (4.9)$$

where

- $\tilde{\beta}_k$ is the current estimate of β_k ;
- $\mathbf{H}_l(\tilde{\beta}_k)$ is the Hessian matrix containing all the second-order partial derivatives of $l(\boldsymbol{\eta})$ w.r.t. β_k evaluated at $\tilde{\beta}_k$ (or some approximation thereof); and
- $\nabla_l(\tilde{\beta}_k)$ is the gradient vector containing the first-order partial derivatives of $l(\boldsymbol{\eta})$ w.r.t. β_k evaluated at $\tilde{\beta}_k$.

Using the chain rule,

$$\begin{aligned}\tilde{\beta}_k^+ &= \tilde{\beta}_k - \mathbf{H}_l^{-1}(\tilde{\beta}_k) \nabla_l(\tilde{\beta}_k) \\ &= \tilde{\beta}_k - \{(\mathbf{G}_k \tilde{\mathbf{v}}_k)^T \mathbf{H}_l(\tilde{\boldsymbol{\eta}}) \mathbf{G}_k \tilde{\mathbf{v}}_k\}^{-1} (\mathbf{G}_k \tilde{\mathbf{v}}_k)^T \nabla_l(\tilde{\boldsymbol{\eta}}),\end{aligned}$$

where $\nabla_l(\tilde{\boldsymbol{\eta}})$ and $\mathbf{H}_l(\tilde{\boldsymbol{\eta}})$ are recalculated in between updating $\tilde{\mathbf{v}}_k$ and $\tilde{\beta}_k$.

The modified version of the Newton-Raphson method for GLM-OS can be summarized as follows.

GLM-OS algorithm:

Initialization: Set $\mathbf{G}_0 = \mathbf{1}_n$ and $\mathbf{v}_0 = \{1\}$, create $\mathbf{G}_1, \dots, \mathbf{G}_p$ based on the data, and initialize the model parameters $\tilde{\beta}_0, \dots, \tilde{\beta}_p$ and $\tilde{\mathbf{v}}_1, \dots, \tilde{\mathbf{v}}_p$.

Cycle: For $k = 0, \dots, p$, do:

Step 1: Calculate the Hessian matrix $\mathbf{H}_l(\tilde{\boldsymbol{\eta}})$ and the gradient vector $\nabla_l(\tilde{\boldsymbol{\eta}})$.

Step 2: If the scaling level of variable k is nonnumeric, calculate the unrestricted estimates of the quantifications of k as

$$\tilde{\mathbf{v}}_k^+ = \tilde{\mathbf{v}}_k - \{(\tilde{\beta}_k \mathbf{G}_k)^T \mathbf{H}_l(\tilde{\boldsymbol{\eta}}) \tilde{\beta}_k \mathbf{G}_k\}^{-1} (\tilde{\beta}_k \mathbf{G}_k)^T \nabla_l(\tilde{\boldsymbol{\eta}}).$$

Apply appropriate scaling restrictions to $\tilde{\mathbf{v}}_k^+$ and standardize the result.

Step 3: Update the Hessian matrix $\mathbf{H}_l(\tilde{\boldsymbol{\eta}})$ and the gradient vector $\nabla_l(\tilde{\boldsymbol{\eta}})$ using the current estimate $\tilde{\mathbf{v}}_k$.

Step 4: Update the estimate for model coefficient β_k as

$$\tilde{\beta}_k^+ = \tilde{\beta}_k - \{(\mathbf{G}_k \tilde{\mathbf{v}}_k)^T \mathbf{H}_l(\boldsymbol{\eta}) \mathbf{G}_k \tilde{\mathbf{v}}_k\}^{-1} (\mathbf{G}_k \tilde{\mathbf{v}}_k)^T \nabla_l(\tilde{\boldsymbol{\eta}}).$$

Convergence: Repeat the cycle until convergence criteria are met.

4.3.3 The relation between the Newton-Raphson method for GLM(-OS)s, IRLS, and OS-regression

The Newton-Raphson method for GLMs is often referred to as Iterative Reweighed Least Squares (IRLS), because the algorithm iteratively solves reweighted least

squares problems. This will be explained in this section. This relation between GLM estimation and least squares is important for GLM-OS, because it accommodates the use of monotone regression and I-splines when finding optimal quantifications.

As was shown in subsubsection 4.3.2, $\tilde{\beta}$ in an ordinary GLM is updated in each iteration as

$$\tilde{\beta}^+ = \tilde{\beta} - \mathbf{H}_l^{-1}(\tilde{\beta}) \nabla_l(\tilde{\beta}).$$

Given that the linear combination in ordinary GLMs is $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, the matrix containing all its partial derivatives w.r.t. $\beta_1, \dots, \beta_k$ is $\mathbf{X}$. Hence, according to the chain rule

$$\mathbf{H}_l(\tilde{\beta}) = \mathbf{X}^T \mathbf{H}_l(\tilde{\boldsymbol{\eta}}) \mathbf{X}$$

and

$$\nabla_l(\tilde{\beta}) = \mathbf{X}^T \nabla_l(\tilde{\boldsymbol{\eta}}).$$

Hence, the updates for the model parameters $\tilde{\beta}$ can be rewritten as

$$\begin{aligned} \tilde{\beta}^+ &= \tilde{\beta} - \mathbf{H}_l^{-1}(\tilde{\beta}) \nabla_l(\tilde{\beta}) \\ &= \tilde{\beta} - \{\mathbf{X}^T \mathbf{H}_l(\tilde{\boldsymbol{\eta}}) \mathbf{X}\}^{-1} \mathbf{X}^T \nabla_l(\tilde{\boldsymbol{\eta}}) \\ &= \{\mathbf{X}^T \mathbf{H}_l(\tilde{\boldsymbol{\eta}}) \mathbf{X}\}^{-1} \mathbf{X}^T \mathbf{H}_l(\tilde{\boldsymbol{\eta}}) \{\mathbf{X}\tilde{\beta} - \mathbf{H}_l^{-1}(\tilde{\boldsymbol{\eta}}) \nabla_l(\tilde{\boldsymbol{\eta}})\} \\ &= \{\mathbf{X}^T \mathbf{H}_l(\tilde{\boldsymbol{\eta}}) \mathbf{X}\}^{-1} \mathbf{X}^T \mathbf{H}_l(\tilde{\boldsymbol{\eta}}) \mathbf{z} \end{aligned}$$

where $\mathbf{z} = \mathbf{X}\tilde{\beta} - \mathbf{H}_l^{-1}(\tilde{\boldsymbol{\eta}}) \nabla_l(\tilde{\boldsymbol{\eta}})$. These updates are exactly the solution to the weighted least squares problem

$$\operatorname{argmin}_{\boldsymbol{\beta}} \|\mathbf{z} - \mathbf{X}\boldsymbol{\beta}\|_{\mathbf{H}_l(\tilde{\boldsymbol{\eta}})}^2,$$

with $\mathbf{H}_l(\tilde{\boldsymbol{\eta}})$ the (diagonal) matrix with weights for each observation. Hence, the Newton-Raphson algorithm iteratively optimizes a weighted least squares problem in which the weights are updated in each iteration. For this reason, it is often called the Iterative Reweighted Least Squares algorithm.

The same reasoning can be used to show that the GLM-OS algorithm iteratively solves the weighted least squares problems

$$\operatorname{argmin}_{\mathbf{v}_k} \left\| \mathbf{z}_k - \tilde{\beta}_k \mathbf{G}_k \mathbf{v}_k \right\|_{\mathbf{H}_l(\tilde{\boldsymbol{\eta}})}^2 \quad (4.10)$$

and

$$\operatorname{argmin}_{\beta_k} \left\| \mathbf{z}_k - \beta_k \mathbf{G}_k \tilde{\mathbf{v}}_k \right\|_{\mathbf{H}_l(\tilde{\boldsymbol{\eta}})}^2, \quad (4.11)$$

with $\mathbf{z}_k = \tilde{\beta}_k \mathbf{G}_k \tilde{\mathbf{v}}_k - \mathbf{H}_l^{-1}(\tilde{\boldsymbol{\eta}}) \nabla_l(\tilde{\boldsymbol{\eta}})$.

Since the GLM-OS alternates between updating the model coefficients and the quantifications, it could be referred to as the Iterative Reweighted Alternating Least Squares (IRALS) algorithm. Note that the weights in $\mathbf{H}_l(\tilde{\boldsymbol{\eta}})$ are recalculated between calculating the updates of $\tilde{\mathbf{v}}_k^+$ and $\tilde{\beta}_k^+$, and that the objects should be weighted accordingly when fitting the step or spline functions.

Although loss functions (4.10) and (4.11) look very similar to loss function (4.3) of the OS-regression algorithm, they are different. In the GLM-OS setting the objects are weighted according to the Hessian entries, while in OS-regression they receive equal weights. Furthermore, in GLM-OS the least squares problems change at each iteration and are subproblems that serve as intermediate steps to get closer to the maximum of the (log-)likelihood. In OS-regression, the minimization of the loss function is the actual optimization problem.

4.3.4 Example: logistic regression with optimal scaling transformations

The GLM-OS algorithm as described previously can be applied to a variety of GLMs. In this paper, we focus on the logistic regression model, which is used when the outcome of interest is dichotomous. It models the probability π_i that observation i has response $y_i = 1$, given observed predictor values $\mathbf{x}_i$. To avoid the probability estimates to be negative or exceed one, a logistic distribution maps the weighted sum of (transformed) predictor variables η_i onto the unit interval, i.e.

$$\pi_i = \frac{1}{1 + \exp(-\eta_i)}, \quad (4.12)$$

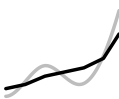
which represents the probability of success ($y_i = 1$) in a Bernoulli trial. The resulting likelihood function is

$$L(\boldsymbol{\eta}) = \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \prod_{i=1}^n \exp(\eta_i)^{y_i} \frac{1}{1 + \exp(\eta_i)}, \quad (4.13)$$

with corresponding log-likelihood

$$l(\boldsymbol{\eta}) = \sum_{i=1}^n y_i \eta_i - \sum_{i=1}^n \log\{1 + \exp(\eta_i)\}. \quad (4.14)$$

We use the modified Newton-Raphson method as described in subsection 4.3.2 to maximize (4.14) to find the optimal estimates for both the model parameters $\boldsymbol{\beta}$ and quantifications $\mathbf{v}_1, \dots, \mathbf{v}_p$. To simplify later calculations, we recast the maximization problem into a minimization problem and find the minimum of the negative of the log-likelihood.



To apply the algorithm, we need to derive the gradient vector $\nabla_{-l}(\tilde{\boldsymbol{\eta}})$ and the Hessian matrix $\mathbf{H}_{-l}(\tilde{\boldsymbol{\eta}})$ of $-l(\boldsymbol{\eta})$ w.r.t. $\boldsymbol{\eta}$ evaluated at the current estimate $\tilde{\boldsymbol{\eta}}$. The gradient is

$$\nabla_{-l}(\tilde{\boldsymbol{\eta}}) = \boldsymbol{\pi} - \mathbf{y}$$

and the Hessian is

$$\mathbf{H}_{-l}(\tilde{\boldsymbol{\eta}}) = \text{diag} \{ \boldsymbol{\pi}(1 - \boldsymbol{\pi}) \},$$

where $\boldsymbol{\pi}$ is the n -vector of probabilities π_i as defined in (4.12). Calculation details are provided in subsection 4.6.1.

The updates for quantifications $\mathbf{v}_k$ and coefficient β_k in each iteration are as follows

$$\tilde{\mathbf{v}}_k^+ = \tilde{\mathbf{v}}_k - \left[(\tilde{\beta}_k \mathbf{G}_k)^T \text{diag} \{ \boldsymbol{\pi}(1 - \boldsymbol{\pi}) \} \tilde{\beta}_k \mathbf{G}_k \right]^{-1} (\tilde{\beta}_k \mathbf{G}_k)^T (\boldsymbol{\pi} - \mathbf{y}) \quad (4.15)$$

and

$$\tilde{\beta}_k^+ = \tilde{\beta}_k - [(\mathbf{G}_k \tilde{\mathbf{v}}_k)^T \text{diag} \{ \boldsymbol{\pi}(1 - \boldsymbol{\pi}) \} \mathbf{G}_k \tilde{\mathbf{v}}_k]^{-1} (\mathbf{G}_k \tilde{\mathbf{v}}_k)^T (\boldsymbol{\pi} - \mathbf{y}), \quad (4.16)$$

where $\boldsymbol{\pi}$ is recalculated before updating $\tilde{\beta}_k$.

This algorithm that integrates OS transformations in the logistic regression model has been implemented in R software environment (R Core Team (2018)) to perform the analyses that will be described in the next section. This implementation of the algorithm uses some methods to speed up the calculations and save memory space by, for example, avoiding matrix multiplications with the sparse matrices $\mathbf{G}_k$.

4.4 Application of GLM with optimal scaling: logistic regression

In this section the GLM-OS method is applied to three different datasets. In all examples, we use a logistic regression model to predict a binary classification from a set of predictors. Each illustration focuses on a particular predictor type, namely categorical, ordinal and mixed data, and on different scaling levels which can be used to analyze these types of data.

4.4.1 Transformation and visualization of categorical predictors

We use a medical dataset to show how the OS methodology deals with categorical data by finding optimal quantifications for each category level. This approach is an alternative to the use of dummy variables, which is the standard approach for categorical predictors in GLMs. We will show how the replacement of dummy variables by quantifications will simplify the visualization and interpretation of the model, while it also benefits the computational process.

Data description

The first dataset is provided by the German multi-center project DINSTAP (Differentialle INdikationsstellung Stationärer und Tagesklinischer Psychotherapie; differential indication for inpatient and day clinic psychotherapy). The aim of the original project was to explore which criteria are used by clinicians to choose between an inpatient or a day clinic psychotherapy treatment.

Data on 25 possible predictors for treatment choice were collected. In the analysis illustrated in this section, we will only include the six most important variables for prediction (Hartmann et al. (2009)); namely *Need for medical care*, *Travel time*, *Need for relief from family conflicts*, *Need for relief from strain*, *Psychological restrictions of mobility*, and *Need to apply therapy in everyday life*.

Since this data analysis is for illustration purposes only, we focus only on the complete cases ($n = 342$). For 53.8% of these patients, clinicians preferred a day clinic treatment ($y = 0$), while for the others (46.2%) an inpatient treatment ($y = 1$) seemed more suitable.

We refer to Zeeck et al. (2009) for a description of the full dataset.

OS transformations with nominal scaling level

In the OS methodology, optimal quantifications for the categories of the predictor variables are found within the restrictions of the chosen scaling level. The least restrictive scaling level is a nominal transformation in which no ordering of the categories is taken into account. This scaling level best resembles the standard approach to handle categorical data, in which first dummy variables are defined to represent the category levels and then model coefficients are estimated for each dummy individually. Namely, if there are C_k categories for variable k , then $C_k - 1$ dummies are defined and hence $C_k - 1$ regression coefficients will be estimated, each indicating the effect of one category in comparison to the reference category. In contrast, optimal scaling assigns quantifications to all categories and estimates a single regression coefficient for each categorical predictor. Namely, the vector $\mathbf{v}_k$ of length C_k contains quantifications for the C_k categories and matrix $\mathbf{G}_k$ contains C_k columns representing all the categories, such that $\mathbf{G}_k \mathbf{v}_k$ gives the transformed predictor which is weighted by one regression coefficient β_k . If no restrictions (nominal scaling level) are applied to the quantifications $\mathbf{v}_k$ it will give similar results as analysis on dummy variables, but these results are represented differently, as shown in Table 4.1

The OS transformations for the six predictor variables are visualized by plotting the estimated quantifications against the original values of the category levels (Figure 4.1) and the estimated regression coefficients are given below each plot.



Category	Dummy coding	Optimal scaling
1		$\beta_k v_{k1}$
2	β_{k2}	$\beta_k v_{k2}$
$\vdots$	$\vdots$	$\vdots$
$C_k - 1$	β_{kC_k-1}	$\beta_k v_{kC_k-1}$
C_k	β_{kC_k}	$\beta_k v_{kC_k}$

Table 4.1: Contributions of the C_k category levels of a categorical variable k to the linear combination of predictor variables for the ordinary regression model with dummy coding and the optimal scaling model.

The lines that connect the dots have no meaning since there are no intermediate categories. However, their slopes visualize useful additional information about the relation between category levels. For example, a steep slope indicates a large difference between consecutive categories, while a small increase or decrease is indicated by a flat slope. In this way, the lines help interpreting the result and are therefore included in the plots.

The interpretation of the influence of a specific variable k in a model with OS transformations is via the estimated model coefficient $\hat{\beta}_k$ which indicates the strength of the effect, and via the estimated quantifications $\hat{v}_k$ which indicate the direction of the effect.

To understand which predictors have the strongest effect on the outcome, we first compare the regression coefficients. Given the values of the estimated coefficients, the predictors can be ordered according to the strength of their effect; i.e. *Travel Time* has the strongest effect ($\hat{\beta} = 1.41$), followed by *Need for relief from family conflicts* ($\hat{\beta} = 1.28$), *Need to apply therapy in everyday life* ($\hat{\beta} = 1.20$), *Need for relief from strain* ($\hat{\beta} = 1.06$), *Psychological restrictions of mobility* ($\hat{\beta} = 0.96$), and *Need for medical care* ($\hat{\beta} = 0.65$). The proportions of the model coefficients can also be used to draw conclusions. For example, we can conclude that the effect of *Need for relief from family conflicts* on the weighted sum of predictors is twice as big compared to the effect of *Need for relief for medical care* ($\hat{\beta} = 1.28$ vs. $\hat{\beta} = 0.65$).

The direction of the effect of a category level is given by the combination of its quantification and the sign of the predictor's model coefficient. For example, the large positive quantifications of the third category (often) of *Need for medical care* in combination with the positive model coefficient of this predictor indicates that if medical care is often required, a patient is more likely to be referred to inpatient treatment than day clinic treatment. Furthermore, this probability

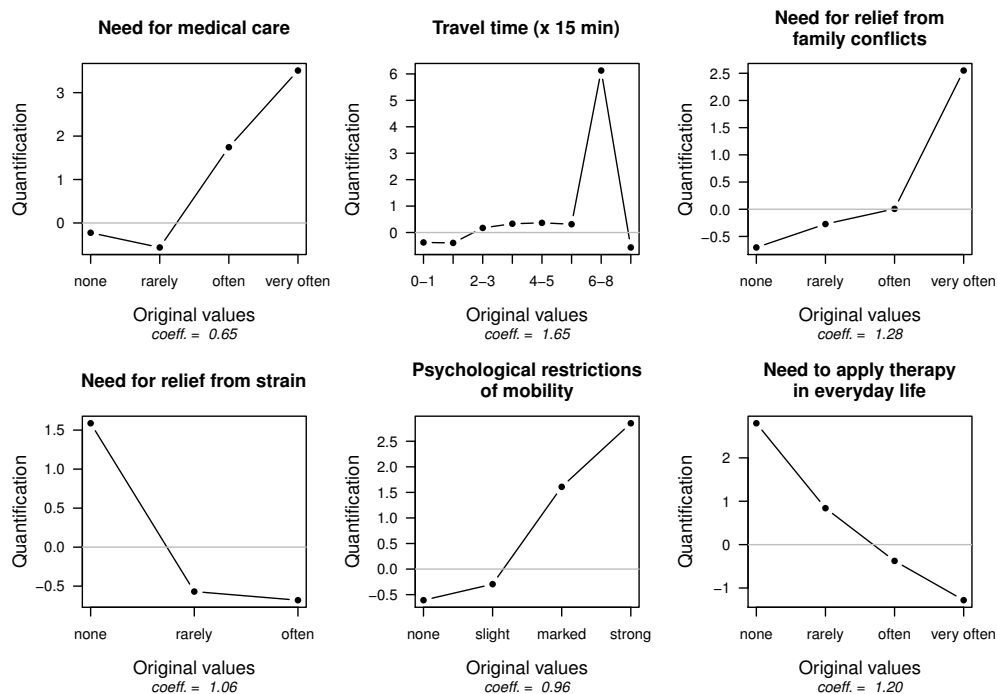


Figure 4.1: *DINSTAP* data: Nominal quantifications estimated for each of the original categories of the six predictor variables. Regression coefficients are provided below the plots of the corresponding predictor. The estimated intercept is 0.56.

increases if medical care is very often needed (fourth category). However, when no medical care is required (first category) or just rarely (second category), this will hardly influence a clinician's choice. A similar pattern is seen for *Psychological restrictions of mobility*. Additionally, only a strong need to be relieved from family conflicts seems to be a reason to choose for an inpatient treatment. Apparently inpatient treatment is believed to give additional mental stress, because this type of treatment is usually only given when there is no need for relief from strain. Moreover, the effect of the need to apply the therapy in everyday life seems to be almost linear in its categories. Surprisingly, a *Travel Time* of 6–8 quarters of an hour seems to be a strong indicator for inpatient treatment, while an even longer travel time is an indicator for day clinic treatment. This is a questionable result which might be due to the small number of observations in these two categories (8 and 6 patients relatively).

Concluding, the visualizations of the quantifications help to interpret the

results. A closer look at their exact values and the model coefficients will give a more detailed interpretation.

Comparison between OS transformations and the use of dummy variables

In standard logistic regression dummy variables are created to estimate the effects of each category level for all predictors on the outcome. The analysis on the DINSTAP data gives the estimates in Table 4.2.

These estimates should always be interpreted in terms of the reference category, thus the estimate -0.220 for category 2 of *Need for medical care* indicates that for patients who are classified in the second level of this predictor, the weighted sum of predictors is 0.220 lower than the weighted sum for those in the first category (= reference level). This coefficient represents the log of the odds ratio between these two categories. To compare the second and third categories, it is necessary to subtract the corresponding coefficients. Hence, to know whether being classified in category *rarely* instead of *often* or in *often* instead of *very often* has a bigger effect on the treatment choice, we have to compare the differences between their corresponding coefficients. Since $1.284 - (-0.220) = 1.504$ and $2.434 - 1.284 = 1.150$, this implies that the step from the second to the third category is larger than the step from the third to the fourth level. The same conclusion could be drawn by looking at the slopes in the quantification plots in Figure 4.1.

The similarity between the results obtained with optimal scaling and the use of dummy variables can be seen through the differences between the category quantifications. For example, the difference in the effect of categories 1 and 2 of *Need for Medical Care* in the optimal scaling result is the difference between the quantifications multiplied by the coefficient, $0.65 \cdot \{-0.563 - (-0.225)\} = -0.220$, which is precisely the coefficient for the corresponding dummy variable.

Hence, the results from ordinary logistic regression with dummy variables are essentially equal to the results of logistic regression with nominal scaling transformations, but they are represented differently. While the result for dummy variables focuses on the numeric coefficients only, OS puts more emphasis on visualization to improve the understanding of the quantification result, and provides regression coefficients for the predictors. The coefficients estimated for each dummy variable could also be plotted and the resulting figures will be very similar to those in Figure 4.1. However, most statistical software do not provide these plots as a default.

Category	1/ref	2	3	4	5	6	7	8
Intercept	3.34							
Need For Medical Care		-0.22	1.28	2.43				
Travel Time		-0.03	0.91	1.17	1.22	1.14	17.10	-0.31
Need For Relief From Family Conflicts		0.55	0.91	4.17				
Need For Relief From Strain		-2.29	-2.40					
Psychological Restrictions Of Mobility		0.30	2.13	3.32				
Need To Apply Therapy In Every Day Life		-2.34	-3.80	-4.88				

Table 4.2: DINSTAP data: Estimated model coefficients of the logistic regression model with dummy variables representing all categories of the categorical predictors.

4.4.2 Monotone transformations to facilitate interpretation

In the next illustration we use survey data to show the differences between nonmonotone and monotone quantifications for both ordered categorical and continuous data. If the prediction accuracy is not reduced significantly, it may be beneficial to put monotonicity constraints on the transformations.

Data description

For this illustration we use a subset of the 1987 National Indonesia Contraceptive Prevalence Survey data (Lim et al. (2000), available from the ICU Machine Learning Repository via <https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice>). The dataset contains several variables collected from married couples and their choice of contraceptive method. The categories of the outcome variable are no, long-term, or short-term use, and we merged the short- and long-term use into one category to create a binary outcome variable indicating whether couples use contraceptive methods ($y = 1$) or not ($y = 0$). There are nine predictor variables of which three are binary, four are categorical with ordered levels, and two are continuous. There are no missing values for any of the variables ($n = 1472$).

Nonmonotone vs. monotone quantifications

Since the values for most predictor variables in this dataset are ordered (namely for four categorical and two continuous variables), this dataset is suitable to compare nonmonotone and monotone transformations. For the categorical variables, either a nonmonotone or monotone step function are fitted, and for the continuous variables we use a (non)monotone spline transformation (of degree two with one interior knot). The results are shown in Figure 4.2.

Most estimated transformations are monotone even without imposing monotonicity, therefore the quantifications of the monotone and nonmonotone analyses are very similar for most predictors. The largest differences can be found for the variables *Education Husband*, *Occupation Husband*, and *Number of previous children*. However, although the results are very similar, it may still be beneficial to apply the monotonicity constraints, since it may simplify the interpretation of the result, or correspond better to the expected relation between the predictor variable and the outcome.

For example, if there are no monotonicity restrictions, the model indicates that if the husband is in the highest category of education, then the couple is less likely to use contraceptive methods, compared to the two middle categories. This result seems unexpected and it is difficult to explain. If monotonic restrictions are imposed, the quantifications of the three highest categories are equal and

very close to zero. This suggests that only a low education level of the husband is an indicator for no use of contraceptive methods (although the overall effect of this variable is small ($\beta = 0.05$)). From this example we see that monotonicity may simplify the interpretation of the quantifications, since there is no need to explain an unexpected decrease in effect.

A similar reasoning can be used for the quantifications of the *Number of previous children*. Namely, it is more plausible that the probability of using contraceptives increases with the number of previous children, than that there is a slight dip after 9 children and then again an increase after 12 children.

Hence, although differences with the nonmonotonic results are small, the monotonic quantifications of *Education Husband*, *Occupation Husband*, and *Number of previous children* are easier to interpret and correspond more to reality than the nonmonotonic ones.

Even though monotonic constraints ease interpretation, imposing too many restrictions on the transformations may hide the true relation between the predictor and outcome variables. Therefore it is important to check the model's performance for future observations before choosing for monotone scaling levels. This check can be done with cross-validation (CV). The results for this dataset are shown in Table 4.3.

As can be expected, the prediction errors based on the test data (EPE) using a 10-fold cross-validation are higher for both models compared to the apparent prediction error (APE) calculated on the training data. The increase is slightly smaller for the model with monotone transformations, but the difference between the models is very small (0.1875 vs. 0.1869). This suggests that applying monotonicity does not hide any important relation between the predictors and the outcome variable.

	APE	EPE	SE(EPE)	MCR (%)
Logistic Regression (linear)	0.206	0.211	0.0047	32.8
GAM (nonmonotone)	0.181	0.188	0.0052	28.3
GLM-OS (nonmonotone)	0.181	0.188	0.0053	28.7
GLM-OS (monotone)	0.181	0.187	0.0053	28.2

Table 4.3: Contraceptive method choice data: Apparent prediction error (APE) for the GLM-OS model with nonmonotone and monotone transformations, together with the 10-fold cross validation results: Expected Prediction Error (EPE) along with its standard error (SE(EPE)) and the Misclassification Rate (MCR). Results from standard logistic regression and GAM are added for comparison.

Relation with ordinary logistic regression and GAMs

In ordinary logistic regression, categorical predictors are included in the model by defining the category levels with dummy variables and by analyzing continuous data linearly. For categorical data, the dummy coding essentially gives the same result as transformations with a nominal scaling levels, although the result is represented differently (see subsection 4.4.1). The main difference between the ordinary logistic regression and nonmonotone GLM-OS results for this dataset is in the restrictions applied on the continuous variables. Namely, the linearity assumption for the logistic regression is more restrictive than the nonmonotone spline transformations in the GLM-OS.

In a GAM analysis, categorical variables are represented as dummy variables, as in ordinary logistic regression. Continuous variables are usually transformed using a nonmonotone spline function, but the algorithm to find the optimal spline is different from the algorithm used in OS. Therefore, the objective of GAMs is similar to nonmonotone GLM-OS, but the results for categorical data are represented differently and the nonmonotone splines are fitted in a slightly different way.

In Table 4.4 restrictions for ordinary logistic regression, (non)monotone GLM-OS, and GAMs are provided for comparison of the models.

The similarity of GAM and nonmonotonic GLM-OS is confirmed by the cross-validation results provided in Table 4.3. The small difference between the fitted splines have little influence on the predicted values of the observations.

Larger differences are seen for ordinary logistic regression. The prediction errors and misclassification percentages for the classic analysis are higher than those for (non)monotone GLM-OS and GAM. This suggests that the linearity assumption seems too strict for this dataset. Hence, imposing monotonicity will enhance interpretation, but imposing linearity (which would simplify the interpretation even more) will hide nonlinear relations between the predictors and outcome that are important for prediction.

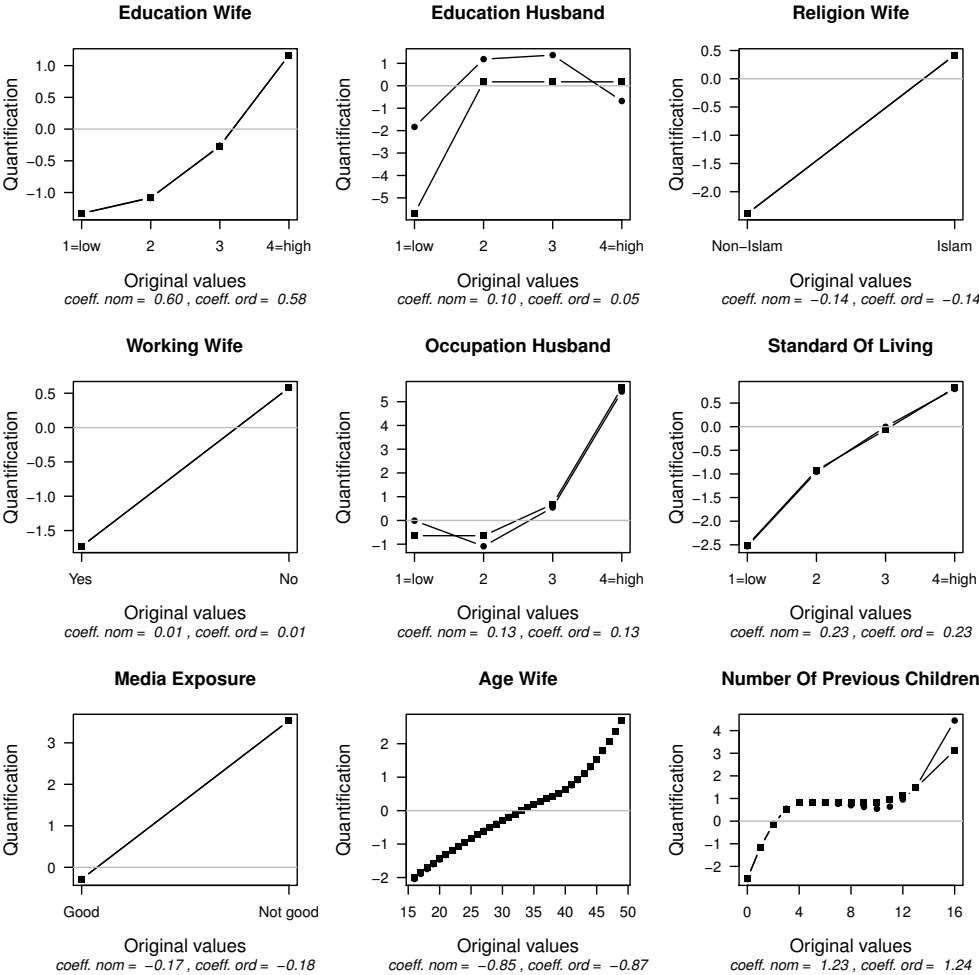
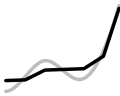


Figure 4.2: Contraceptive method choice data: Nominal (circles) and ordinal (squares) quantifications estimated for each of the original categories of the nine predictor variables. Categorical variables are transformed using step functions and continues predictors are transformed using splines. Estimated regression coefficients are provided below the plots of the corresponding predictor. The estimated intercepts are 0.30 for both the nonmonotone and monotone analyses.



Restrictions	Categorical predictors	Numeric predictors (or with many categories)
none	• Logistic regression with dummies • GAMs with dummies • GLM-OS with nonmonotonic step functions	• GAMs with nonmonotonic splines • GLM-OS with nonmonotonic splines
monotonic	• GLM-OS with monotonic step functions	• GLM-OS with monotonic splines
linear		• Logistic regression • GAMs with linear transformations • GLM-OS with linear transformations

Table 4.4: *Classification of the four models according to the restrictions applied to categorical and numeric data.*

4.4.3 Mixed scaling levels

Although monotone quantifications are usually easier to interpret, choosing a monotone scaling level is only correct if the predictor is also measured on an ordinal scale. For example, it would not make sense to impose monotonicity on the transformations of nominal categorical variables like countries, color, or blood type. On the other hand, nonmonotone restrictions can be applied to an ordered variable (categorical or numeric) if a nonmonotone relation is expected between this variable and the outcome. Therefore, it is important to choose each scaling level in accordance with the measurement level of the predictor.

In OS a different scaling level can be selected for each individual predictor. Usually this results in a model with a mix of scaling levels most suitable for the data. We will illustrate a mixed scaling level model with a medical dataset.

Data description

For this illustration we use the breast cancer recurrence dataset (M. Zwitter & M. Soklic, University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia; available from the ICU Machine Learning Repository via <https://archive.ics.uci.edu/ml/datasets/breast+cancer>). This dataset contains information on the binary response variable which indicates whether a patient experienced recurrence-events ($y = 1$) or not ($y = 0$). The aim is to predict the probability of recurrence-events from nine categorical and numerical predictor variables.

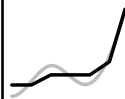
The predictor variables were measured on different scales. Variables *Node caps*, *Irradiation*, and *Breast* are categorical with two unordered categories. *Breast quadrant* and *Menopause* are categorical with more than two unordered categories. The *Degree of malignancy* is indicated by three category levels. These levels are ordered and a higher level indicates more abnormal cells. Finally, there are three numeric predictors that were discretized into categories; *Inv-nodes*, *Age*, and *Tumor size*. Unfortunately, the dataset does not include the original numeric values, so we can only use the discretized results.

The dataset contains 276 complete cases and the distribution of these observations over the predictors' categories is shown in Figure 4.3.

GLM-OS with scaling levels according to measurement level

Given that all predictors have different measurement levels, they require a different type of transformation in the logistic regression analysis. In the OS setting this can easily be done by selecting an appropriate scaling level for each variable.

For binary variables, all scaling levels will result in the same quantifications. Hence, for the three binary predictors, any scaling level could be chosen. To



reduce calculation time it is best to choose the numeric scaling level.

The categorical variables *Breast quadrant*, *Menopause*, and *Degree of malignancy* contain up to five categories. Since the levels of the first two predictors are unordered, a nonmonotone step function is the most appropriate. For *Degree of malignancy* a monotone step function is more suitable since its category levels are ordered.

The last three numerical predictors were summarized into categories. Predictors *Inv-nodes* and *Age* were discretized into 6 and 5 levels respectively and restrictions for a monotone step function are used to fit their quantifications. *Tumor size* was discretized into eleven categories and we choose a smooth transformation by fitting a monotone spline (quadratic, 1 interior knot).

Results based on the logistic regression analysis with OS transformations are given in Figure 4.4. Several conclusions can be drawn from this exploratory analysis.

The values of the estimated coefficients suggest that whether the cancer metastasizes to a lymph node (*Node caps*) has little influence on the probability of a recurrence-event; nor does the use of irradiation therapy. Furthermore, tumor location (indicated by *Breast* and *Breast Quadrant*) has a small effect.

From the quantifications of the categories of *Menopause*, it seems that the *lt40* stage is protective against recurrence-events. However, this result was based on only five observations (see corresponding bar plot in Figure 4.3), so more information should be collected from patients in this menopause stage to verify this result.

The ordinal predictors seem quite informative. For example, patients who were in the third degree of malignancy were more likely to get recurrence-events compared to those who were in one of the two lower levels. Furthermore, the transformations of the numerical predictors indicate that recurrence-events are more likely to occur if lymph nodes contain metastatic breast cancer (*Inv-nodes*), or if the tumor size was large. But, although the probability of a recurrence-event increases with tumor size, it barely increases once a tumor has reached size 25. Moreover, especially women of age 60 or older experience recurrence-events.

Comparison with nonmonotone scaling level and linearity restrictions

In the current analysis, all scaling levels are chosen to preserve all properties of the data. However, scaling levels with less restrictions may be chosen. For example, although *Age* was calculated on an ordinal scale, it does not necessarily imply that the relation between *Age* and the probability of recurrence-events is monotonic. Therefore, we may check whether a nonmonotone scaling level is more suitable for an ordinal predictor as well.

A cross-validation is used to compare the prediction accuracy from the previous model with those from the less restrictive analysis with only nonmonotone transformations. In the latter model, nonmonotone step functions were used to transform all variables except for *Tumor size*, for which a nonmonotone spline function (quadratic, 1 interior knot) was chosen. Results are shown in Table 4.5.

	APE	EPE	SE(EPE)	MCR(%)
GLM-OS (nonmonotone)	0.156	0.196	0.0154	28.6
GAM (nonmonotone)	0.150	0.193	0.0154	26.5
Logistic regression (1 variable linear)	0.154	0.189	0.0152	26.1
Logistic regression (4 variables linear)	0.166	0.187	0.0137	27.9
GLM-OS (mixed scaling levels)	0.156	0.179	0.0142	25.4

Table 4.5: *Breast cancer recurrence data: Apparent prediction error (APE) for the GLM-OS model with nonmonotone and monotone transformations, together with the 10-fold cross validation: Expected Prediction Error (EPE) along with its standard error (SE(EPE)) and the Misclassification Rate (MCR). Results from standard logistic regression and GAM are added for comparison.*

Cross-validation shows that the analysis with monotonicity restrictions produce smaller prediction errors on the test data and a smaller misclassification rate. This result suggests that a nonmonotone approach yields overfitting. So, in addition to easing interpretation of the quantifications, monotone transformations can prevent overfitting.

We also estimated a GAM on this data set. In this analysis we fitted a nonmonotonic spline transformation for *Tumor size* and all other predictors were defined with dummy variables. With these settings, GAM analysis closely resembles the nonmonotonic GLM-OS approach. This resemblance is supported by the similarity of the cross-validation results (Table 4.5).

We also estimated two ordinary logistic regressions with linearity assumptions. In the first analysis, we put linearity constraints on the *Tumor size* and included all the other variables as categorical data by defining dummy variables. In the second analysis, we put linearity constraints on all four ordinal predictors (i.e. on *Degree of malignancy*, *Inv-nodes*, *Age*, and *Tumor size*). The latter is the standard approach used if researchers want to preserve the category ordering. The cross-validation results for these models are also shown in Table 4.5.

When comparing the two logistic regression models with linearity assumptions, the cross-validation results show that the prediction error for the full dataset (APE) is much larger when the linearity restrictions are put on all four variables compared to only on *Tumor size*. However, in the cross-validation the prediction

errors (EPE) are almost similar, although the misclassification rate is slightly higher for the model with most restrictions.

When comparing the results from ordinary logistic regression to the GLM-OS results, we see that the prediction accuracy is in between the results of the models with nonmonotone and monotone scaling levels. This results suggests that applying no restrictions on the ordering will give the worst predictions. The prediction error can be improved by imposing the strict linearity restrictions for only one or four predictors. However, the most beneficial option is to impose monotonicity instead of linearity.

Concluding, analyzing the data with mixed scaling levels that are appropriate for the measurement levels of the predictors can help improve the results. When choosing the most suitable scaling level a cross-validation study is helpful to prevent overfitting.

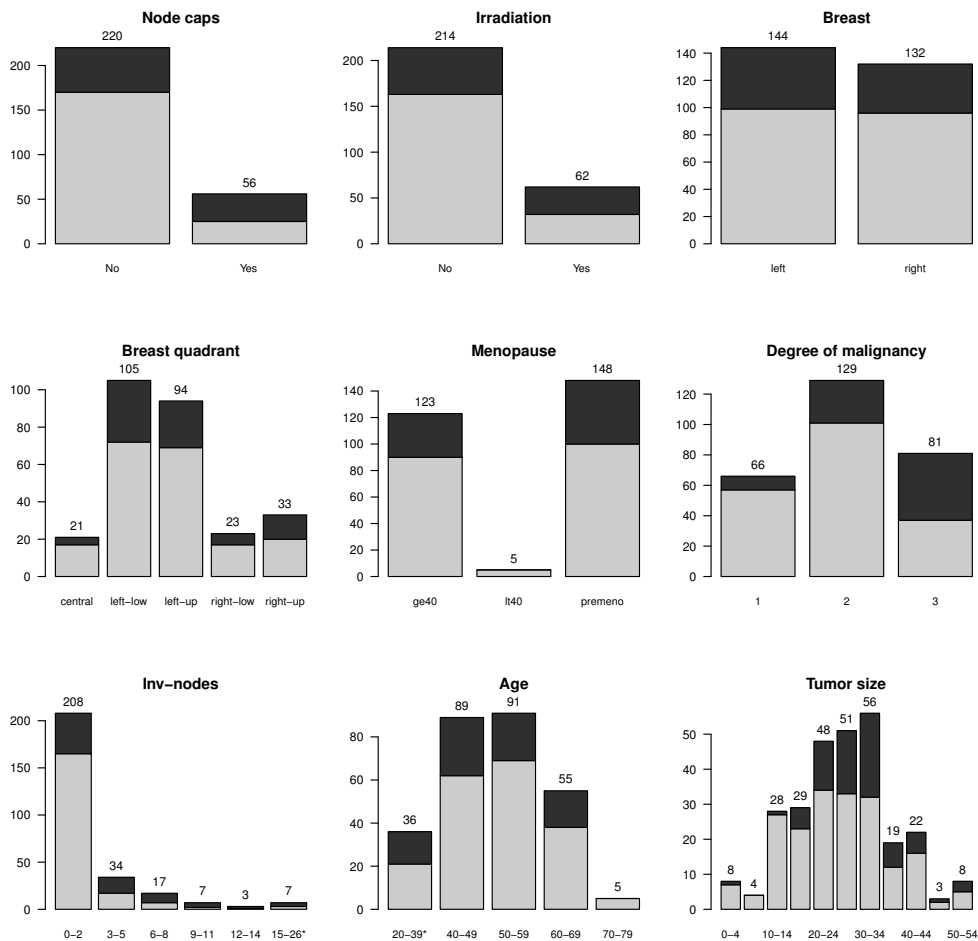


Figure 4.3: Breast cancer recurrence data: Distribution of observation over the categories of the predictor variables, split by outcome value (dark grey = recurrence events, light grey = no recurrence events).
*This category is a union of two categories that were merged because one of the original categories contained only one observation.

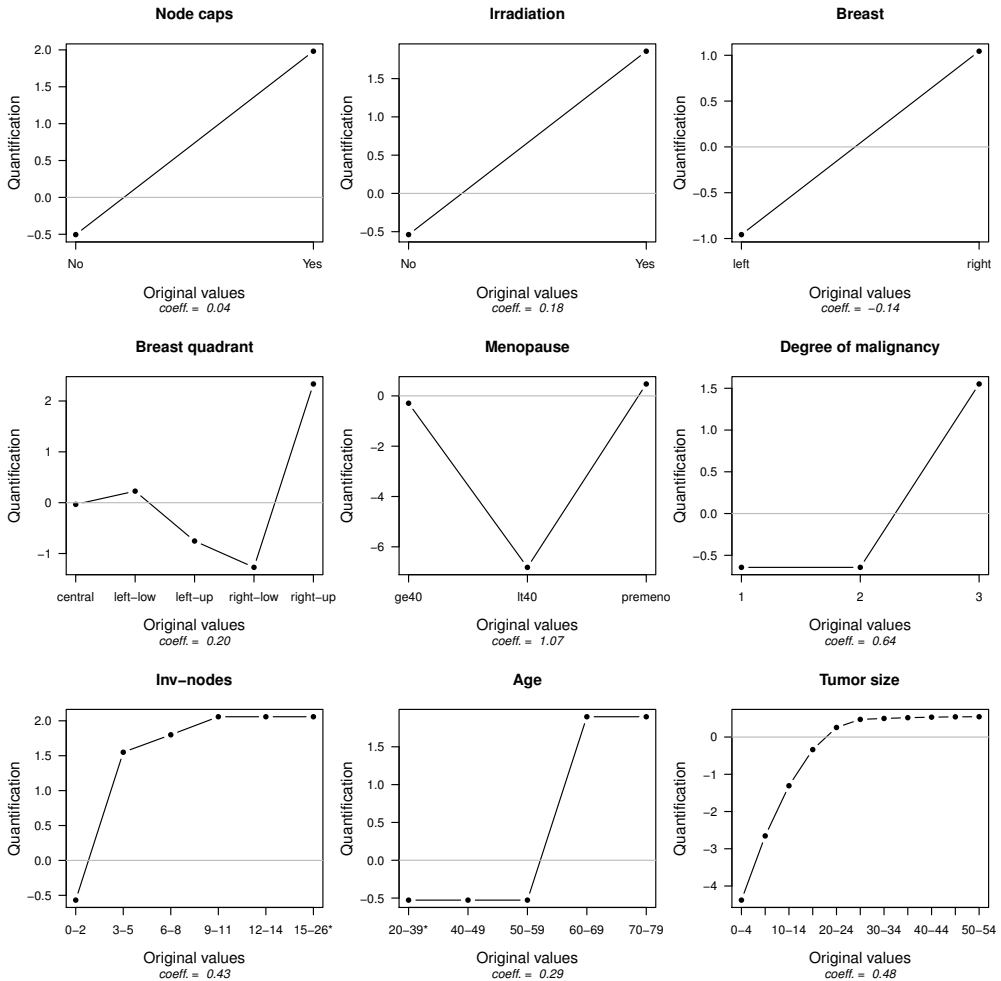


Figure 4.4: Breast cancer recurrence data: Quantifications estimated for each of the original categories of the nine predictor variables. Unordered categorical variables are transformed using nonmonotone step functions and ordered categorical variables are transformed using monotone step functions. Variable Tumor size is transformed using a monotone spline. Regression coefficients are provided below the plots of the corresponding predictor. The estimated intercept is -1.24 .

*This category is a union of two categories that were merged because one of the original categories contained only one observation.

4.5 Discussion

In this paper we have shown how OS transformations can be integrated in GLMs to transform predictors to optimize model fit and prediction accuracy. OS allows for nonlinear transformations that can be either nonmonotonic or monotonic, and are fit with a step or a spline function.

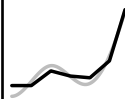
Transformations of the predictor variables have been integrated in GLMs before (Hastie and Tibshirani, 1990). However, the OS methodology has several benefits compared to other methods.

The strong focus of OS on categorical predictors results in a more flexible analysis method and an easier interpretation of the results for categorical data. While models like ordinary GLMs and GAMS use dummy variables to include categorical predictors, in the OS setting the predictor's categories are given optimal numerical values (quantifications) such that they can be interpreted in the same way as numerical data. Since the categories are quantified directly, no dummy variables are required, and, consequently, no reference category has to be chosen. This simplifies the interpretation. The quantifications are plotted against the original category levels to visualize the transformations and simplify interpretation. Quantifications and model coefficients are also provided numerically. Hence, while ordinary GLMs focus on the fitted numerical results only, OS puts emphasis on visualizing the result.

Another advantage of OS is the possibility to impose monotonicity restrictions on a transformation to preserve the ordering of category levels. This monotonicity restriction can be beneficial in two ways. First of all, a monotone transformation makes interpretation easier since an increase in category level implies an increase or decrease in response. Furthermore, by imposing more restrictions on the transformation, there is a smaller risk of overfitting on the training data, which may reduce the prediction error for new data.

GAMs usually apply nonmonotone transformations on numeric data and the standard approach for categorical data in which dummy variables are defined does by definition not preserve the order of the categories. Hence, the ability to apply monotonicity constraints to both categorical and continuous predictors is a unique property of OS.

In a GLM-OS, the scaling level can be individually chosen for each predictor variable in the model. Hence, the most appropriate combination of transformation restrictions can be selected for each individual predictor. GLM-OS with mixed scaling levels is a provides a flexible analysis method that can be applied to a large variety of data types, ranging from unordered categorical data to (ordered) numerical data.



Another feature of the OS technique is its group treatment of the category levels of a predictor variable. Namely, in OS a regression coefficient is obtained for each predictor to indicate its overall effect on the outcome (as in linear logistic regression) while no such diagnostic is obtained for logistic regression with dummy variables. In other words, in the OS setting the categories are no longer analyzed individually as is done when using dummy variables, but together as a group.

This grouping is extremely useful when applying regularization techniques. Namely, in an OS analysis, regularization can be done directly on the regression coefficients since these are estimated separately from the quantifications. Three regularization methods, Ridge regression, the Lasso, and the Elastic Net, were already implemented in OS-regression (Regularized Optimal Scaling Regression; ROS Regression (Meulman et al., 2019)), and these techniques can be implemented in GLM-OS in a similar manner. In other models that transform predictor variables in a GLM, the regression coefficients are incorporated in the variables' transformation and therefore regularization cannot be applied directly to the coefficients. Alternatives like Group Lasso (Yuan and Lin, 2006) and Blockwise Sparse Regression (Kim et al., 2006), to regularize a group or block of instead of the individual variables, have been suggested to remedy this. However, applying regularization directly to the regression coefficients in the OS model is more straightforward and gives the same model fit. Hence, the incorporation of regularization techniques is a useful future extension of GLM-OS.

Acknowledgements

Clinicians and researchers involved in the the DINSTAP-project are gratefully acknowledged for providing the dataset used in this chapter (J. von Wietersheim, H. Weiss, I. Sammet, E. Gaus, E. Semm, D. Hartms, A. Eisenberg, R. Rahm, and J. Küchemhoff).

4.6 Supplementary material

4.6.1 Calculating the gradient and Hessian of the negative log-likelihood function of the logistic regression model

In a logistic regression function, the outcome is binary, i.e. $Y \in \{0, 1\}$. The probability π_i of having outcome $y_i = 1$, given observed predictor variables $\mathbf{x}_i$, is modeled. To avoid that the probability estimates are negative or exceed one, a logit link function maps the linear combination of predictor variables, $\eta_i = \mathbf{x}_i\boldsymbol{\beta}$, onto the unit interval, i.e.

$$P(y_i = 1) = \pi_i = \frac{1}{1 + \exp(-\eta_i)} = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}. \quad (4.17)$$

Using this representation, the probability distribution for Y_i is

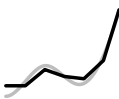
$$p(y_i) = P(Y_i = y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}.$$

Since observations are assumed to be independent, the likelihood function is product of marginal probabilities, i.e.

$$\begin{aligned} L(\boldsymbol{\eta}) &= \prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= \prod_{i=1}^n \left(\frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i) \\ &= \prod_{i=1}^n \exp(\eta_i)^{y_i} \left[1 - \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \right] \\ &= \prod_{i=1}^n \exp(\eta_i)^{y_i} \left[\frac{1 + \exp(\eta_i) - \exp(\eta_i)}{1 + \exp(\eta_i)} \right] \\ &= \prod_{i=1}^n \exp(\eta_i)^{y_i} \frac{1}{1 + \exp(\eta_i)}, \end{aligned}$$

and the corresponding log-likelihood function is

$$\begin{aligned} l(\boldsymbol{\eta}) &= \log \left[\prod_{i=1}^n \exp(\eta_i)^{y_i} \frac{1}{1 + \exp(\eta_i)} \right] \\ &= \sum_{i=1}^n y_i \eta_i - \sum_{i=1}^n \log[1 + \exp(\eta_i)]. \end{aligned}$$



To simplify computations the negative log-likelihood

$$l^*(\boldsymbol{\eta}) = \sum_{i=1}^n \log[1 + \exp(\eta_i)] - \sum_{i=1}^n y_i \eta_i$$

is minimized. The gradient of l^* is the vector with elements

$$\begin{aligned} \frac{\partial l^*(\boldsymbol{\eta})}{\partial \eta_i} &= [\log(1 + \exp(\eta_i))]' - [y_i \eta_i]' \\ &= \frac{1}{1 + \exp(\eta_i)} [\exp(\eta_i)]' - y_i \\ &= \frac{1}{1 + \exp(\eta_i)} \exp(\eta_i) - y_i \\ &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} - y_i \\ &= \pi_i - y_i. \end{aligned}$$

Since these partial derivatives are independent of η_j for $j \neq i$ all second-order mixed partial derivatives are zero. Hence, the Hessian is a diagonal matrix with diagonal elements

$$\begin{aligned} \frac{\partial^2 l^*(\boldsymbol{\eta})}{\partial \eta_i^2} &= \frac{[\exp(\eta_i)]'(1 + \exp(\eta_i)) - \exp(\eta_i)[1 + \exp(\eta_i)]'}{(1 + \exp(\eta_i))^2} - 0 \\ &= \frac{\exp(\eta_i)(1 + \exp(\eta_i)) - \exp(\eta_i) \exp(\eta_i)}{(1 + \exp(\eta_i))^2} \\ &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} - \frac{\exp(\eta_i)^2}{(1 + \exp(\eta_i))^2} \\ &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \left(1 - \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \right) \\ &= \pi_i(1 - \pi_i). \end{aligned}$$

In matrix notation,

$$\begin{aligned} \boldsymbol{\nabla}(\boldsymbol{\eta}) &= \boldsymbol{\pi} - \mathbf{y}; \\ \mathbf{H}(\boldsymbol{\eta}) &= \text{diag} \{ \boldsymbol{\pi}(1 - \boldsymbol{\pi}) \}, \end{aligned}$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n)$ as defined in (4.17).

VARIABILITY IN THE INTERPRETATION OF PROBABILITY PHRASES USED IN DUTCH NEWS ARTICLES - A RISK FOR MISCOMMUNICATION

5

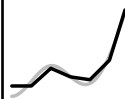
Verbal probability phrases are often used in science communication to express estimated risks in words instead of numbers. In this study we look at how laypeople and statisticians interpret Dutch probability phrases that are regularly used in news articles. We found that there is a large variability in interpretations, even if the phrases are given in a neutral context and even among statisticians. We conclude that experts and the media should be careful in using verbal probability expressions.

5.1 Introduction

Every day people make decisions based on estimated probabilities and risks. These decisions range from choices with little consequences (“*Should I bring my umbrella to avoid getting wet in the rain?*”) to important life decisions (“*Which treatment will most likely cure my cancer without causing too many undesirable side-effects?*” or “*Should I evacuate for the approaching storm?*”). Many of our decisions rely on risks expressed by others (weather forecasters, oncologists, or scientists). Due to this dependence of the decision maker on the information provider, it is important that the message is understood as intended in order to minimize the risk of miscommunication.

Many estimated probabilities are communicated verbally and in that case it is important that the interpretation of the verbal probability phrase is the same for both sender and receiver. For example, an oncologist may predict, based on his or her experience, that there is a 90–95% probability of a cure for a particular patient. The doctor may then use the expression *very likely* to communicate this

This chapter is submitted as Willems, S. J. W., Albers, C. J., and Smeets, I. *Variability in the interpretation of probability phrases used in Dutch news articles - a risk for miscommunication.*



Probability term	Subjective Probability range
Almost certain	99–100%
Extremely likely	95–99%
Very likely	90–95%
Likely	66–90%
About as likely as not	33–66%
Unlikely	10–33%
Very unlikely	5–10%
Extremely unlikely	1–5%
Almost impossible	0–1%

Table 5.1: *Approximate probability scale recommended for harmonized use in the European Food Safety Authority (EFSA) to express uncertainty about questions or quantities of interest (European Food Safety Authority et al., 2019).*

probability. However, if the patient interprets this verbal probability expression as a 75% probability, he or she may unnecessarily have lower expectations.

Some organizations and bureaucrats use probability scales as in Table 5.1 to standardize their risk communication. But how well do these translations match with how people actually interpret these phrases?

Early studies on the interpretation of probability phrases typically used the ‘how likely’ approach; respondents were asked to give their interpretation of a probability expression as a single value or range on a scale of 0–1 or 0–100%, or were asked to rank them. The phrases were either presented out-of-context or in sentences describing a particular situation. Many of these studies were summarized in the literature reviews by Druzdzel (1989) and Visschers et al. (2009), and the meta-analysis by Theil (2002).

The overall conclusion from these studies was that, although individuals seem to be internally consistent in their ranking of probability phrases (Budescu and Wallsten, 1985) and their perception of them over time (Bryant and Norman, 1980), the interpretation of these phrases varies greatly among individuals. This interpretation variability is especially large for phrases expressing a probability in the range from 20% to 80%. For words that express extreme probabilities, such as *always*, *certain*, *never*, and *impossible*, consensus was highest. This variability of interpretations is represented by the varying widths of the subjective probability ranges in probability scales as in Table 5.1. These wide ranges complicate communication, because it is impossible to express a very specific probability.

Several studies also showed that the numerical interpretations of some prob-

ability phrases overlap or are very similar. For example, Reagan et al. (1989) concluded that *likely* is synonymous with *probable*, and *low chance* with *unlikely* and *improbable*. Synonymous words have overlapping probability ranges which would complicate a probability scale. The codification presented in Table 5.1 seems to avoid this complication by limiting the vocabulary to phrases with nonoverlapping ranges.

Furthermore, translation issues for verbal probability expressions are important for all international organizations that publish their documents in more than one language. For example, a question that may arise within the European Food Safety Authority is whether their probability scale (Table 5.1) translates directly to other European languages, or whether the subjective probability ranges in the second column should be adjusted, and consequently, the expressions in the documentation text.

Most research on the numerical interpretation of probability phrases was conducted in English. There have been some replication studies in other languages, among which the Dutch language. Most of the Dutch studies are over twenty years old. For instance, Eekhof et al. (1992) focused on the interpretation of 30 Dutch phrases. However, all phrases in this study expressed frequencies instead of probabilities. In a later study by Timmermans (1994) some probability phrases were included, usually in combination with an adverb like *quite* or *rather*. Unfortunately, the paper is written in English and does not provide the Dutch expressions used in the study, hence it is unclear exactly which Dutch expressions and adverbs were investigated. In a study by Pander Maat and Klaassen (1996), focus was on the interpretation of uncertainty in information leaflets that come with medicine. Although their main interest was not in the numerical values associated with verbal probability phrases, they did investigate this for three phrases. Renooij and Witteman (1999) did several experiments to develop a probability scale containing both words and numbers. Their focus was on ranking seven probability phrases and developing their corresponding numerical scale. Given that the first study included many phrases but only frequencies, and the other three studies included only a few probability phrases, usually in combination with adverbs, many Dutch probability expressions still needed to be studied.

In addition to replication studies in other languages, several studies have been done to compare the interpretation variability of English probability phrases with the interpretations of their translations to other languages. Three studies, comparing English with French (Davidson and Chrisman, 1994), German (Dounnik and Richter, 2003), and Chinese (Harris et al., 2013), showed that on average the numerical interpretations of the English phrases differ from the interpretation of their counterparts in the three other languages. Additionally,



in French and Chinese, the standard deviations of the numerical values related to the probability phrases were much larger than those of the original English wording.

These results show that the meaning of probability expressions can get lost in translation from one language to another.

5.2 Background

5.2.1 The communication mode preference paradox

Until recently, it was generally believed that information providers, the senders of a message, prefer to express probabilities verbally, namely by using verbal probability expressions as *unlikely*, *usually* and *maybe*, while decision makers favor numeric expressions like percentages. Druzdzel (1989) reasoned that senders prefer verbal expressions because these convey some amount of uncertainty. Including this uncertainty in the expression is favored by senders, because probability estimates are usually based on empirical data and therefore not sufficiently precise to be translated into exact numerical statements. Hence, if a numerical value is given its suggested precision may be misleading. On the other hand, decisions makers prefer this precision of numerical expressions, since numeric values are easier to compare and to draw conclusions from. Erev and Cohen (1990) referred to this difference in preference as the *communication mode preference paradox*. In more recent studies, researchers have challenged this theory, but the results are not conclusive. For example, Juanchich and Sirota (2019) concluded that people favor verbal phrases in general, but in some contexts or for specific purposes numerical expressions are preferred.

5.2.2 Asymmetry

A complication in the interpretation of probability phrases is asymmetry. For example, based on the discovery of the synonymous pair *low chance* with *unlikely* and *improbable*, Reagan et al. (1989) also expected *high chance* to be synonymous with *likely* and *probable*. However, their data indicated that actually *very likely* and *very probable* are its synonyms. This unbalanced result shows that there is some asymmetry in the interpretation of mirrored probability phrases.

This phenomenon of asymmetry is studied and confirmed by many researchers. In most studies, this imbalance is investigated on a group level by comparing the group means or medians of two complementary phrases. For instance, Lichtenstein and Newman (1967) concluded that the interpretations of *likely* and *unlikely* are asymmetric, since their means sum to $(72\% + 18\% =) 90\%$ and their medians sum to $(75\% + 16\% =) 91\%$ instead of 100%. This asymmetry was

confirmed by both Reagan et al. (1989) (medians sum to 90%) and Stheeman et al. (1993) (medians sum to 80%). Furthermore, Lichtenstein and Newman (1967) focused on the influence of adverbs (such as *very*, *quite* and *fairly*) and found that, for instance, the means of the numeric probabilities given to *quite likely* and *quite unlikely* sum to $(79\% + 11\% =) 90\%$ instead of 100%. Previous studies have also shown that some terms actually are (almost) symmetrical. For example, *very likely* and *very unlikely* (mean interpretations sum to 96%, Lichtenstein and Newman (1967)), and *almost always* and *almost never* (median interpretations sum to 98%, Stheeman et al. (1993)).

Some mirrored terms have a clear linguistic explanation for their asymmetry. For example, Mosteller and Youtz (1990) studied the terms *possible* and *impossible* and found that the interpretation of *impossible* is stable (around 3% for all participants of the study), while *possible* has distinct meanings for different groups of people. Namely, some respondents used the literal interpretation of *possible* and indicated that it could express any percentage between 0% and 100%, and others associated it with rare events that only scarcely occur (as in *barely possible*). Hence, the different interpretations of *possible* causes the strong asymmetry with its mirrored expression *impossible*. The asymmetry in the interpretation of *certain* and *uncertain* can be explained in a similar way.

The asymmetry in the interpretation of verbal probability expressions complicates the development of probability tables. For example, the symmetry of the probability scale in Table 5.1 simplifies the use of the table, but it does not necessarily represent the actual probability ranges of its terms.

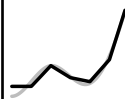
All these research results show that the interpretations of verbal probability expressions vary too much to translate them into a (symmetrical) probability scale of which the numerical probability ranges would be supported by everyone. Therefore, many researchers who initially intended to make a translation table, concluded that such a codification is practically impossible (Lichtenstein and Newman (1967); Mosteller and Youtz (1990); Timmermans and Mileman (1993); Weber and Hilton (1990)), or realized that their currently used table was actually not conveying the intended probabilities (Pander Maat and Klaassen, 1996). Yet, still many organizations are using tables like this.

5.2.3 Context dependence

The interpretation of a probability phrase is influenced enormously by its context. For instance, compare your numerical interpretation of the word *likely* in the next two statements:

- *It is likely that it will rain in Manchester, England, next June;*
- *It is likely that it will rain in Barcelona, Spain, next June.*

Probably, your numerical interpretation of *likely* in the first statement is higher



than in the second. Wallsten et al. (1986) used this example and, based on their research, predicted a difference in the numerical interpretation of these statements. Namely, in their study, they showed that an individual's expected base-rate expectation of a context scenario influences this person's interpretation of the probability phrase. In this example the base-rate for the first scenario is higher (in spring rain is more probable in England than in Spain) and this influences the interpretation of the word *likely*.

This hypothesis on the base-rate effect was confirmed by Weber and Hilton (1990), who additionally provided evidence that other variables may be affecting the interpretation as well. According to their findings, the perceived severity or consequentiality of an event and its emotional valence will also influence the judged probability.

Since it was shown that context may influence the interpretation of probability phrases, many researchers decided to investigate them out-of-context. However, it was argued by Druzdzel (1989) that, if no specific context is provided, participants may invent their own context. Due to these self-created contexts, participants' responses will portray the interpretation of the probability phrases in many completely different contexts instead of out-of-context. These different scenarios may cause extra variability in the data which makes it more difficult to draw conclusions from the results.

5.2.4 Differences between sub-populations

In most studies, data on the interpretation of probability phrases was gathered within specific sub-populations. Participants were, for instance, physicians (Bryant and Norman, 1980), science writers (Mosteller and Youtz, 1990), radiologists (Stheeman et al., 1993), biological scientists (MacLeod and Pietravalle, 2017), or patients (Pander Maat and Klaassen, 1996). Although all these studies showed variability in the perception of probability phrases within these sub-populations, one might wonder whether there are any differences between these groups as well. For example, Theil (2002) argued that there may be a difference between professionals, who regularly make and communicate probability estimations, and persons who are inexperienced in this respect. However, his meta-analysis did not provide evidence for this hypothesis.

In studies on the use of jargon, it has been shown that there is a significant difference in the interpretation of medical terms between doctors and patients (Boyle, 1970) and of hydrological vocabulary between experts and laypeople (Venhuizen et al., 2019). Experts may be unaware of this difference (Castro et al., 2007) and, hence, their use of jargon may cause a miscommunication of information.

Given these results on the different interpretations of jargon, there is reason

to believe that there may be differences between the numerical interpretations of probability expressions of experts and laypeople as well, as Theil (2002) suggested. If this hypothesis is correct, experts may be misunderstood if they express probabilities verbally.

5.2.5 Gaps in the literature

Summarizing, we see that despite ongoing interest in and usage of verbal probability expressions, there are large gaps in the literature. Furthermore, in most studies on this topic, the sample sizes were quite small. For instance, the number of participants in the Dutch studies lay between 78 (Timmermans, 1994) and 101 (Eekhof et al., 1992). The English studies have comparable sample sizes, for example, in the nine studies mentioned by Theil (2002) the median number of participants is 52 and the mean is 170.

Therefore, we set up a large-scale study for the interpretation of Dutch verbal probability expressions, presented in a neutral context which are based on ordinary events. In this way, we investigate whether the variability in interpretation is also high when the context is barely susceptible to prior beliefs.

Additionally, we check for synonymous phrases and asymmetry since these two characteristics are well studied in English but have not yet been analyzed in Dutch studies. Furthermore, we compare the results of statisticians with those of laypeople to check whether experts use different interpretations.

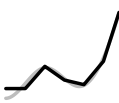
5.3 Methods

We used a survey design where probability phrases were presented in a neutral sentence to participants, and they could give their interpretation as a point estimate on a 0–100% scale.

5.3.1 Choice of phrases

There are many Dutch probability and frequency phrases that can be studied. To make a selection for our study, we first listed the phrases used in the English studies and translated them to Dutch. For translation Google Translate (Google, 2018) and the leading Dutch dictionary Van Dale (Van Dale Uitgevers, 2018) were used. If more than one translation was appropriate, both were added to the list. Then we added the expressions from previous Dutch studies (Eekhof et al., 1992; Pander Maat and Klaassen, 1996; Renooij and Witteman, 1999). This resulted in a list of 131 phrases.

This list was too long to use in one survey, so a selection had to be made. Since the most frequently used phrases are also the most relevant, we selected



the verbal probability expressions that were used at least 100 times in all online available articles of the popular Dutch news website *nu.nl*. To prevent too much overlap with the research by Eekhof et al. (1992), only the ten most commonly used frequency phrases were selected. Furthermore, some combinations of adverbs with a probability phrase were removed from the list to prevent too much overlap with the study by Timmermans (1994), and to prevent repetitions of very similar phrases. Additionally, the word *undecided* was removed, since it was mostly used in sport results where it has a different meaning.

This method of phrase selection resulted in a list of 29 frequency and probability expressions. These phrases, and their English translations, are given in Table 5.2 of the supplementary material (subsection 5.7.1). In the text of this paper, we will use the English translations. Please keep in mind that all given numerical interpretations for these phrases are actually for their Dutch counterparts.

5.3.2 Context

As described before, the interpretation of a probability expression may be influenced by a person's prior expectations of the phrase's context. To avoid these base-rate effects, our aim was to formulate sentences that are neutral in the sense that everyone can imagine the situation but has little prior expectations about it. Some examples of the statements, formulated with the probability phrase *likely*, are

- *It is **likely** that this plan succeeds.*
- *It is **likely** that this hotel is fully booked.*
- *It is **likely** that the team wins a match.*

We tried to minimize the base-rate effect by not specifying a specific plan, hotel, or team. We developed twelve sentences like these. The complete list of these contexts is given in Table 5.2 of the supplementary material (subsection 5.7.2). In each sentence the verbal probability expression was printed in bold to direct more attention to it.

5.3.3 Numeric interpretations

For each probability expressions in the survey, participants gave the point estimate of their numerical interpretation in percentages (0–100%) by using a slider. After the statement, each survey item was formulated as a question. For example, the questions related to the three statements above were formulated as:

- *What is the probability (expressed in percentages) that this plan succeeds?*
- *What is the probability (expressed in percentages) that this hotel is fully booked?*
- *What is the probability (expressed in percentages) that the team wins a match?*

All probability phrases were presented individually and in a random order, and participants were required to answer each question before continuing to the next. In this way, missing data was prevented.

5.3.4 Randomization

To prevent a systematic influence of the context on the interpretation of the probability phrase, 12 different versions of the survey were created. In every version, the probability phrase was formulated in a different context and contexts were repeated two or three times in each survey version (since 29 is not divisible by 12). All survey versions were evenly and randomly distributed among the participants by the survey software Qualtrics (2005-2018).

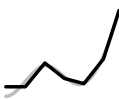
5.3.5 Personal characteristics

After giving their interpretation of the 29 phrases, participants were asked for some personal information. This included whether they are a statistician, their highest completed education level, their age, and their gender. Statisticians were self-reported and this was questioned as “*Are you a statistician or do you perform statistical analyses on a weekly or monthly basis?*”. Education was categorized in six common categories of degrees in the Netherlands. Age was categorized in intervals of 20 years. These wide intervals were chosen to protect the anonymity of the participants and because the exact ages were not of particular interest for this research. However, age was included to check whether both young and older people participated. As with age, gender is not of particular interest for this study, but it was included to check whether participants are almost equally distributed among the genders.

All personal characteristics were asked as multiple-choice questions and participants could select one of the given categories. Participants were allowed to refrain from providing their age and gender.

5.3.6 Pilot

A pilot study showed that the length of the survey was reasonable (approximately ten minutes) and that the explanation was clear. We noticed that some participants had the tendency to base their interpretation of a phrase on their interpretations of previous phrases. This confirms that randomization of the phrases is necessary. Additionally, it supported our decision to present one phrase at the time and to not allow participants to change their answers to previous questions. If we had permitted this, participants may have ranked their answers instead of giving the interpretations individually, which may have influenced the results. Based on the pilot study, we decided to make the original question “*Are*



you a statistician or do you perform statistical analyses on a regular basis?" more specific by changing "*on a regular basis*" into "*on a weekly or monthly basis*".

5.3.7 Survey distribution

We obtained permission to distribute this survey from the ethical committee of the Faculty of Behavioural and Social Sciences of the University of Groningen in the Netherlands (17451-O). Since we wanted to compare the interpretations of Dutch-speaking statisticians with those of non-statisticians, the survey was distributed among both groups. Statisticians were invited to participate via the mailing list of the Netherlands Society for Statistics and Operations Research (VVSOR) and the Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS). To reach non-statisticians, the survey invitation was distributed via the personal Twitter (Twitter Inc., 2018) accounts of the three authors (one of the authors is a public figure and has over 60,000 followers, many of which are not in the academic community). Their followers were asked to participate and to share the survey in their network.

5.4 Results

5.4.1 Participants' characteristics

The survey was open for participation for almost four months, namely between July 18th, 2018 and November 8th, 2018. During this time, 1004 persons started the survey, of which 115 did not finish it. These incomplete observations were removed from the data. Another 8 participants were excluded from the analysis, because their native language was not Dutch. As a result, the data contains the responses of 881 participants.

The participants are evenly distributed among the genders (430 male vs. 440 female). There were many more non-statisticians than (self-reported) statisticians (655 vs. 226). Their distribution among the age groups and education levels is displayed in Figure 5.1. The first bar plot indicates that most participants were equally distributed among the two middle age groups (20–39 years and 40–60 years). The second bar plot shows that many of the participants were highly educated. This is partially explained by the fact that most statisticians have an academic education (94%), but even among the non-statisticians, the proportion of academically educated persons is large (58%). Furthermore, there are more males than females among the statisticians (59% male) and more females among the non-statisticians (55% female).

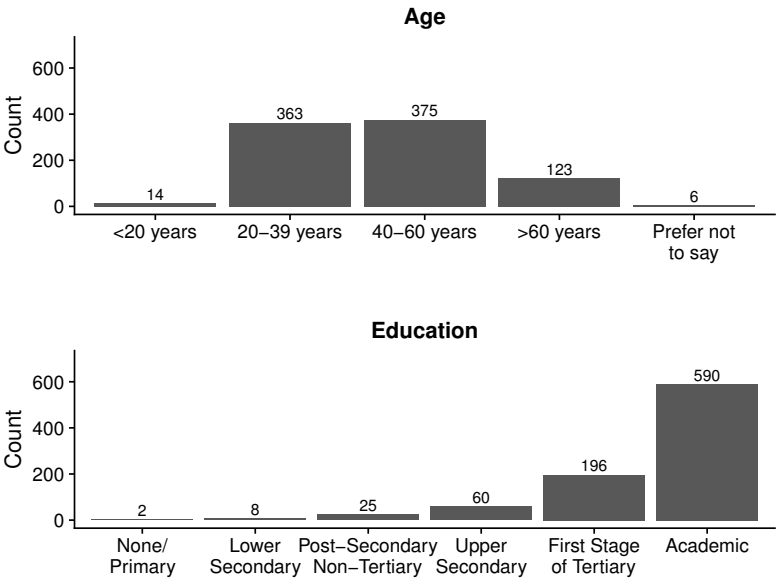


Figure 5.1: Bar graphs of the number of participants in each of the category levels of variables Age and Education.

5.4.2 Interpretation of probability phrases

The distributions of the interpreted percentages of each probability phrase are displayed by the density plots in Figure 5.2 and the mean values and 5% and 95% percentiles are listed on the right side of the plots. The 5% and 95% percentiles indicate the range of interpretations of 90% of the participants.

There seems to be some consensus about the interpretation of extreme words like *always*, *certain*, and *impossible*. Namely, the intervals between their 5% to 95% percentiles have a width of about 20 percentage points. Surprisingly, the 95% percentile of the extreme phrase *never* is at 32%, which seems high for this expression.

There is even less consensus for phrases that do not represent an extreme probability. Namely, their numerical interpretations have percentile ranges with widths up to 50 percentage points. For example, 90% of the respondents interpreted the verbal probability expressions *sometimes*, *probable*, and *almost always* between, respectively, 11–55%, 41–86%, and 70–96%.

Other things to notice are the small peaks in the density plots which indicate that participants often express probabilities as multiples of ten. Also, there was no phrase in our survey that represents 50%. The candidates *liable to happen*, *chance*, *uncertain*, *maybe*, and *possible*, for which 50% is the most frequently

chosen interpretation, all have a large tail to the left and percentile ranges of 42–50 percentage points.

5.4.3 Asymmetry

For the usability of verbal probability expressions, (a)symmetry in the interpretation of mirrored verbal probability expressions are of interest. The imbalance in their interpretation is often investigated by reviewing whether the group means or group medians of the interpretations of two complementary words sum to 100%. The groups means from our data are listed in Figure 5.2, and show that, as in English, asymmetry is present for the Dutch translations of *likely* and *unlikely*. Namely, the mean interpretation of *likely* in our data is 75% and the mean for *unlikely* is 16%, and hence these sum to 91%. Symmetry is found for phrases as *very likely* and *very unlikely* (sum to 95%), *almost always* and *almost never* (sum to 100%), and *often* and *not often* (sum to 97%).

The results from previous studies and those listed above are based on the results on a group level (group means). We also looked at the results on an individual level by plotting the density of the sums of complementary phrases, see Figure 5.3. These plots show that there are some mirrored pairs which interpretation sums up to about 100% for most participants, for example *(almost) always* and *(almost) never*, and *very likely* and *very unlikely*. Other complementary phrases were interpreted asymmetrically by many participants and usually sum up to slightly less than 75% to 100%, for example *likely* and *unlikely*, and *often* and *not often*.

As explained in the introduction, in some cases asymmetry has a linguistic cause. Our results on the interpretation of possible and impossible confirm the findings of Mosteller and Youtz (1990). Namely, Figure 5.2 shows that *impossible* has a stable interpretation that is close to 0%, while *possible* has a broad interpretation from 20% to 70% which peaks around 50%. The asymmetry is also confirmed by the distribution of their sums in Figure 5.3.

A similar pattern is found for *certain* and *uncertain*; there is a consensus on the interpretation of *certain* (around 100%) while the perception of *uncertain* varies a lot and is comparable to *maybe*'s interpretation, namely some value between 20% to 50% (see Figure 5.2). As a result, the percentages of *certain* and *uncertain* always sum to more than 100% and together peak at 150% (see Figure 5.3).

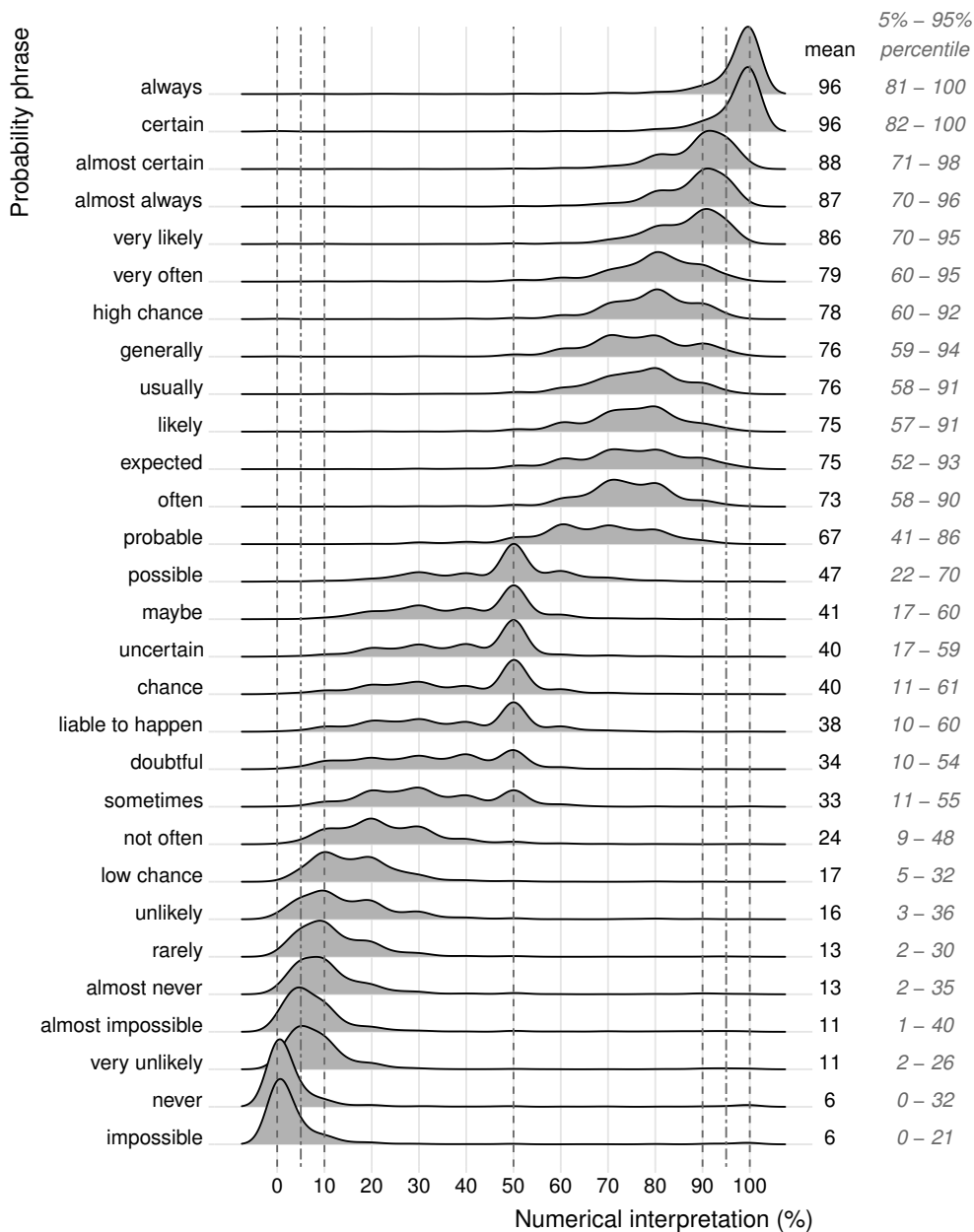


Figure 5.2: Density plots and mean values of the numerical interpretations (in percentages) given by all participants for each phrase in the survey. Note that density plots are a smooth variant of histograms and may therefore be positive outside the data range of 0–100%.

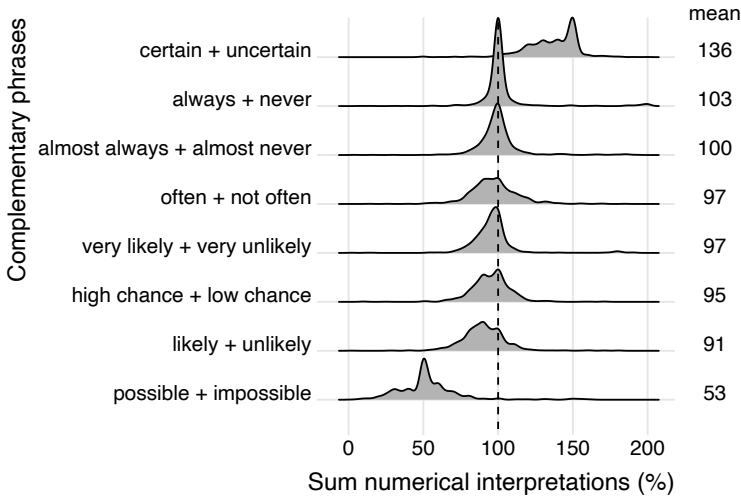


Figure 5.3: *Density plots and mean values of the sums of the numerical interpretations (in percentages) given by all participants for the complementary phrase pairs in the survey.*

5.4.4 Context

One of our concerns was that the context of the sentences influences the perception of the probability phrases. To avoid the base-rate effect, we tried to formulate the context sentences as neutral as possible.

To check whether we succeeded in our intention, we investigated the variability of the interpretation of phrases among different contexts. Figure 5.4 shows the mean percentages given by the participants to each probability phrase, grouped by context. This plot shows that, in general, the means of a phrases are very similar for each context, with a maximum of 20 percentage points difference between contexts. Most of this variability appears for words that represent 30% to 80%.

Most importantly, although the plots show some influence of context on the interpretations, they do not suggest that any of the sentences is systematically interpreted differently (higher/lower or more/less extreme) from the others.

5.4.5 Differences between sub-populations

One of the aims of this research was to make a comparison of the interpretation of probability phrases of different sub-populations, namely to compare interpretations of experts (statisticians) with those of laypeople. Figure 5.5 shows the density plots of the statisticians and non-statistician for a selection of five probability phrases. These expressions were selected from different ranges of

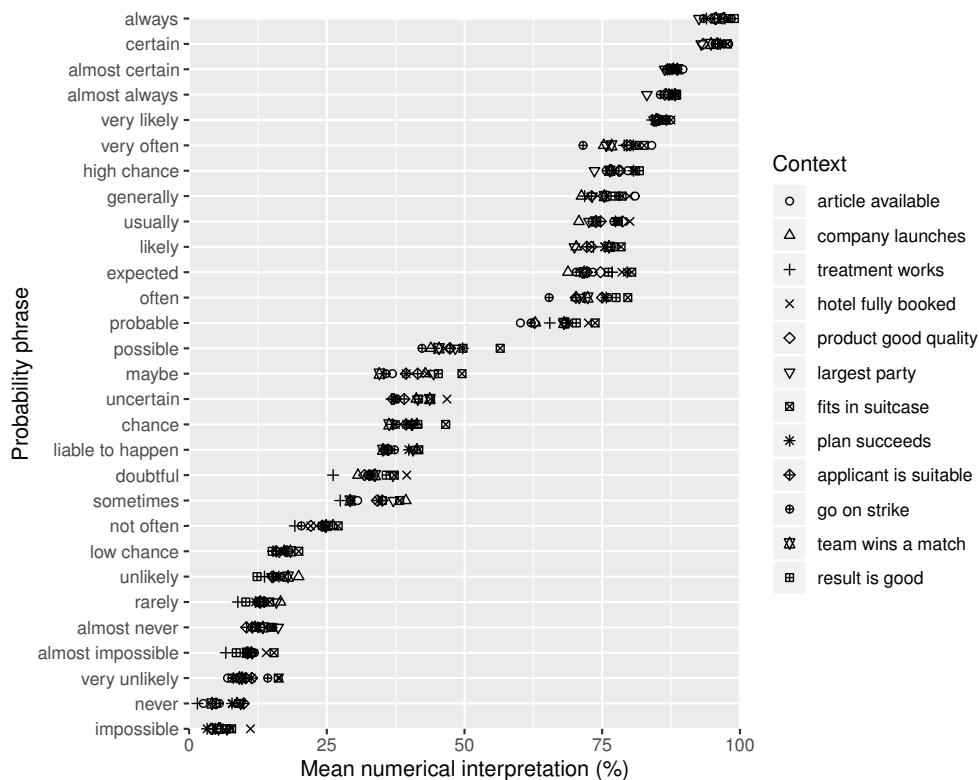
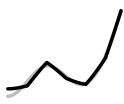


Figure 5.4: Means of the numerical interpretations (in percentages) given by all participants for each phrase in the surveys, grouped by the context of the sentences. Listed contexts in the legend are abbreviations of the originals (see Table 5.3 of the supplementary material (subsection 5.7.2)).

numerical interpretations. Results for all phrases are shown in Figure 5.6 of the supplementary material (subsection 5.7.3).

These density plots show that the interpretations of the probability phrases are very similar for both statisticians and non-statisticians. This similarity is represented by the overlapping regions of the plots. The nonoverlapping regions are relatively small, which suggests that there are no big differences between the groups. This is supported by the group means, since the maximum difference between statisticians and non-statisticians is four percentage points.

Although the differences are small, the density plots of *very likely* and *almost never* in Figure 5.5 may suggest that statisticians agree more on the interpretation of verbal probability expressions expressing an extreme probability. This phenomenon is also seen for other extreme phrases (see Figure 5.6 of the supplementary material, subsection 5.7.3), but not for phrases expressing



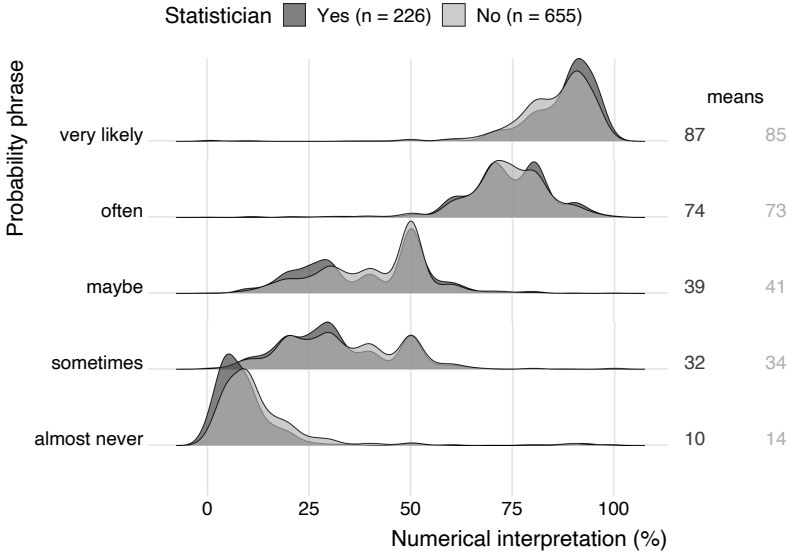


Figure 5.5: *Density plots and mean values of the numerical interpretations (in percentages) given by statisticians and non-statisticians for a selection of five phrases from the survey. Note that density plots are a smooth variant of histograms and may therefore be positive outside the data range of 0–100%.*

percentages closer to 50%. However, the difference between the group means is small for these phrases, so the group effect (if present) is weak.

We also investigated whether there were differences in responses by men and women but found no notable differences (see Figure 5.7 of the supplementary material, subsection 5.7.3).

Note that we do not statistically test for group differences because the interpretations have very irregular distributions which makes statistical testing complicated.

5.5 Discussion

In this study we have investigated the variability of the interpretation of Dutch probability and frequency phrases. The set-up of our survey was comparable to previous surveys on the interpretation of English phrases, but it filled some gaps in the research on Dutch probability phrases. For example, we included many Dutch expressions that were not studied before and represented them in a neutral context. Furthermore, we verified asymmetries in the interpretation of mirrored phrases, and checked for differences in interpretation between statisticians and non-statisticians.

Our results showed that, as in English, there is a large variability in the interpretation of Dutch probability and frequency phrases. Although there is some agreement about extreme words as *always*, *certain*, *never*, and *impossible*, there is no consensus about words that describe a less extreme probability.

As mentioned before, Eekhof et al. (1992) already studied many Dutch frequency expression. Ten of those were also included in our study so we could compare the results. For nine of these phrases, the mean interpretations differed a maximum of three percentage points. Only the interpretation of *sometimes* differed more, namely a difference of 8 percentage points (mean of 33% in our study vs. 25% in their study).

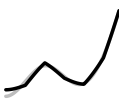
Besides comparing our results to those of previous studies in Dutch, we can verify our results with studies that included the English counterparts of the phrases in our survey. Theil (2002) listed the mean interpretations for ten probability phrases found in ten studies, seven of which overlapped with our list of verbal probability expressions. The mean interpretations that we measured for these phrases are all between the lower and upper bounds of the means measured in these ten studies. However, the ranges of those means were quite wide for some expressions. Due to this large amount of variability in the English results, it is not possible to conclude from this that there are no differences between the interpretations of Dutch phrases and their English translations.

Additionally, our data confirms the previous results on asymmetry in the interpretation of verbal probability expressions, also on an individual level. For example, usually an individual's numerical interpretations of *likely* and *unlikely* do not sum to 100%.

Previous studies in English showed that the asymmetry in the interpretation of some mirrored pairs (as *possible* and *impossible*, and *certain* and *uncertain*) has a linguistic cause. Our study confirms that similar asymmetries are found for the Dutch translations of these phrases.

Another phenomenon that has previously been shown to have an influence on the interpretations of verbal probability expressions is context. Therefore, we tried to present the expressions in neutral contexts. Although the mean interpretations varied among the contexts (see Figure 5.4), our results did not show a structural difference between contexts. Hence, there were probably no strong base-rate effects, indicating that our chosen contexts were neutral enough.

Only after analyzing our data we realized that of the twelve contexts that we used, ten presented a positive outcome (for example *this treatment will work* and *this plan succeeds*), but two presented a negative outcome (namely *they will go on strike* and *this hotel is fully booked*). Phrasing a risk positively or negatively may also influence interpretation. However, we found no structural differences between the mean interpretations of these positive and negative contexts.



To test for differences in interpretations between experts and laypeople this survey was distributed among both statisticians and non-statisticians. Our data showed large variability within each group and no structural differences between them. Hence, it seems that regularly making and communicating probability estimations does not increase agreement about the interpretations of probability expressions. This justifies our analyses on the complete sample.

The size of the complete sample (881 participants) is one of the strengths of this study. In most studies sample sizes were quite small; the mean number of participants in the previous Dutch studies listed in this paper was 93 and, for example, in the English studies mentioned by Theil (2002) it was 170.

Participants were invited via Twitter, which is a convenient way to reach a lot of people. Although this resulted in our large sample size, the convenience sample includes a disproportionate distribution of education levels and, hence, our results on the non-statisticians may not generalize to the population. However, the results from this study are still valuable, since they showed that, even within this homogeneous sample, interpretations of probability expressions differed enormously. This indicates that interpretations are dissimilar even among more like-minded persons. If the sample had been more heterogeneous, the interpretations may have varied even more.

5.6 Conclusion

From our results we conclude that the interpretations of Dutch verbal probability expressions are comparable to those of their English translations. Therefore, all challenges regarding communicating with English verbal probability expressions also apply in Dutch. For example, making a translation table from Dutch verbal probability phrases to numeric values is infeasible.

Although it was generally believed that information providers prefer to express probabilities verbally, a solution might be to convey estimated risks using either numerical values instead of verbal expressions, or both. This may prevent the intended probability from getting lost in its translation from one language to another.

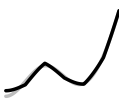
Recently, studies have been done on this topic. For example, Jenkins et al. (2018) studied whether the order (either verbal - numeric or numeric - verbal) may influence the interpretation of a verbal probability expression, and Wintle et al. (2019) studied four different methods of presenting numeric probabilities along with a verbal probability expression. Since research on this topic is not finished, we would advise to keep an eye open for them. Hopefully an optimal mode of presentation of estimated probabilities will be found that minimizes the risks of miscommunication.

5.7 Supplementary material

5.7.1 Translations of probability phrases

Dutch phrase	English translation
onmogelijk	impossible
nooit	never
zeer onwaarschijnlijk	very unlikely, very improbable
bijna onmogelijk	almost impossible
bijna nooit	almost never
zelden	rarely, seldom
onwaarschijnlijk	unlikely, improbable
kleine kans	low chance
niet vaak	not often
soms	sometimes, once in a while
twijfelachtig	doubtful
kan gebeuren	liable to happen
kans	chance
onzeker	uncertain
misschien	maybe
mogelijk	possible
vermoedelijk	probable
vaak	often
te verwachten	expected
waarschijnlijk	likely, probably
meestal	usually
doorgaans	generally, usually
grote kans	high chance
heel vaak	very often
zeer waarschijnlijk	very likely, very probable
bijna altijd	almost always
bijna zeker	almost certain
zeker	certain
altijd	always

Table 5.2: *The 29 Dutch frequency and probability phrases used in the survey, with their English translations used in this paper. The phrases are presented in the same order as in Figure 5.2 in this paper.*



5.7.2 Context sentences

Dutch context sentences and their English translations (italic)
Het is waarschijnlijk dat alles in de koffer past. <i>It is likely that everything fits in the suitcase.</i>
Het is waarschijnlijk dat het team een wedstrijd wint. <i>It is likely that the team wins a match.</i>
Het is waarschijnlijk dat deze behandeling aanslaat. <i>It is likely that this treatment will work.</i>
Het is waarschijnlijk dat een sollicitant geschikt is voor de baan. <i>It is likely that this applicant is suitable for the job.</i>
Het is waarschijnlijk dat bedrijf A het product eerder lanceert dan bedrijf B. <i>It is likely that company A launches the product before company B does.</i>
Het is waarschijnlijk dat zij gaan staken. <i>It is likely that they will go on strike.</i>
Het is waarschijnlijk dat de uitslag goed is. <i>It is likely that the result is good.</i>
Het is waarschijnlijk dat deze partij de grootste wordt bij de verkiezingen. <i>It is likely that this party will be the largest in the elections.</i>
Het is waarschijnlijk dat dit plan slaagt. <i>It is likely that this plan succeeds.</i>
Het is waarschijnlijk dat deze producten van goede kwaliteit zijn. <i>It is likely that these products are of good quality.</i>
Het is waarschijnlijk dat dit hotel is volgeboekt. <i>It is likely that this hotel is fully booked.</i>

Table 5.3: The 12 Dutch context sentences used in the survey, with their English translations used in this paper.

5.7.3 Differences between sub-populations

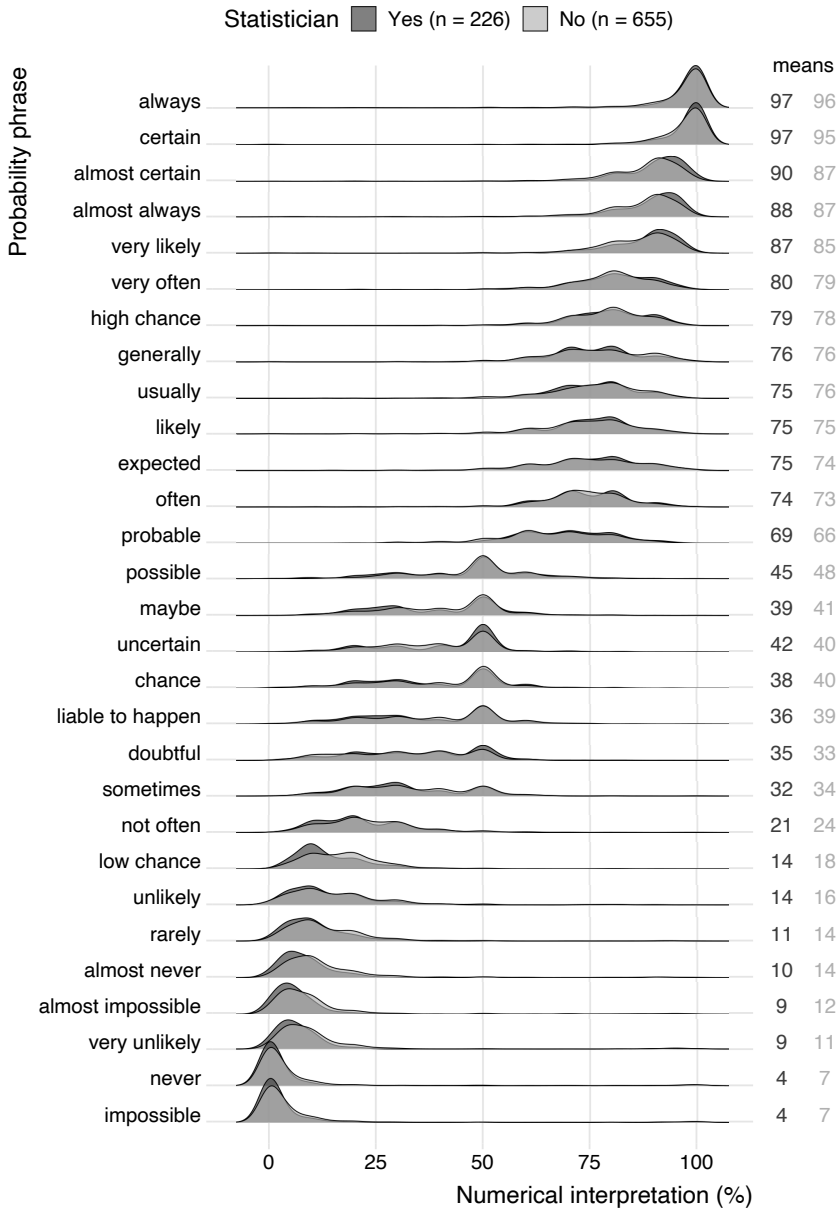


Figure 5.6: Density plots and mean values of the numerical interpretations (in percentages) given by statisticians and non-statisticians for each phrase in the survey. Note that density plots are a smooth variant of histograms and may therefore be positive outside the data range of 0–100%.

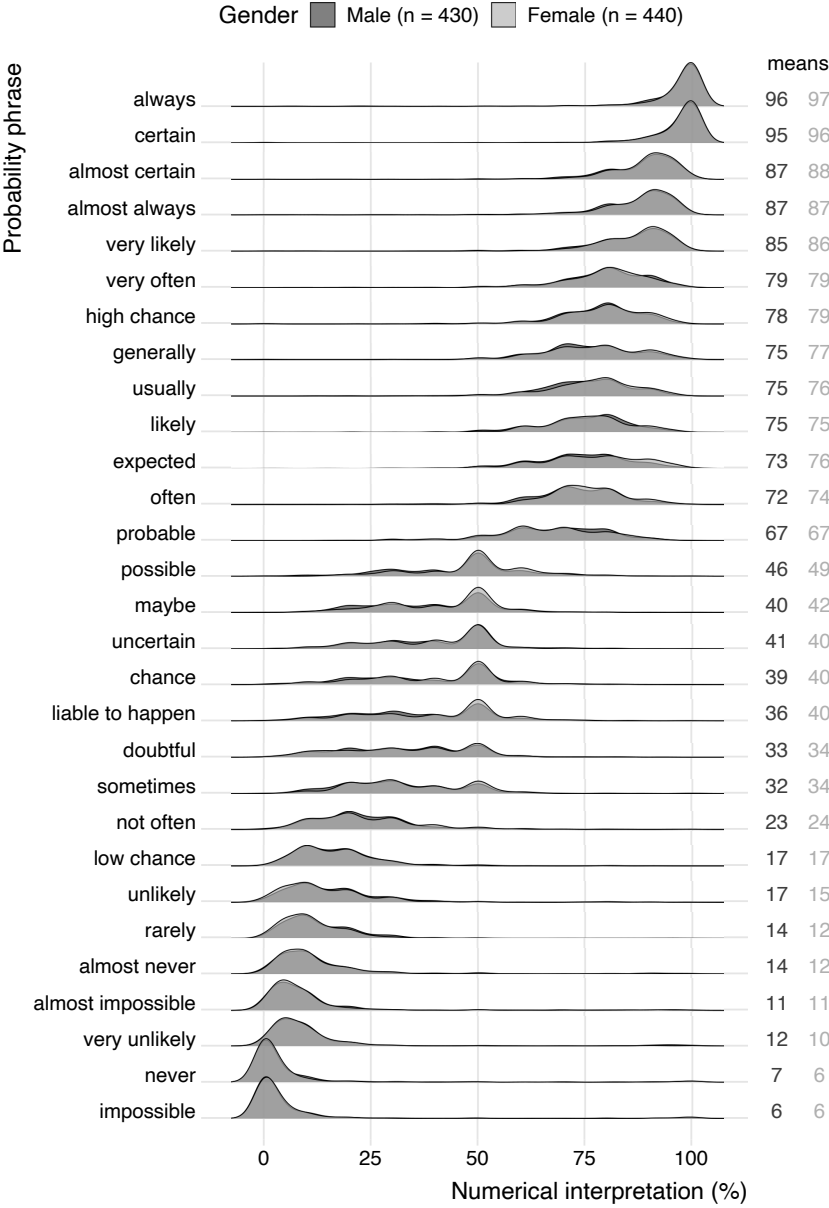


Figure 5.7: *Density plots and mean values of the numerical interpretations (in percentages) given by males and females for each phrase in the survey. Category level Other/Prefer not to say was omitted, because there were only 11 observations in this group. Note that density plots are a smooth variant of histograms and may therefore be positive outside the data range of 0–100%.*

BIBLIOGRAPHY

- Ajzen, I. *The theory of planned behaviour*. Organisational behaviour and human decision processes, 50:179–211, 1991. – Cited on page 55.
- Andersen, P. K. and Gill, R. D. *Cox's Regression Model for Counting Processes: A Large Sample Study*. The Annals of Statistics, 10(4):1100–1120, 1982. – Cited on page 33.
- Bender, R., Augustin, T., and Blettner, M. *Generating survival times to simulate Cox proportional hazards models*. Statistics in Medicine, 24(11):1713–1723, 2005. – Cited on page 38.
- Boršič, D. and Kavkler, A. *Modeling unemployment duration in Slovenia using Cox regression models*. Transition Studies Review, 16(1):145–156, 2009. – Cited on page 45.
- Boyle, C. M. *Difference Between Patients' and Doctors' Interpretation of Some Common Medical Terms*. British Medical Journal, 2(5704):286–289, 1970. – Cited on page 140.
- Breslow, N. E. *Contribution to the Discussion of Paper by D.R. Cox*. Journal of the Royal Statistical Society. Series B (Methodological), 34:216–217, 1972. – Cited on pages 81 and 99.
- Brouwer, S., Bakker, R. H., and Schellekens, J. M. H. *Predictors for re-employment success in newly unemployed: A prospective cohort study*. Journal of Vocational Behaviour, 89:32–38, 2015. – Cited on pages 46, 47, 51, and 53.
- Bryant, G. D. and Norman, G. R. *Expressions of probability: words and numbers*. New England Journal of Medicine, 302(7):411, 1980. – Cited on pages 136 and 140.
- Budescu, D. V. and Wallsten, T. S. *Consistency in interpretation of probabilistic phrases*. Organizational Behavior and Human Decision Processes, 36(3):391–405, 1985. – Cited on page 136.

- Castro, C. M., Wilson, C., Wang, F., and Schillinger, D. *Babel babble: Physicians' use of unclarified medical jargon with patients*. American Journal of Health Behavior, 31(Suppl 1):S85–95, 2007. – Cited on page 140.
- Chang, I. M., Gelman, R., and Pagano, M. *Corrected group prognostic curves and summary statistics*. Journal of Chronic Diseases, 35(8):669 – 674, 1982. ISSN 0021-9681. – Cited on page 30.
- Cox, D. R. *Regression Models and Life Tables (with Discussion)*. Journal of the Royal Statistical Society B, 34:187–220, 1972. – Cited on pages 8 and 53.
- Davidson, R. A. and Chrisman, H. H. *Translations of uncertainty expressions in Canadian accounting and auditing standards*. Journal of International Accounting, Auditing and Taxation, 3(2):187 – 203, 1994. – Cited on page 137.
- Doupnik, T. S. and Richter, M. *Interpretation of uncertainty expressions: a cross-national study*. Accounting, Organizations and Society, 28(1):15 – 35, 2003. – Cited on page 137.
- Druzdzel, M. J. *Verbal Uncertainty Expressions: Literature Review*. Technical Report, 1989. – Cited on pages 10, 136, 138, and 140.
- Eekhof, J. A. H., Mol, S. S. L., and Pielage, J. C. *Is doorgaans vaker dan dikwijls; of hoe vaak is soms?* Nederlands Tijdschrift voor Geneeskunde, 136(1):41, 1992. – Cited on pages 137, 141, 142, and 151.
- Erev, I. and Cohen, B. L. *Verbal versus Numerical Probabilities: Efficiency, Biases, and the Preference Paradox*. Organizational Behavior and Human Decision Processes, 45:1–18, 1990. – Cited on page 138.
- European Food Safety Authority, Hart, A., Maxim, L., Siegrist, M., da Cruz, C., Merten, C., Lahaniatis, M., and Smith, A. *Outcome of the Public Consultation on the draft Guidance on Communication of Uncertainty in Scientific Assessments*. EFSA Journal, 17(1):1–73, 2019. – Cited on page 136.
- Fox, J. and Carvalho, M. S. *The RcmdrPlugin.survival Package: Extending the R Commander Interface to Survival Analysis*. Journal of Statistical Software, 49(7):1–32, 2012. – Cited on page 34.
- Friedman, J. H. and Stuetzle, W. *Projection Pursuit Regression*. Journal of the American Statistical Association, 76(376):817–823, 1981. – Cited on pages 102 and 105.
- Gifi, A. *Nonlinear Multivariate Analysis*. John Wiley & Sons, Ltd., 1990. Original work published 1981. – Cited on pages 9, 69, 102, and 105.

- Google. *Google Translate*. <https://translate.google.com>, June 2018. – Cited on page 141.
- Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. *Assessment and comparison of prognostic classification schemes for survival data*. *Statistics in Medicine*, 18:2529 – 2545, 1999. – Cited on page 65.
- Green, P. J. *Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives*. *Journal of the Royal Statistical Society: Series B*, 46(2):149–192, 1984. – Cited on page 99.
- Harrell, F. E., Lee, K. L., and Mark, D. B. *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors*. *Statistics in Medicine*, 15:361–387, 1996. – Cited on page 65.
- Harris, A. J. L., Corner, A., Xu, J., and Du, X. *Lost in translation? Interpretations of the probability phrases used by the Intergovernmental Panel on Climate Change in China and the UK*. *Climatic Change*, 121(2):415–425, 2013. – Cited on page 137.
- Hartmann, A., Van der Kooij, A. J., and Zeeck, A. *Exploring nonlinear relations: Models of clinical decision making by regression with optimal scaling*. *Psychotherapy Research*, 19(4-5):482–492, 2009. – Cited on page 115.
- Hastie, T. J. and Tibshirani, R. J. *Generalized Additive Models*. *Monographs on Statistics & Applied Probability*. Chapman & Hall/CRC, 1990. – Cited on pages 102, 105, and 131.
- Jenkins, S. C., Harris, A. J. L., and Lark, R. M. *Understanding ‘Unlikely (20% Likelihood)’ or ‘20% Likelihood (Unlikely)’ Outcomes: The Robustness of the Extremity Effect*. *Journal of Behavioural Decision Making*, 31, 2018. – Cited on page 152.
- Jolliffe, I. T. *Principal component analysis*. New York, NY: Springer-Verlag, 2002. – Cited on page 52.
- Juanchich, M. and Sirota, M. *Do people really prefer verbal probabilities?* *Psychological Research*, Jun 2019. – Cited on page 138.
- Kaplan, E. L. and Meier, P. *Nonparametric Estimation from Incomplete Observations*. *Journal of the American Statistical Association*, 53:457–481, 1958. – Cited on pages 15 and 35.

- Kavkler, A., Dănaćică, D. E., Babucea, A. G., Bićanić, I., Böhm, B., Tevdovski, D., Toševska, K., and Boršič, D. *Cox regression models for unemployment duration in Romania, Austria, Slovenia, Croatia, and Macedonia*. Romanian Journal of Economic Forecasting, 10(2):81–104, 2009. – Cited on page 45.
- Kim, Y., Kim, J., and Kim, Y. *Blockwise Sparse Regression*. Statistica Sinica, 16(2):375–390, 2006. – Cited on page 132.
- Klein, J. P. and Moeschberger, M. L. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer, second edition edition, 2003. – Cited on pages 53 and 71.
- Kleinbaum, D. G. and Klein, M. *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health. Springer, third edition edition, 2012. – Cited on page 16.
- Kruskal, J. B. *Nonmetric multidimensional scaling: a numerical method*. Psychometrika, 29:115–129, 1964. – Cited on pages 74, 78, 103, and 106.
- Lichtenstein, S. and Newman, J. R. *Empirical scaling of common verbal phrases associated with numerical probabilities*. Psychonomic Science, 9(10):563–564, 1967. – Cited on pages 10, 138, and 139.
- Lim, T. S., Loh, W. Y., and Shih, Y. S. *A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms*. Machine Learning, 40(3):203–228, Sep 2000. – Cited on page 120.
- Linting, M. and van der Kooij, A. *Nonlinear Principal Components Analysis With CATPCA: A Tutorial*. Journal of Personality Assessment, 94(1):12–25, 2012. – Cited on pages 52, 54, and 55.
- Linting, M., Meulman, J. J., Groenen, P. J., and van der Kooij, A. J. *Nonlinear principal components analysis: introduction and application*. Psychological Methods, 12(3):336–358, September 2007. – Cited on pages 44, 52, 70, and 95.
- MacLeod, A. and Pietravallo, S. *Communicating risk: variability of interpreting qualitative terms*. EPPO Bulletin, 47(1):57–68, 2017. – Cited on page 140.
- McCullagh, P. and Nelder, J. A. *Generalized Linear Models*. Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, 1989. – Cited on pages 9 and 102.
- Meulman, J. J., Van der Kooij, A. J., and Heiser, W. J. *Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data*. In Kaplan, D., editor, *Handbook of Quantitative Methodology for the*

- Social Sciences.*, pages 49–70. Sage, London, 2004. – Cited on pages 10, 44, 52, 70, and 95.
- Meulman, J. J., Van der Kooij, A. J., and Duisters, K. L. W. *ROS Regression: Integrating Regularization with Optimal Scaling Regression*. Statistical Science, 34:361–390, 2019. – Cited on pages 9, 70, 75, 104, and 132.
- Montgomery, S. A. and Åsberg, M. *A new depression scale designed to be sensitive to change*. The British Journal of Psychiatry, 134(4):382–9, 1979. – Cited on page 22.
- Mooney, C. Z. *Monte Carlo Simulation (Quantitative Applications in the Social Sciences)*. SAGE Publications, April 1997. ISBN 0803959435. – Cited on page 38.
- Mosteller, F. and Youtz, C. *Quantifying Probabilistic Expressions*. Statistical Science, 5(1):2–12, 1990. – Cited on pages 139, 140, and 146.
- Pander Maat, H. and Klaassen, R. *De invloed van de formulering van informatie over bijwerkingen op interpretaties door patiënten en op gemelde bijwerkingen*. Nederlands Tijdschrift voor Geneeskunde, 140(8):424–427, 1996. – Cited on pages 137, 139, 140, and 141.
- Putter, H., Fiocco, M., and Geskus, R. B. *Tutorial in biostatistics: competing risks and multi-state models*. Statistics in Medicine, 26(11):2389–2430, 2007. – Cited on page 53.
- Qualtrics. *Qualtrics*. <https://www.qualtrics.com>, 2005-2018. – Cited on page 143.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL <https://www.R-project.org/>. – Cited on pages 11, 18, and 114.
- Ramsay, J. O. *Monotone Regression Splines in Action*. Statistical Science, 3(4): 425–442, November 1988. – Cited on pages 75, 103, and 106.
- Reagan, R. T., Mosteller, F., and Youtz, C. *Quantitative meanings of verbal probability expressions*. Journal of Applied Psychology, 74(3):433–442, 1989. – Cited on pages 10, 137, 138, and 139.
- Renooij, S. and Witteman, C. *Talking probabilities: communicating probabilistic information with words and numbers*. International Journal of Approximate Reasoning, 22(3):169–194, 1999. – Cited on pages 137 and 141.

- Robins, J. M. *Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers*. Proceedings of the Biopharmaceutical Section, American Statistical Association, pages 24–33, 1993. – Cited on pages 14 and 16.
- Robins, J. M. and Finkelstein, D. M. *Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests*. Biometrics, 56:779–788, 2000. – Cited on pages 14, 16, and 18.
- Robins, J. M. and Rotnitzky, A. *Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers*. In Jewell, N. P., Dietz, K., and Farewell, V. T., editors, *AIDS Epidemiology*, pages 297–331. Birkhäuser Boston, 1992. ISBN 978-1-4757-1231-5. – Cited on pages 14 and 16.
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. *Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent*. Journal of Statistical Software, 39(5):1–13, 2011. – Cited on pages 80, 81, 97, 99, and 109.
- SPSS Inc. *SPSS Statistics for Windows, Version 17.0*. Chicago, 2008. – Cited on page 52.
- Stheeman, S. E., Mileman, P. A., van ’t Hof, M. A., and van der Stelt, P. F. *Blind chance? An investigation into the perceived probabilities of phrases used in oral radiology for expressing chance*. Dento maxillo facial radiology, 22:135–139, 1993. – Cited on pages 10, 139, and 140.
- Takane, Y., Young, F. W., and de Leeuw, J. *The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features*. Psychometrika, 49:279–281, 1978. – Cited on page 10.
- Theil, M. *The role of translations of verbal into numerical probability expressions in risk management: a meta-analysis*. Journal of Risk Research, 5(2):177–186, 2002. – Cited on pages 136, 140, 141, 151, and 152.
- Therneau, T. M. *A Package for Survival Analysis in S*, 2013. URL <http://CRAN.R-project.org/package=survival>. R package version 2.37-4. – Cited on pages 14 and 18.
- Therneau, T. M. and Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. – Cited on pages 14 and 18.
- Timmermans, D. *The Roles of Experience and Domain of Expertise in Using Numerical and Verbal Probability Terms in Medical Decisions*. Medical Decision Making, 14:146–156, 1994. – Cited on pages 137, 141, and 142.

- Timmermans, D. R. M. and Mileman, P. A. *Lost for words: using verbal terms to express probabilities in oral radiology*. Dento maxillo facial radiology, 22(4): 171–172, 1993. – Cited on page 139.
- Tutkun, N. A. and Karasoy, D. *Applying frailty models to analyse the duration of unemployment in Turkey*. International Journal of Statistics & Economics, 17(2):44–57, 2016. – Cited on page 45.
- Twitter Inc. *Twitter*. <https://twitter.com>, 2018. – Cited on page 144.
- Van Dale Uitgevers. *Van Dale*. <https://www.vandale.nl>, June 2018. – Cited on page 141.
- Van der Kooij, A. J. *Prediction Accuracy and Stability of Regression with Optimal Scaling Transformations*. PhD thesis, Leiden University, 2007. – Cited on page 104.
- Van der Kooij, A. J. and Meulman, J. J. *Regression with Optimal Scaling*. In Meulman, J., Heiser, W., and Inc., S., editors, *SPSS Categories 10.0.*, pages 1–8, 77–101, 239–246. Chicago: SPSS Inc, 1999. – Cited on pages 9, 52, and 102.
- van Swieten, J. C., Koudstaal, P. J., Visser, M. C., Schouten, H. J., and van Gijn, J. *Interobserver agreement for the assessment of handicap in stroke patients*. Stroke, 19:604–607, 1988. – Cited on page 68.
- Venhuizen, G. J., Hut, R., Albers, C., Stoof, C. R., and Smeets, I. *Flooded by jargon: how the interpretation of water-related terms differs between hydrology experts and the general audience*. Hydrology and Earth System Sciences, 2019. Accepted for publication. – Cited on page 140.
- Visschers, V. H. M., Meertens, R. M., Passchier, W. W. F., and de Vries, N. N. K. *Probability Information in Risk Communication: A Review of the Research Literature*. Risk Analysis, 29(2):267–287, 2009. – Cited on page 136.
- Vroom, V. H. *Work and motivation*. John Wiley & Sons, New York, 1964. – Cited on page 55.
- Wallsten, T. S., Fillenbaum, S., and Cox, J. A. *Base rate effects on the interpretations of probability and frequency expressions*. Journal of Memory and Language, 25(5):571–587, 1986. – Cited on pages 11 and 140.
- Wanberg, C. R., Hough, L. M., and Song, Z. *Predictive validity of a multidisciplinary model of re-employment success*. Journal of Applied Psychology, 85: 1100–1120, 2002. – Cited on page 45.

- Weber, E. U. and Hilton, D. J. *Contextual Effects in the Interpretations of Probability Words: Perceived Base Rate and Severity of Events*. Journal of Experimental Psychology: Human Perception and Performance, 16(4):781–789, 1990. – Cited on pages 139 and 140.
- Wijnhoven, M. A. and Havinga, H. *The Work Profiler: A digital instrument for selection and diagnosis of the unemployed*. Local Economy, 29(6-7):740–749, 2014. – Cited on page 46.
- Willems, S. J. W., Fiocco, M., and Meulman, J. J. *Optimal scaling for survival analysis with ordinal data*. Computational Statistics & Data Analysis, 115:155 – 171, 2017. – Cited on page 109.
- Winsberg, S. and Ramsay, J. O. *Monotonic transformations to additivity using splines*. Biometrika, 67(3):669–674, 1980. ISSN 0006-3444. – Cited on page 102.
- Wintle, B. C., Fraser, H., Wills, B. C., Nicholson, A. E., and Fidler, F. *Verbal probabilities: Very likely to be somewhat more confusing than numbers*. PLoS ONE, 14(4):1–18, 04 2019. – Cited on page 152.
- Yang, Y. and Zou, H. *A cocktail algorithm for solving the elastic net penalized Cox’s regression in high dimensions*. Statistics and Its Interface, 6:167–173, 2013. – Cited on page 85.
- Young, F. W. *Quantitative analysis of qualitative data*. Psychometrika, 46(4): 357–388, 1981. – Cited on page 102.
- Young, F. W., de Leeuw, J., and Takane, Y. *Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features*. Psychometrika, 41:505–528, December 1976. – Cited on pages 9, 102, and 105.
- Yuan, M. and Lin, Y. *Model Selection and Estimation in Regression With Grouped Variables*. Journal of the Royal Statistical Society: Series B, 68:49–67, 02 2006. – Cited on page 132.
- Zeeck, A., Hartmann, A., Küchenhoff, J., Weiss, H., Sammet, I., Gaus, E., Semm, E., Harms, D., Eisenberg, A., Rahm, R., and von Wietersheim, J. *Differenzielle Indikationsstellung stationärer und tagesklinischer Psychotherapie: die DINSTAP-Studie[1]*. Psychotherapie, Psychosomatik, Medizinische Psychologie, 59(9/10):354–363, 2009. – Cited on page 115.

INDEX

- (G)AM: (Generalized) Additive Model
 - application to breast cancer recurrence data, 128–130
 - application to contraceptive method choice data, 124–127
 - Method, 104, 107
- Censoring, 14
 - dependent, 14–16
 - correct for, *see* IPCW
 - informative, 14
 - noninformative, 14
 - time, 15, 53, 71, 80
- Cox' proportional hazards model
 - for censoring mechanism, 16, 18, 69
 - for survival mechanism, 71, 81, 86
 - IRLS estimation, 80–82
 - method, 53
 - with optimal scaling transformations, *see* Cox-OS
 - with ties, 81
 - with weighted objects, 81
- Cox-OS: Cox model with optimal scaling transformations
 - application to simulated data, 87–92
 - method, 70, 82–85
- Data types
 - Categorical data
 - nominal, 68–69
 - ordinal, 68–69
 - Numeric/continuous data, 68, 69
 - see* Scaling Levels, 1
- Datasets
 - breast cancer recurrence data, 127
 - contraceptive method choice data, 122
 - DINSTAP data, 117
 - ROM data, 22–24
 - simulated survival data with an ordinal predictor, 86–87
 - simulated survival data with dependent censoring, 26, 38–40
 - toy survival data, 18
 - UWV data, 46–48
- Dummy coding, 71–72, 117
- Factor analysis
 - application to UWV data, 47, 51
 - method, 44, 62–64
- GLM-OS: GLM with optimal scaling transformations
 - application to breast cancer recurrence data, 127–130
 - application to contraceptive method choice data, 122–127
 - application to DINSTAP data, 116–122
 - logistic regression, 115–116
 - method, 106, 109–113
- GLM: Generalized Linear Model
 - method, 104, 110–111

- Hazard
 - rate, 15–16, 18, 26, 71
 - ratio, 53–54, 71
 - method, 45, 64–65, 135–136
 - OS, *see* GLM-OS, logistic regression
- Integer coding, 72–73
- Interpretation of verbal probability phrases
 - (a)symmetry, 138–141, 148, 153
 - communication mode preference paradox, 140
 - context dependence, 141–142, 144, 150, 153–154
 - difference between sub-populations, 142–143, 150–152, 154, 158–159
 - other languages, 139–140, 153
 - probability scales, 138–139, 141
 - variability, 138, 147–148, 153
- IPCW: Inverse Probability Censoring
 - Weighting
 - application to ROM data, 23–25
 - application to simulated data, 26–28
 - application to toy data, 18–21, 33, 36, 37
 - assumption, 17–18
 - implementation in R, 32–37
 - method, 14, 16–21
 - weights, 16–17, 20
- Kaplan-Meier Estimator, *see* Product-Limit Estimator
- Linear Regression
 - method, 75, 104
 - with OS, *see* Optimal scaling regression
- Logistic regression
 - application to breast cancer recurrence data, 128–130
 - application to contraceptive method choice data, 124–127
 - application to UWV data, 46, 51, 53–54
- Optimal scaling
 - Cox, *see* Cox-OS
 - GLM, *see* GLM-OS
 - logistic regression, *see* GLM-OS, logistic regression
 - PCA, *see* OS-PCA
 - quantifications, 45, 69, 73, 75–76, 105
 - regression, 69–70, 74–80, 105–109
- OS-PCA: Optimal scaling principal components analysis
 - application to UWV data, 52–56, 58–59
 - method, 44–45, 52, 62–64, 70
- PCA: Principal Components Analysis
 - method, 52
 - nonlinear, *see* OS-PCA
- Product-Limit Estimator
 - for censoring mechanism, 19
 - method, 15, 20
- Scaling levels
 - monotone spline function, 74–75, 105, 108, 112
 - nominal step function, 69, 74, 78, 105, 108, 112, 117
 - nonmonotone spline function, 105, 108, 112
 - numeric, 105
 - ordinal step function, 74–75, 78, 105, 108, 112
- Survival analysis, 45, 53, 64–65
 - application to UWV data, 54, 56–59
 - censoring, *see* Censoring
 - competing risks, 53
 - hazard rate/ratio, *see* Hazard

SUMMARY

This thesis consists of five scientific papers on several topics in statistics, namely survival analysis, optimal scaling transformations, and statistics communication or combinations of these topics. Each research project was motivated by either a specific dataset or an interest in a particular aspect of data.

Chapter 1

The research project described in the first chapter was motivated by a question from the Department of Psychiatry of Leiden University Medical Center. Doctors collected data on patients with mood or anxiety disorders to study the course of their suicidal ideation. In particular they wanted to estimate the time between the diagnosis of a patient's condition and the moment of remission of suicidal ideation. The goal of estimating this recovery time was to determine which patients are at risk of persistent suicidal thoughts.

Statistical methods used to answer this type of questions are called *survival analysis* methods. These methods are used to model the time between a time origin and a specific event of interest. In the setting of estimating recovery time the time origin is the moment of diagnosis and the event of interest is remission. Besides recovery time, survival analysis methods can be used to model any time interval, for example the survival time of patients (time from diagnosis until death) or unemployment duration (time between losing a job and starting a new one).

In order to determine the recovery time of patients, these patients need to be monitored over a long period of time, namely until the patient's remission is observed. Since it takes time to collect the data there is a high risk that the event times will not be registered for some persons. For example, if a patient moves from one city to another and therefore changes hospital during treatment, researchers at the initial hospital will not observe the patient's recovery time. This phenomenon is called *censoring* because the event time of a person is unknown and is thus *censored*. Although the data on a censored person are incomplete, the available information is still valuable. For example, of the patient who moved it is known that he or she had not yet recovered at the time of the

last hospital visit. This indicates that the recovery might occur *after* this visit.

Since censoring is common when collecting data for time estimations, methods have been designed to also include this incomplete but valuable information in the survival analysis. To keep these analysis methods simple, they are designed under the assumption that censoring occurs randomly. This means that everyone should have the same probability of being censored. Therefore this probability should not depend on the person's characteristics or event time. For example, if a patient changes hospital because he or she moved to a new city, then it can be assumed that this could have happened to any of the patients. In other words, the cause of censoring is *independent* of the patient's health. This independence is important for a correct analysis.

However, the doctors who were estimating the recovery time suspected that especially those patients who started to feel better stopped visiting the clinic. Hence, the probability of censoring was suspected to be related to the patient's health status and therefore not equal for all patients. Hence, the independent censoring assumption may be incorrect and therefore the standard survival methods may give inaccurate results. Namely, the data would be biased because it contains more information on patients with long recovery times and less information on those with short recovery. As a result the recovery time may be overestimated. To avoid bias in the result due to dependent censoring special corrections can be applied in the survival analysis.

This first chapter focuses on one specific correction method called the *Inverse Probability Censoring Weighting Estimator*. This method compensates for the loss of information if an observation is censored by giving extra weight in the analysis to the individuals who most closely resemble the censored person and who are still under observation.

As part of this thesis, a user-friendly implementation of this method was developed for the frequently used statistical software program named *R*. This implementation makes the method more easily accessible for researchers who have less experience in programming.

The method was then applied to the dataset on patients with mood and anxiety disorders to correct for the suspected dependent censoring. The results showed that without the correction the recovery time was indeed overestimated, but the difference was very small. This suggests that the dependent censoring was either not problematic or that important data on patient characteristics are missing which prevented a correct distribution of weights among the remaining patients.

To investigate the performance of this method in the presence of dependent censoring, a simulation study was performed. This simulation study showed that the correction method can indeed correct for dependent censoring. However,

as is to be expected, the method performs best when the probability of being censored is estimated correctly. Thus, to get accurate estimates for the recovery time, data on many persons should be collected and all characteristics associated to censoring should be registered.

Chapter 2

The second chapter focusses on an application of survival and optimal scaling techniques to identify predictors for unemployment duration. The dataset comes from the Dutch Employee Insurance Agency (UWV). In the Netherlands, job seekers who recently became unemployed can apply for unemployment benefits and counseling at the UWV. Since the budget is limited, the counseling service expenses must be reduced. One of the strategies is to reduce face-to-face counseling. Therefore, UWV has developed an online survey to select individuals who will experience difficulties in finding a job and could therefore benefit from counseling. When face-to-face counseling is only given to the people who really need it, this will greatly reduce UWV's costs.

In the online survey UWV collects data on the unemployed on 17 different factors, including *job search behaviour*, *job search attitude*, *job acceptance readiness*, and *self-efficacy w.r.t. job interview preparation*. Each factor is measured by 1–5 items on the questionnaire. For example the last mentioned factor is measured by the three following statements.

- *I am well able to find information about vacancies.*
- *I can explain my strengths and weaknesses for a job well.*
- *I can write a good (digital) application.*

Job seekers are asked to indicate to what extent they agree with each of the statements and the provided answer options are *strongly disagree*, *disagree*, *don't disagree/agree*, *agree*, or *strongly agree*.

The given answers are usually closely related to each other. For example, if a person is able to independently search for a job and send applications, then his or her answer to the three statements listed above will probably be *agree* or *strongly agree*, and vice versa. In other words, the answers to the questions within one factor will probably be very similar. Some factors are also closely related. For example if a person has a positive attitude towards searching for a job, then he or she will probably also actively search for a job. As a result, this person will agree with most of the statements about both the factor *job search attitude* and the factor *job search behaviour*.

For statistical analyses it is better if strongly related variables are not included individually in a model but are first summarized into a small number of summary scores. A very simple way to provide the scores is by taking the (weighted)

average of the questions related to the same factor. The weighted averaging approach was used to summarize each factor and *factor analysis* was applied to calculate the optimal weights for each question. Furthermore, since it is impossible to do calculations with category names as *strongly agree*, the five category options were replaced by the integers 1–5 such that these values could then be used in the calculations.

Next, the summary scores and some additional available information including age, gender, and education, were used to model the probability that an unemployed person would find a new job. For this analysis *logistic regression* was used which can classify the job seekers into one of two groups. The first group contains people who will probably find a job within one year and the second group are those who will probably not find a job. Besides the unemployment status after one year, the UWV was interested in this status at 6 and 9 months. Therefore three analyses were performed to estimate the re-employment probability at 6, 9, and 12 months.

Although this analysis already gives good insights in the factors influencing the probability of finding a new job, there are alternative methods to prepare and analyze the data. Alternative methods for both steps are proposed in the second chapter of this thesis and applied to UWV's dataset.

For data preparation the method of *optimal scaling principal components analysis* (OSPCA) is proposed. As factor analysis, this method aims to summarize many variables into a smaller set of summary scores. However, in OSPCA these scores are not calculated for each factor individually but are weighted averages over *all* the original variables. The weights are chosen in such a way that the summary scores together explain most of the variation between the persons in the dataset.

The main benefit of this method is that it also finds optimal numeric values to represent the category levels before calculating the weighted average. For example, it may conclude that 1.5, 2, 3, 4, and 4.5 are better replacements for the categories instead of 1, 2, 3, 4, and 5 because the step from having no opinion (*don't disagree/agree*) to (dis)agreeing is larger than the step between (dis)agreeing and strongly (dis)agreeing. The step to the more extreme opinion is smaller and is therefore represented by a smaller step size in this example. In this way the replacement values preserve the property that the category levels are ordered, but does not force equal distances between the consecutive categories as was done by the replacement with integer values.

Furthermore, OSPCA can give insight in the relations between the factors. Namely, this method will summarize the UWV data via weighted averages over all the statements in the questionnaire instead of doing so for one factor at the time. As a result these summary scores also give insight in the relation between

the 17 factors. For example, it can be used to verify the expected relation between job search attitude and job search behaviour. This ability to study relations between factors is another benefit of OSPCA. However, it can also be a disadvantage since the summary scores are more complicated to interpret when they are calculated over all variables instead of per factor.

The proposed alternative method for the final analysis is survival analysis. This method (as described in the first chapter) can estimate the probability of remaining unemployed over an interval of time as opposed to at a few prespecified time points such as 6, 9, and 12 months. A *survival curve* can be drawn to visualize this probability over the time interval. An example of a survival curve is shown in Figure 1; it shows that about 60% of the people found a job within 12 months.

In this chapter the differences between the two combinations of methods are described and for several situations it is discussed which combination is most suitable.

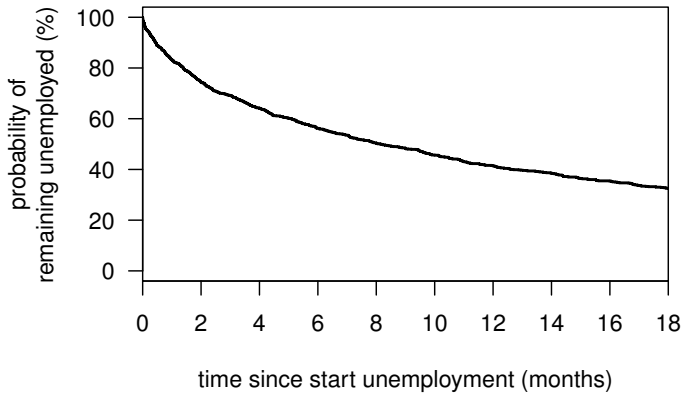


Figure 1: Example of a survival curve which shows the estimated probability of remaining unemployed until a certain time point. For example, this probability is 40% at 12 months which indicates that 60% of the people found a job within this time.

Chapter 3

The first two chapters of this thesis were motivated by questions from researchers and the provided datasets contain several different types of variables. Some variables, such as age, are *numeric* variables measured on a continuous scale. Other variables are *categorical* and indicate a category level. This type of data can be further classified to either *nominal* categorical data for which the category

levels are not ordered, such as gender, or *ordinal* categorical data for which the categories do have an ordering, such as education level or a person's level of agreement with a statement (*strongly disagree* – *disagree* – *don't disagree/agree* – *agree* – *strongly agree*).

Before any calculations can be done on categorical data, they first need to be transformed into numeric values. In other words, the category levels need to be *quantified*. The resulting numeric values are called *quantifications*. For example, in chapter 2 the category levels were replaced by integer values 1–5 such that the weighted average could be calculated. However, integer values may not be suitable replacements. Finding optimal quantifications is not a straightforward process and several quantification strategies are available.

In survival analysis two different quantification strategies are applied. The first option is to replace each category level by a new variable which takes either value 0 or 1 and the second option is to replace the levels by integer values, as was done in chapter 2.

If the first method is applied each person in the dataset gets a value for each new variable that represents a category level. This value is 1 if the person is in the particular category and it is 0 if he or she is not. If all category levels are coded in this way, each person gets a 0 for each of the new variables except for the one representing the specific person's category level. Hence, many additional variables are introduced by using this strategy. The effect of each variable on the survival time is then estimated individually without taking the other category levels into consideration. If the effects are calculated individually in this way, then any ordering of category levels may be lost. This may be a problem for ordinal categorical data if the order of the categories needs to be preserved. For example, if patients' depression is estimated on a scale from *low*, *medium*, to *high* then it may be expected that the patients with the lowest classification also have the shortest recovery time, and vice versa. If the effects are estimated individually then this ordering may not be retrieved.

The second option of using integer values is often used for ordinal data because it preserves the ordering of the category levels. Namely, the correct ordering is enforced by replacing the levels by increasing integers (*low* = 1, *medium* = 2, *high* = 3). However, these quantifications also enforce equal distances between the effects of consecutive categories which may be unsuitable for the data.

Another quantification strategy is *optimal scaling* which aims to simultaneously find optimal quantifications for the category levels and estimate the effect of each variable on the outcome. This strategy has previously been implemented in several statistical methods, such as the optimal scaling principal components analysis (chapter 2).

For each categorical variable the optimal scaling method starts with finding

quantifications for each category without taking the other categories into account. Hence, it may result in quantifications that are in a different order than the original category levels. This is called the *nominal scaling level* result. An example output of this first step for the scale on depression is shown in Figure 2a. In this example the ordering is wrong because patients with a low level of depression are given a higher quantifications than patients with a medium level.

If the ordering of the category levels needs to be preserved the algorithm continues with the *ordinal scaling level* step. In this step the nominal quantifications which are in a decreasing order are replaced by nondecreasing values. The result for the depression scale is shown in Figure 2b. The first two categories which are in the incorrect order are replaced by their average value. They get similar quantifications and as a result all quantifications are in a nondecreasing order and the original ordering is preserved.

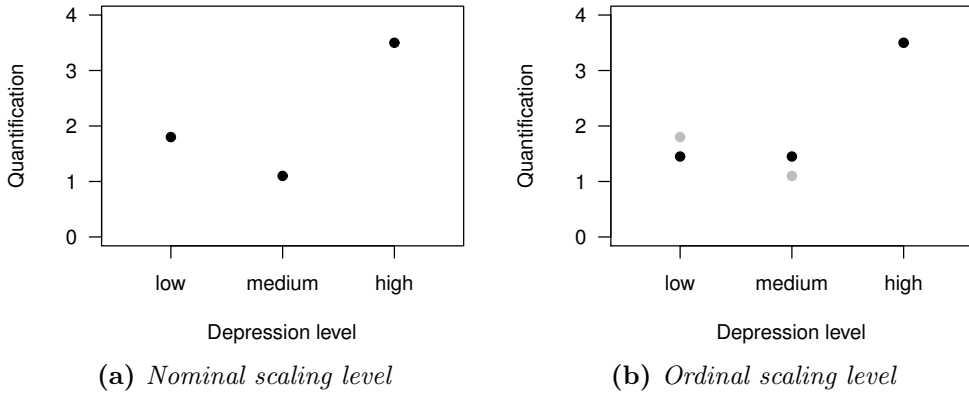


Figure 2: Examples of nominal and ordinal quantifications for the depression scale ranging from low, medium, to high. The nominal quantifications as given in figure (a) are shown as grey dots in figure (b).

In the third chapter the optimal scaling quantification method is incorporated in survival analysis and details about the algorithm are provided. Furthermore a simulation study is performed to compare its performance to the that of the other two quantification strategies. Survival data with an ordinal categorical variable were simulated for this.

From this simulation study it was concluded that the first strategy of estimating the effects individually has a good performance for ordinal data unless the correct ordering of the category levels is difficult to detect. If the ordering is obvious or easy to detect due to the large amount of data, this method can

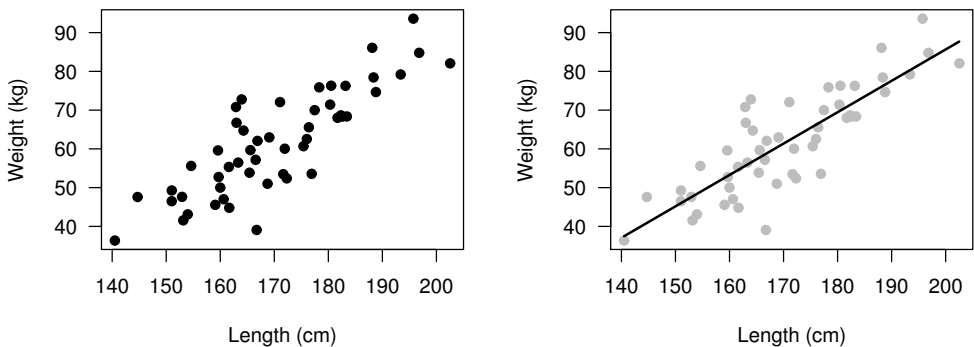
find the correct ordering. In other situations it may be beneficial to enforce the correct ordering by applying the optimal scaling method.

While the first strategy shows good performance in many cases, the simulation study showed that integer coding is a bad quantification strategy for most data. Namely, since it uses equal distances between the levels, it will try to average the effect over all category levels. As a result it will hide any differences between the effects of categories.

Integer coding seems to be a bad strategy and either new variables should be introduced for each category level or optimal scaling should be applied to get insight in the differences in the categories' effects.

Chapter 4

The first three chapters of this thesis focussed on models to analyze survival data. However, there are many other types of models to analyze data. For example, a commonly used method is *linear regression*. This model describes the relationship between a set of variables and a numeric outcome. For example, data may be collected on the length and weight of teenage girls to investigate how the weight of a girl can be predicted from her length. This data example is shown in Figure 3a where each dot represents the combination of the length and weight of one girl. For example, the tallest girl in the dataset is just over two meters tall and weighs about 80 kilograms. The linear regression model will then summarize the relation between the length and weight of girls by fitting a *linear* (straight) line that most closely resembles the observed data points, as shown in Figure 3b.



(a) Observed combinations of weight and length of teenage girls. (b) The estimated straight line that best resembles the observed data (grey dots).

Figure 3: Example of linear regression on data.

Although this fitted line quite closely resembles the relation between length and weight, a straight line is not suitable to explain all possible relations. A typical example is that between the age of a driver and the number of accidents caused by him or her. In general, younger and older drivers cause more accidents than drivers in the middle-age group. Hence, this relation has a u-shape which cannot be captured by standard linear regression. Namely, the straight line that would best fit the data would underestimate the number of accidents caused by young and old drivers and overestimate this number for the middle-age drivers (see Figure 4a). A more flexible model is required for this *nonlinear* relation such that the u-shape can be estimated correctly (see Figure 4b).

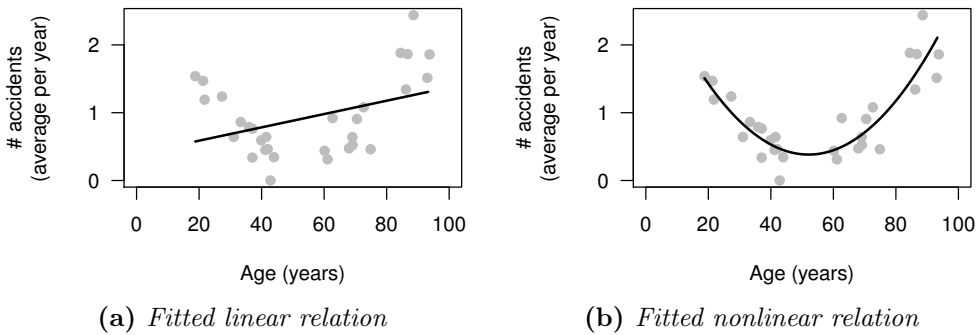


Figure 4: Examples of linear regression and nonlinear regression on data concerning the age of a driver and the number of accidents caused by this person (grey dots).

The method of *optimal scaling* described in chapter 3 has been extended to also quantify numeric data. With this extension this method can be used to model more complicated relations between variables. In the algorithm, each unique numeric value observed in the dataset is assumed to be a separate category. As for categorical data, the method starts by estimating the effect of each of these categories individually. Hence, this step will estimate quantifications for the numeric values in the original dataset. For example, if the age of patients is collected and the first step of optimal scaling is performed on the observed ages, this may result in the quantifications as shown in Figure 5a.

Since not all possible ages between 20 and 40 years were observed, quantifications were only calculated for a limited number of ages in that interval. For example, there are no estimates for 27 to 30 years. To fill this gap, the optimal scaling method will fit a smooth continuous curve through the nominal quantifications. This curved can be either increasing, or both increasing and decreasing, as shown by the solid and dashed lines in Figure 5b. This curve then also provides quantifications for ages that were not observed.

The optimal scaling method to quantify both numeric and categorical data has been implemented in the linear regression model to allow for nonlinear relations. However, the linear regression model can only be used for numeric outcome variables such as body weight and number of accidents. For other types of outcome variables generalized versions of the linear model are used, called *generalized linear models*. The logistic regression model that classifies observations into one of two classes is an example of such a generalized model. This model was used in chapter 2 to estimate the probability of finding a job after one year.

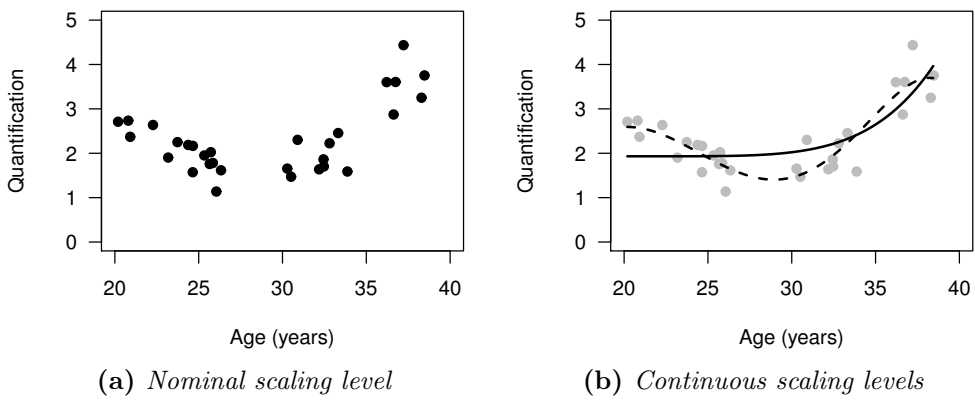


Figure 5: Examples of fitted smooth quantification curves, either increasing (solid line) or both increasing and decreasing (dashed line), on numeric data.

In the fourth chapter of this thesis it is described how the method of optimal scaling can be implemented in generalized linear models. Special focus is on the implementation in the logistic regression model and this combination is applied to three different datasets to demonstrate its use and benefits.

Chapter 5

As increasing amounts of data are collected, statistical results are more frequently used in decision making. Therefore, an important task of statisticians is to communicate their findings in an understandable way. If the results are misunderstood, then all the work put into collecting the data, developing analysis methods, and applying them, is wasted.

Since statistical models are often used to make predictions, clear communication of the estimated probabilities is important. Previous research showed that people who communicate estimated probabilities prefer to express these

verbally by using probability expressions such as *unlikely*, *usually*, and *maybe* because these expressions convey some amount of uncertainty. This preference indicates that a translation step is needed from the estimated numeric probability to an appropriate verbal phrase. In some cases probability scales are used to standardize this translation step. For example, a probability scale may state that the phrase *very likely* should be used for probabilities in the range of 90–95% and *extremely likely* for 95–99%. Usually these scales are symmetrical. So, if *very likely* represents the range 90–95%, then *very unlikely* indicates 5–10%.

Extensive research has been done on the interpretation of English verbal probability phrases. This research showed that there is a large variation between people in their interpretation of these phrases and that it is often asymmetric. For example, the mean of perceived percentages of mirrored expressions such as *likely* and *unlikely* do not sum to 100%. Additionally, interpretation is influenced by the base rate expectation of the statement in which the phrase is placed. For example, the numeric interpretation of *likely* in the statement “*It is likely that it will rain in Manchester, England, next June*” is usually higher than in “*It is likely that it will rain in Barcelona, Spain, next June*”.

These research results show that it is impossible to summarize verbal probability expressions into a (symmetrical) probability scale that would be supported by everyone. Yet, many organizations are using these scales.

Although extensive research has been done on English expressions, it is important to study them in other languages as well. Many international organizations publish documents in more than one language and the meaning of verbal probability expressions may get lost in translation.

In the fifth and final chapter of this thesis a research project on the interpretation of phrases from the Dutch language is described. This study includes both probability phrases such as *waarschijnlijk* (*probably*) and *misschien* (*maybe*) and frequency phrases including *soms* (*sometimes*) and *doorgaans* (*usually*).

The results from this study were very similar to the conclusions of the English studies. Namely, a large variation among the interpretations of the probability phrases was observed and also an asymmetry in the interpretation of mirrored phrases. This shows that Dutch probability scales would not be supported either.

In the study, the interpretations of laypeople was also compared to those of statisticians and no structural differences were found between these groups. Even persons who regularly make and communicate probability estimations do not agree on the interpretation of the phrases. From the results of this study, it was concluded that the use of verbal phrases to express estimated probabilities is a risk for miscommunication.

SAMENVATTING

Dit proefschrift bestaat uit vijf wetenschappelijke artikelen over verschillende statistische onderwerpen, namelijk overlevingsanalyse, optimale-schalingstransformaties en statistiekcommunicatie, of een combinatie van deze onderwerpen. Elk onderzoek is gemotiveerd door een specifieke dataset of door een interesse in een bepaalde eigenschap van data.

Hoofdstuk 1

De motivatie voor het onderzoeksproject dat wordt beschreven in het eerste hoofdstuk is een vraag van de afdeling Psychiatrie van het Leids Universitair Medisch Centrum. Doctoren op deze afdeling verzamelden data over patiënten met stemmings- en angststoornissen om het beloop van hun suïcidale gedachten te onderzoeken. In het bijzonder wilden zij op basis van de eigenschappen van een patiënt voorspellen hoeveel tijd er zit tussen de diagnose en de remissie van suïcidale gedachten. Het doel van het schatten van deze hersteltijd was om te achterhalen welke patiënten een verhoogde kans hebben op aanhoudende suïcidale gedachten.

De statistische methoden die gebruikt worden om dit soort vragen te beantwoorden worden *overlevingsanalysemethoden* genoemd. Deze methoden worden gebruikt om de tijd te modelleren tussen een specifiek startpunt en het moment waarop de gebeurtenis waar de interesse naar uitgaat plaatsvindt. Bij het schatten van de hersteltijd is het startpunt dus het moment waarop de patiënt de diagnose krijgt en de interesse gaat uit naar het moment van remissie van suïcidale gedachten. Overlevingsanalyse kan naast hersteltijd ook gebruikt worden om andere tijdsintervallen te modelleren, bijvoorbeeld de overlevingsduur van patiënten (tijd vanaf diagnose tot overlijden) of werkloosheidsduur (tijd tussen het verliezen van een baan het starten van een nieuwe).

Om de hersteltijd van patiënten te bepalen moeten deze patiënten over lange tijd geobserveerd worden, namelijk tot het moment dat de remissie wordt geconstateerd. Het kost dus tijd om alle benodigde data te verzamelen en dit verhoogt het risico dat de remissie voor sommige personen niet geobserveerd zal worden. Als een patiënt bijvoorbeeld gedurende het onderzoek verhuist naar

een andere stad en daarom overstapt naar een ander ziekenhuis, dan zullen de onderzoekers in het eerste ziekenhuis het herstel van deze patiënt niet waarnemen. Dit fenomeen wordt *censurering* genoemd omdat het herstel van deze persoon niet geobserveerd wordt en dus *gecensureerd* is. Hoewel de data van een gecensureerd persoon incompleet zijn, is de beschikbare informatie wel waardevol. Van de verhuisde patiënt is het bijvoorbeeld wel bekend dat hij of zij nog niet hersteld was op het moment van het laatste ziekenhuisbezoek. Dit geeft aan dat herstel op zijn vroegst *ná* dit bezoek plaats zal vinden.

Aangezien censurering regelmatig voorkomt wanneer data worden verzameld om tijdschattingen te maken, zijn de overlevingsanalysemethoden zo ontworpen dat zij ook dit soort incomplete maar waardevolle data kunnen meetellen. Om deze modellen simpel te houden zijn ze ontwikkeld onder de aanname dat de censurering willekeurig is. Dit betekent dat iedereen dezelfde kans moet hebben om gecensureerd te worden. Die kans mag daarom niet afhangen van de eigenschappen van een persoon of het moment waarop zijn of haar gebeurtenis plaatsvindt. Als een patiënt bijvoorbeeld van ziekenhuis verandert na een verhuizing dan kunnen we ervan uitgaan dat dit net zo goed bij een andere patiënt had kunnen gebeuren. In andere woorden, de oorzaak van censurering is *onafhankelijk* van de gezondheid van de patiënt. Deze onafhankelijkheid is belangrijk voor een correcte analyse.

De artsen wie de hersteltijd probeerden te schatten vermoedden echter dat juist voornamelijk de patiënten wie zich beter begonnen te voelen niet meer naar de kliniek kwamen. Dus, het vermoeden was dat de kans op censurering gerelateerd is aan de gezondheid van een patiënt en dat deze dus niet voor alle patiënten gelijk is. Daarom zou de aanname van onafhankelijke censurering incorrect kunnen zijn en daardoor kunnen de standaard overlevingsanalysemethoden onnauwkeurige resultaten geven. De data zouden namelijk partijdig zijn omdat deze meer informatie bevatten over patiënten met een lange hersteltijd en minder informatie over degenen met een kort herstel. Hierdoor kan de hersteltijd overschat worden. Om te voorkomen dat de door afhankelijke censurering veroorzaakte partijdigheid ook in de resultaten doordringt kunnen er speciale correcties worden toegepast in de overlevingsanalyse.

Dit eerste hoofdstuk richt zich op een specifieke correctiemethode genaamd *Inverse Probability Censoring Weighting Estimator*. Wanneer een observatie gecensureerd is compenseert deze methode voor het verlies aan informatie door in de analyse extra gewicht te geven aan de individuen wie het meest lijken op de gecensureerde persoon en nog wel geobserveerd worden.

Als onderdeel van dit proefschrift is een gebruiksvriendelijke implementatie van deze correctiemethode ontwikkeld voor het veelgebruikte statistische softwareprogramma *R*. Deze implementatie maakt de methode makkelijker beschikbaar

voor onderzoekers wie zelf weinig programmeerervaring hebben.

De methode is daarna toegepast op de data van patiënten met stemmings- en angststoornissen om te corrigeren voor de vermoedelijke afhankelijke censurering. De resultaten lieten zien dat de hersteltijd inderdaad overschat werd zonder deze correctie, maar het verschil was erg klein. Dit geeft aan dat de afhankelijke censurering ofwel niet problematisch was ofwel dat belangrijke informatie over de patiënten mistte waardoor de gewichten niet correct verdeeld konden worden over de patiënten die nog wel geobserveerd werden.

De methode is daarna getest op data met afhankelijke censurering door middel van een simulatiestudie. Deze simulatiestudie toonde aan dat de correctiemethode inderdaad kan corrigeren voor afhankelijke censurering. Echter, zoals te verwachten is, presteert deze methode het beste als voor iedereen de kans op censurering correct kan worden geschat. Om de hersteltijd goed te kunnen schatten moet daarom data over veel personen worden verzameld en moeten ook alle eigenschappen die geassocieerd zijn met censurering geregistreerd worden.

Hoofdstuk 2

In het tweede hoofdstuk ligt de focus op een toepassing van overlevingsanalyse en optimale-schalingstechnieken om mogelijke voorspellers van werkloosheidsduur te identificeren. De dataset die gebruikt is komt van het Nederlandse Uitvoeringsinstituut Werknemersverzekeringen (UWV). In Nederland kunnen werkzoekenden wie recent werkloos zijn geworden een werkloosheidsuitkering aanvragen bij het UWV en ook begeleiding en advies krijgen bij het zoeken naar een nieuwe baan. Omdat het budget beperkt is moeten de kosten van deze adviesdienst gereduceerd worden. Een besparingsstrategie is het verminderen van één-op-één persoonlijke begeleiding. Het UWV heeft daarom een online vragenlijst ontwikkeld om werklozen te selecteren wie moeite hebben om zelf een baan te vinden en daarom baat zullen hebben bij persoonlijke begeleiding. Het UWV zal veel kosten besparen wanneer de persoonlijke begeleiding alleen aangeboden wordt aan de personen wie dit echt nodig hebben.

Via de online vragenlijst verzamelt het UWV data over de werkzoekenden op 17 verschillende factoren, zoals *werkzoekgedrag*, *werkzoekattitude*, *acceptatiebereidheid* en *zelf-effectiviteit m.b.t. sollicitatievoorbereiding*. Deze factoren worden gemeten met 1–5 vragen in de vragenlijst. De laatstgenoemde factor wordt bijvoorbeeld gemeten met behulp van de drie volgende uitspraken:

- *Ik kan goed informatie vinden over vacatures.*
- *Ik kan mijn sterke en zwakke punten voor een baan goed uitleggen.*
- *Ik kan een goede (digitale) sollicitatie schrijven.*

Werkzoekenden worden gevraagd om aan te geven in hoeverre zij het eens zijn

met elk van deze stellingen en de gegeven antwoordopties zijn *zeer mee oneens*, *mee oneens*, *niet eens/niet oneens*, *mee eens* en *zeer mee eens*.

De gegeven antwoorden zijn meestal sterk aan elkaar gerelateerd. Als een persoon bijvoorbeeld in staat is om zelfstandig een baan te zoeken en sollicitaties te versturen, dan zullen zijn of haar antwoorden op de bovenstaande drie uitspraken allemaal *mee eens* of *zeer mee eens* zijn, en vice versa. In andere woorden, de antwoorden op de vragen binnen een factor zijn veelal hetzelfde. Sommige factoren zijn ook sterk aan elkaar gerelateerd. Als een persoon bijvoorbeeld positief staat tegenover het zoeken van een baan, dan zal hij of zij waarschijnlijk ook actief op zoek gaan. Daarom zal deze persoon het ook eens zijn met de uitspraken over zowel de factor *werkzoekattitude* als de factor *werkzoekgedrag*.

Voor statistische analyses is het beter als sterk gerelateerde variabelen niet individueel in een model worden opgenomen, maar eerst worden samengevat in een klein aantal samenvattingsscores. Een simpele manier om deze scores te krijgen is door het (gewogen) gemiddelde te berekenen over de vragen binnen een factor. Deze aanpak met gewogen gemiddelden is gebruikt om elke factor samen te vatten en *factor analyse* is toegepast om de optimale gewichten te berekenen voor elke vraag. Omdat het onmogelijk is om berekeningen te doen met namen van categorieën zoals *zeer mee eens*, werden de vijf antwoordopties vervangen door de gehele getallen 1–5 zodat deze waarden gebruikt konden worden in de berekeningen.

Daarna werden de samenvattingsscores en een aantal aanvullende variabelen, zoals leeftijd, gender en opleiding, gebruikt om de kans dat een werkloze een baan vindt te modelleren. Voor deze analyse werd *logistische regressie* gebruikt welke de werkzoekende kan classificeren in één van twee groepen. De eerste groep bevat de personen wie waarschijnlijk binnen een jaar een baan zullen vinden en de tweede groep bevat de personen wie dat waarschijnlijk niet lukt. Naast de werkstatus na één jaar was het UWV ook geïnteresseerd in deze status na 6 en 9 maanden. Daarom werden er drie analyses gedaan, namelijk om de kansen te berekenen op het vinden van een nieuwe baan binnen 6, 9 en 12 maanden.

Ook al geeft deze analyse al een goed inzicht in welke factoren de kans op het vinden van een nieuwe baan beïnvloeden, zijn er ook andere methodes die gebruikt kunnen worden om de data voor te bereiden en te analyseren. In het tweede hoofdstuk van dit proefschrift wordt voor deze beide stappen een alternatieve methode voorgesteld en toegepast op de dataset van het UWV.

Voor de datavoorbereiding wordt de methode *optimale-schaling principale componentenanalyse (OSPCA)* voorgesteld. Zoals bij factor analyse heeft deze methode als doel om de vele variabelen samen te vatten in een kleinere verzameling van samenvattingsscores. Echter worden deze scores in OSPCA niet individueel per factor berekend, maar zijn het gewogen gemiddelden over *alle*

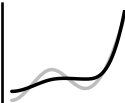
oorspronkelijke variabelen. De gewichten worden op zo een manier gekozen dat de samenvattingsscores samen zo veel mogelijk van de variatie tussen de personen in de dataset beschrijven.

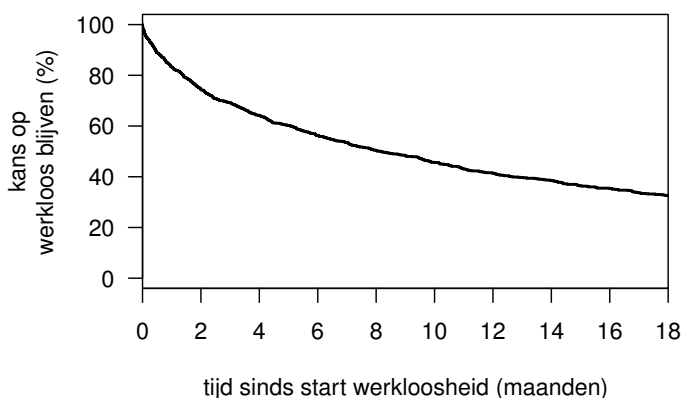
Het grootste voordeel van dit model is dat het ook optimale waarden vindt voor de categorieniveaus voordat het gewogen gemiddelde berekend wordt. Het zou bijvoorbeeld kunnen blijken dat de waarden 1,5, 2, 3, 4 en 4,5 betere vervangers zijn voor de categorieën dan de gehele getallen 1, 2, 3, 4 en 5, omdat de stap van het niet hebben van een mening (*niet eens/niet oneens*) naar het (on)eens zijn met een stelling groter is dan de stap tussen het (on)eens zijn en zeer (on)eens zijn met een stelling. De stappen naar de extremere mening zijn kleiner en worden daarom in dit voorbeeld weergegeven door een kleinere stapgrootte. Op deze manier wordt de volgorde van de categorieniveaus behouden door de vervangende waarden zonder dat er gelijke afstanden tussen de opeenvolgende niveaus worden geforceerd zoals dat werd gedaan bij de vervanging door gehele getallen.

Verder kan OSPCA ook inzicht geven in de relaties tussen de verschillende factoren. Deze methode zal de data van het UWV namelijk samenvatten door gewogen gemiddelden te berekenen over de antwoorden op alle stellingen in plaats van dit per factor te doen. Op deze manier geven de samenvattingsscores ook inzicht in de onderliggende relatie tussen de 17 factoren. Dit kan bijvoorbeeld gebruikt worden om de verwachte relatie tussen *werkzoekattitude* en *werkzoekgedrag* te verifiëren. Dit vermogen om ook de onderlinge relaties tussen factoren te bestuderen is een ander voordeel van OSPCA. Het kan echter ook een nadeel zijn aangezien de samenvattingsscores lastiger te interpreteren zijn als zij berekend worden over alle variabelen in plaats van per factor.

De voorgestelde alternatieve methode voor de uiteindelijke analyse is overlevingsanalyse. Deze methode (welke ook in het eerste hoofdstuk is gebruikt) kan de kans om werkloos te blijven schatten over een heel tijdsinterval in plaats van op een specifiek tijdstip zoals één jaar. Het verloop van deze kans over het tijdsinterval wordt dan gevisualiseerd met een *overlevingsfunctie*. Een voorbeeld van een overlevingsfunctie is gegeven in Figuur 1; deze toont aan dat ongeveer 60% van de personen binnen 12 maanden een baan heeft gevonden.

In dit hoofdstuk worden de verschillen tussen de twee combinaties van methoden beschreven en voor een aantal situaties wordt besproken welke combinatie het meest geschikt is.





Figuur 1: Voorbeeld van een overlevingsfunctie welke de geschatte kans om tot een bepaald moment werkloos te blijven aangeeft. Deze kans is bijvoorbeeld 40% na 12 maanden wat aangeeft dat 60% van de werkzoekende een baan vindt binnen die tijd.

Hoofdstuk 3

De motivatie achter de eerste twee hoofdstukken van dit proefschrift waren vragen van onderzoekers en de geleverde datasets bevatte verschillende typen variabelen. Sommige variabelen, zoals leeftijd, zijn *numerieke* variabelen gemeten op een continue schaal. Andere variabelen zijn *categorisch* en duiden een categorieniveau aan. Dit type data kan verder worden uitgesplitst naar *nominale* categorische data waarvan de categorieniveaus geen rangorde hebben, zoals gender, of *ordinale* categorische data welke wel een rangorde hebben, zoals opleidingsniveau of de mate waarin een persoon instemt met een stelling (*zeer mee oneens*, *mee oneens*, *niet eens/niet oneens*, *mee eens* en *zeer mee eens*).

Voordat er enige berekeningen kunnen worden gedaan op categorische data moeten deze worden omgezet in numerieke waarden. In andere woorden, elke van de categorieniveaus moet worden *gekwantificeerd*. De resulterende numerieke waarden worden ook wel *kwantificaties* genoemd. In hoofdstuk 2 werden de categorieniveaus bijvoorbeeld vervangen door de gehele getallen 1–5 zodat het gewogen gemiddelde berekend kon worden. Echter, gehele getallen zijn mogelijk geen goede vervangers. Het optimaal kwantificeren van categorische data is geen eenvoudig proces en er zijn meerdere kwantificatiestrategieën beschikbaar.

In de overlevingsanalyse worden twee verschillende kwantificatiestrategieën toegepast. De eerste optie is om elk categorieniveau te vervangen door een nieuwe variabele die de waarde 0 of 1 aanneemt en de tweede optie is om de niveaus te vervangen door gehele getallen, zoals werd gedaan in hoofdstuk 2.

Wanneer de eerste methode wordt toegepast krijgt iedere persoon in de dataset een waarde voor elke nieuwe variabele die een categorieniveau vertegenwoordigt. Deze waarde is 1 als deze persoon onder dat specifieke niveau valt, en zo niet dan is de waarde 0. Als alle categorieniveaus op deze manier worden gecodeerd krijgt iedere persoon de waarde 0 voor elke nieuwe variabele behalve degene die zijn of haar categorie vertegenwoordigt. Met deze strategie worden dus vele extra variabelen aangemaakt. Het effect van elk van deze variabelen op de overlevingskans wordt daarna individueel berekend zonder daarbij rekening te houden met de andere categorieniveaus. Wanneer deze effecten op deze manier individueel worden berekend, dan kan het zijn dat een eventuele rangorde van de categorieën verloren gaat. Dit kan een probleem zijn wanneer de rangorde van ordinale categorieën behouden moet blijven. Als bijvoorbeeld de depressie van een patiënt gemeten wordt op een schaal van *laag*, *medium* naar *hoog* dan is het waarschijnlijk dat de patiënten in de laagste groep ook de kortste hersteltijd hebben, en vice versa. Als de effecten individueel worden berekend, dan kan het gebeuren dat deze rangorde niet teruggevonden wordt.

De tweede optie, waarbij gehele getallen worden gebruikt voor de kwantificering, wordt vaak toegepast op ordinale data omdat het de rangorde van de categorieën behoudt. De correcte rangorde wordt namelijk afgedwongen door de niveaus te vervangen voor oplopende gehele getallen (*laag* = 1, *medium* = 2, *hoog* = 3). Echter introduceren deze kwantificaties ook gelijke afstanden tussen de opeenvolgende niveaus wat mogelijk niet geschikt is voor de data.

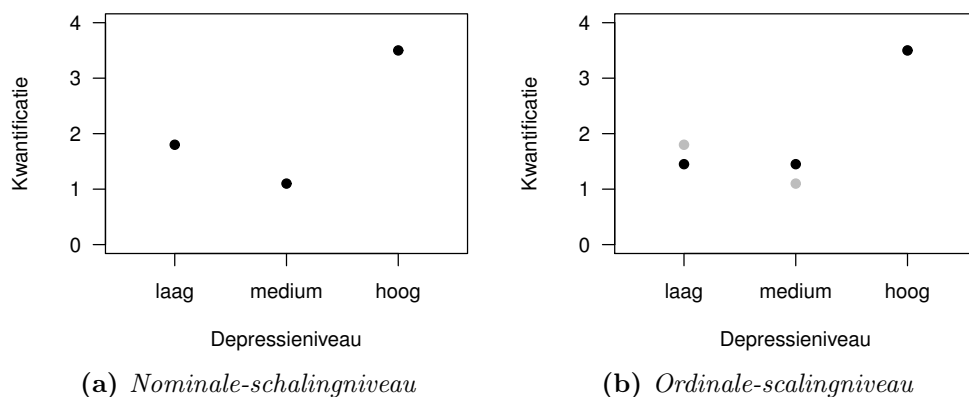
Een andere kwantificatiestrategie is *optimale-schaling* welke als doel heeft om optimale kwantificaties te vinden voor elk categorieniveau en tegelijkertijd het effect van elke variabele op de uitkomst schat. Deze strategie is voorheen al in een aantal statistische methoden geïmplementeerd, zoals in optimale-schaling principale componentenanalyse (hoofdstuk 2).

De optimale-schalingsmethode begint met het vinden van kwantificaties voor elk categorieniveau zonder daarbij rekening te houden met de andere niveaus. Dit kan dus resulteren in kwantificaties die een andere rangorde hebben dan de originele categorieniveaus. Dit wordt het *nominale-schalingniveau* genoemd. In Figuur 2a wordt een voorbeeld van de uitkomst van deze stap weergegeven voor de depressieschaal. In dit voorbeeld is de rangorde onjuist omdat patiënten met een laag depressieniveau een hogere kwantificatie hebben gekregen dan patiënten met een medium niveau.

Als de rangorde van de categorieniveaus behouden moet blijven dan gaat het algoritme verder met de volgende stap waarin kwantificaties worden berekend voor het *ordinale-schalingniveau*. In deze stap worden de nominale kwantificaties die in een dalende volgorde liggen vervangen voor waarden die niet dalen. Het resultaat voor de depressieschaal is weergegeven in Figuur 2b. De eerste twee categorieën zijn in de verkeerde volgorde en worden vervangen door hun gemiddelde. Zij



krijgen dezelfde kwantificatie waardoor alle kwantificaties in een niet-dalende volgorde liggen, zodat de originele rangorde behouden blijft.



Figuur 2: Voorbeelden van nominale en ordinale kwantificaties voor de depressieschaal van laag, medium naar hoog. De nominale kwantificaties uit figuur (a) zijn als grijze stippen weergegeven in figuur (b).

In het derde hoofdstuk wordt de optimale-schalingmethode geïmplementeerd in overlevingsanalyse en dit algoritme wordt in detail uitgelegd. Verder wordt er een simulatiestudie gedaan om de prestaties van deze methode te vergelijken met de twee andere kwantificatiestrategieën. Hiervoor werd overlevingsdata gesimuleerd met een ordinale categorische variabele.

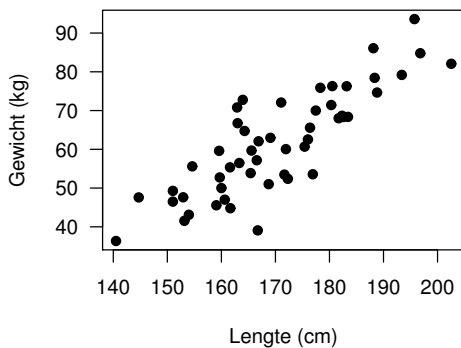
Uit deze simulatiestudie werd geconcludeerd dat de eerste strategie waarin de effecten individueel worden geschat goed presteert voor ordinale data, tenzij de correcte rangorde van de categorieniveaus lastig is op te sporen. Als de rangorde overduidelijk is of makkelijk te vinden is door de grote hoeveelheid data dan kan deze methode de juiste volgorde vinden. In andere situaties kan het gunstig zijn om de juiste rangorde af te dwingen met de optimale-schalingmethode.

Terwijl de eerste strategie dus prima presteert in een aantal situaties, blijkt uit de simulatiestudie dat het gebruik van gehele getallen in veel gevallen een slechte keuze is. Omdat het gelijke afstanden creëert tussen opeenvolgende niveaus wordt het effect van alle categorieniveaus namelijk gemiddeld. Daardoor zal het resultaat eventuele verschillen in effect tussen de niveaus verbergen.

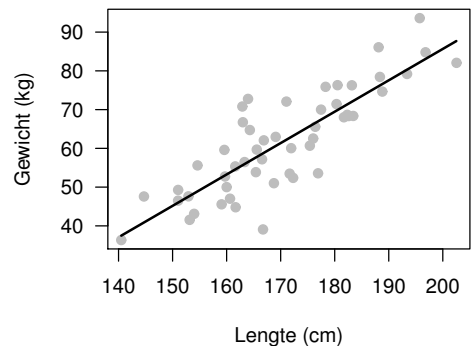
Dus het gebruik van gehele getallen lijkt een slechte strategie te zijn en ofwel nieuwe variabelen kunnen worden geïntroduceerd voor elk categorieniveau ofwel optimale-schaling kan worden toegepast om een beter inzicht te krijgen in de verschillende effecten van categorieniveaus op de overlevingskans.

Hoofdstuk 4

De eerste drie hoofdstukken van dit proefschrift waren gericht op modellen om overlevingsdata te analyseren. Er zijn echter veel meer soorten modellen om data te analyseren. Een voorbeeld is het veelgebruikte *lineaire regressiemodel*. Dit model beschrijft de relatie tussen een groep variabelen en een numerieke uitkomst. Er kan bijvoorbeeld data zijn verzameld over de lengte en het gewicht van tienermeisjes om te onderzoeken hoe het gewicht van een meisje kan worden voorspeld op basis van haar lengte. Dit datavoorbeeld is weergegeven in Figuur 3a waarin elke punt de combinatie van lengte en gewicht van één meisje weergeeft. Het langste meisje in de dataset is bijvoorbeeld iets langer dan twee meter en weegt ongeveer 80 kilogram. Het lineaire regressiemodel zal de relatie tussen de lengte en het gewicht van meisjes samenvatten door een *lineaire* (rechte) lijn te vinden die de geobserveerde data zo goed mogelijk beschrijft, zoals weergegeven in Figuur 3b.



(a) Geobserveerde combinaties van gewicht en lengte van tienermeisjes

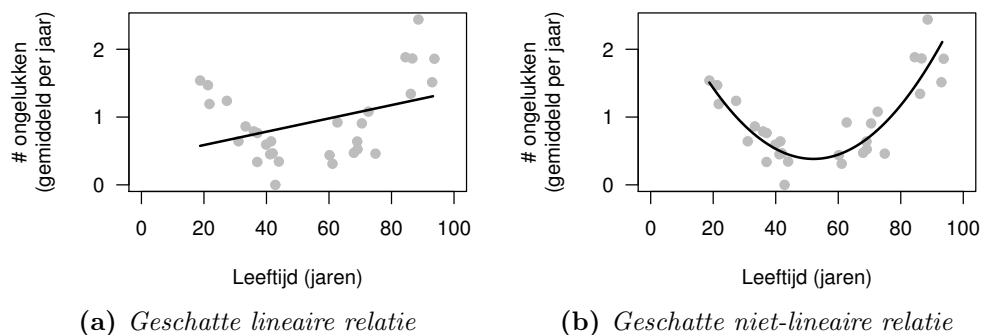


(b) Geschatte rechte lijn die de geobserveerde data (grijze stippen) het beste beschrijft

Figuur 3: Voorbeeld van lineaire regressie op data.

Ook al geeft deze lijn een goed beeld van de relatie tussen lengte en gewicht, is een rechte lijn niet geschikt om alle mogelijke relaties te beschrijven. Een typisch voorbeeld hiervan is de relatie tussen de leeftijd van een bestuurder en het aantal ongelukken dat hij of zij veroorzaakt. In het algemeen veroorzaken jongeren en ouderen meer ongelukken dan mensen in de leeftijdsgroep daartussen. Hierdoor heeft deze relatie een u-vorm welke niet beschreven kan worden met standaard lineaire regressie. De rechte lijn die het beste bij deze data zou passen zou namelijk het aantal ongelukken veroorzaakt door jonge en oude bestuurders onderschatten en het aantal voor de leeftijdscategorie daartussen overschatten

(zie Figuur 4a). Een meer flexibel model is daarom nodig voor deze *niet-lineaire* relatie zodat de u-vorm correct beschreven wordt (zie Figuur 4b).



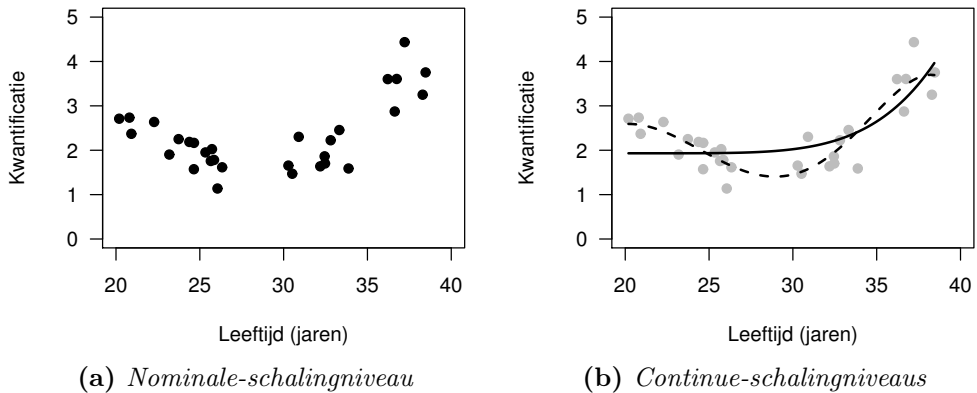
Figuur 4: Voorbeelden van lineaire regressie en niet-lineaire regressie op data over de leeftijd van een bestuurder en het aantal ongelukken dat hij of zij veroorzaakt (grijze stippen).

De *optimale-schalingmethode* beschreven in hoofdstuk 3 is uitgebreid zodat het ook numerieke data kan kwantificeren. Met deze uitbreiding kan de methode gebruikt worden om complexere relaties tussen variabelen te modelleren. In het algoritme wordt elke geobserveerde unieke numerieke waarde gezien als een aparte categorie. Zoals bij categorische data start de methode met het individueel schatten van het effect van elke van deze categorieën. Zo worden in deze stap kwantificaties berekend voor iedere numerieke waarde in de originele dataset. Als bijvoorbeeld de leeftijd van patiënten is verzameld en de eerste stap van de optimale-schalingmethode wordt toegepast op de geobserveerde leeftijden, dan kan dit de kwantificaties opleveren die in Figuur 5a worden weergegeven.

Aangezien niet alle mogelijke leeftijden tussen 20 en 40 jaar zijn geobserveerd zijn er alleen kwantificaties berekend voor een beperkt aantal leeftijden in dit interval. Er zijn bijvoorbeeld geen schattingen voor 27 tot 30 jaar. Om dit gat te dichten zal de optimale-schalingmethode een gladde continue functie tekenen door de nominale kwantificaties. Deze functie kan ofwel stijgend zijn of zowel stijgend als dalend, zoals weergegeven door de ononderbroken lijn en de stippellijn in Figuur 5b. Deze functie geeft dan ook kwantificaties voor de leeftijden die niet waren geobserveerd.

De optimale-schalingmethode om zowel numerieke als categorische data optimaal te kwantificeren is al toegepast in het lineaire regressiemodel om niet-lineaire relaties te modelleren. Dit lineaire regressiemodel kan echter alleen gebruikt worden voor numerieke uitkomsten zoals lichaamsgewicht en aantal ongelukken. Voor andere soorten uitkomsten worden vaak gegeneraliseerde versies van het lineaire model gebruikt, genaamd *gegeneraliseerde lineaire modellen*. Het logisti-

sche model dat in hoofdstuk 2 werd gebruikt om de kans te berekenen dat iemand na één jaar een baan had gevonden en is een voorbeeld van een gegeneraliseerd lineair model. Voor dit model zijn namelijk maar twee uitkomstopties, want iemand vindt binnen één jaar een nieuwe baan of niet.



Figuur 5: Voorbeelden van gladde continue kwantificatiefuncties, ofwel stijgend (ononderbroken lijn) ofwel zowel stijgend als dalend (stippellijn), op numerieke data.

In het vierde hoofdstuk van dit proefschrift wordt beschreven hoe de optimale schalingsmethode kan worden geïmplementeerd in gegeneraliseerde lineaire modellen. Speciale aandacht wordt gegeven aan de implementatie in het logistische regressiemodel en deze combinatie wordt toegepast op drie verschillende datasets om te laten zien hoe de methode gebruikt kan worden en welke voordelen het heeft.

Hoofdstuk 5

Nu er steeds meer data worden verzameld worden statistische resultaten ook steeds vaker gebruikt in besluitvorming. Een belangrijke taak voor statistici is daardoor om hun bevindingen begrijpelijk te communiceren. Als de resultaten verkeerd worden begrepen dan is al het werk gestoken in het verzamelen van de data, het onwikkelen van de analysemethoden en het toepassen daarvan voor niets geweest.

Omdat statistische modellen vaak worden gebruikt om voorspellingen te maken is het duidelijk communiceren van de geschatte kansen belangrijk. Eerder onderzoek heeft aangetoond dat personen wie deze kansen communiceren dat liever verbaal doen met kanswoorden zoals *onwaarschijnlijk*, *meestal* en *misschien*,

omdat deze uitspraken ook een zekere mate van onzekerheid overbrengen. Deze voorkeur geeft aan dat er een vertaaltap nodig is van de geschatte numerieke kans naar een passende uitdrukking. In sommige gevallen worden kansschalen gebruikt om deze vertaaltap te standaardiseren. Een kansschaal geeft bijvoorbeeld aan dat de uitdrukking *zeer waarschijnlijk* gebruikt moet worden voor kansen tussen 90% en 95% en *extreem waarschijnlijk* voor kansen tussen 95% en 99%. Dit soort schalen zijn meestal ook symmetrisch. Dus, als *zeer waarschijnlijk* 90–95% uitdrukt, dan duidt *zeer onwaarschijnlijk* op 5–10%.

Er is al uitgebreid onderzoek gedaan naar de interpretatie van Engelse verbale kansuitdrukkingen. Dit onderzoek toonde aan dat er tussen mensen grote verschillen zijn in hun interpretatie van deze uitdrukkingen en dat deze vaak asymmetrisch is. Bijvoorbeeld, de geïnterpreteerde kansen van gespiegelde kanswoorden zoals *waarschijnlijk* en *onwaarschijnlijk* tellen gemiddeld niet op tot 100%. Verder wordt de interpretatie ook beïnvloed door de aanvankelijke verwachting over een uitspraak waarin een kansuitdrukking geplaatst wordt. De numerieke interpretatie van het woord *waarschijnlijk* in de uitspraak “*Het is waarschijnlijk dat het in juni zal regenen in Manchester, Engeland*” is bijvoorbeeld meestal hoger dan in de zin “*Het is waarschijnlijk dat het in juni zal regenen in Barcelona, Spanje*”.

Deze onderzoekresultaten geven aan dat het onmogelijk is om verbale kansuitdrukkingen samen te vatten in een (symmetrische) kansschaal op een manier zodat iedereen het eens is met deze schaal. Toch gebruiken veel organisaties dit soort schalen.

Ook al is er uitgebreid onderzoek gedaan naar Engelse uitspraken, is het belangrijk om deze ook in andere talen te bestuderen. Veel internationale organisaties publiceren documenten in meer dan één taal en de betekenis van de verbale kansuitdrukkingen zou verloren kunnen gaan bij de vertaling.

In het vijfde en laatste hoofdstuk van dit proefschrift word een onderzoeksproject naar de interpretatie van Nederlandse uitdrukkingen beschreven. In de studie werden zowel kansuitdrukkingen zoals *waarschijnlijk* en *misschien* bestudeerd als uitdrukkingen over frequenties zoals *soms* en *doorgaans*.

De onderzoekresultaten zijn erg vergelijkbaar met de conclusies uit de Engelse studies. Er werden namelijk ook grote verschillen gevonden in de interpretatie van de kanswoorden en ook de asymmetrie bij gespiegelde uitspraken werd waargenomen. Dit toont aan dat ook Nederlandstalige kansschalen niet door iedereen zouden worden ondersteund.

In dit onderzoek werden ook de interpretaties van leken vergeleken met die van statistici en er werden geen structurele verschillen gevonden tussen die groepen. Zelfs de personen wie vaak kansen schatten en communiceren zijn het onderling niet eens over de interpretatie van de kansuitdrukkingen. Uit de resultaten

van deze studie werd geconcludeerd dat het gebruik van kansuitdrukkingen om geschatte kansen over te brengen een risico op miscommunicatie geeft.



ACKNOWLEDGEMENTS

I would like to begin by thanking my supervisors Prof. dr. Jaqueline Meulman and Dr. Marta Fiocco for their guidance and advice during my research. Jacqueline, thank you for supporting me in pursuing different aspects of statistics. I am very grateful that you gave me time to explore what I am interested in, allowed me to do a research project in the communication of statistics, and supported me in participating in committees, boards, and other organizational activities. Marta, thank you for pushing me to work harder when needed, for reading the many versions of my articles very thoroughly (usually at night) and giving numerous useful comments on them, and for introducing me to researchers in your network.

Furthermore, I would like to thank my co-authors Prof. dr. Ionica Smeets and Prof. dr. Casper Albers for our collaboration on the communication project. Ionica, thank you for involving me in your research group and for introducing me to Casper when he proposed an interesting research topic to you. Thank you both for your prompt replies to my many questions during our project.

I would also like to thank my colleagues at the MI for creating the relaxed atmosphere at our institute. I appreciate that initiatives as the PhD colloquium and Mathletico soccer team are supported by the management. These activities have strengthened the bond between PhD students and resulted in many joined activities outside of the institute. I would like to thank all my colleagues with whom I enjoyed sports classes, dancing lessons, Mathletico matches, movie nights, drinks, dinners, et cetera. Special thanks go to my office mates Maarten and Stéphanie who gave me both support for and distraction from my research. Maarten, thank you for your advice on many different topics, for climbing the office chairs to hang up our postcards, and for tolerating our excessive Christmas decorations and the occasional “foute” music. Stéphanie, thank you for the cocktail, movie, theater, and wellness nights, office celebrations, as well as your advice on career issues, dating, books, and time management theories and tools. I also want to thank Anna for providing the template for the design of this thesis.

Annemarie, Moniek, Maxwell, jullie wil ik bedanken voor onze vriendschap. Met jullie heb ik de reis van het studentenleven naar het “echte” werkende leven gemaakt, waarbij de vele late avonden in de sportkantine transformeerden in filmavonden en etentjes, de feestoutfits veranderden in chillpants met sloffen, en Hertog Jan Pilsener werd vervangen door speciaalbier, cocktails of thee. Sinds het einde van onze Helios-taken en we verder uit elkaar wonen zien we elkaar minder vaak, maar met de ingevoerde tradities zal onze vriendschap waarschijnlijk nog lang voortduren. Jelle, jou wil ik bedanken voor al onze leuke gesprekken onder het genot van, bij voorkeur, sushi. Ingrid, bedankt voor al onze gezellige dates.

Julian, jou wil ik ook enorm bedanken. Jij lijkt mij eigenlijk al beter te kennen en te snappen dan ik mezelf. Jij merkt dan ook eerder dan ik wat ik nodig heb; van basisbehoeften als eten en slaap tot schildercursussen en Kaodokus-puzzels. Vooral jouw kennis over hoe je mij, ondanks een eventuele afstand, weer rustig krijgt in tijden van stress en onzekerheid is in de laatste maanden van mijn promotie erg goed van pas gekomen.

Ik wil ook van deze gelegenheid gebruik maken om iedereen van Willems 50, met Jan en Nanny in het bijzonder, nog te bedanken voor alle steun die wij vroeger van hen hebben gekregen. Ik vind het erg fijn dat wij zo vaak bij jullie over de vloer kwamen om samen te eten of te kletsen. Onze gezinnen zijn destijds eigenlijk een beetje samengesmolten en gelukkig is dat altijd zo gebleven.

Ook wil ik Jorine, Johan en mijn ouders bedanken. Jorine, jouw voorbeeldpad naar een PhD heeft mij doen volgen, en Johan jouw zichtbare trots op je zussen werkt nog altijd erg aanmoedigend. Mama, ook al was jij er al niet meer toen ik voor de wiskunde koos, jouw stem heeft altijd in mijn hoofd herhaald dat school en studie belangrijk zijn en dat heeft vast bijgedragen aan de keuzes die geleid hebben tot een PhD-onderzoek. Papa, jou wil ik bedanken voor het altijd klaar staan om ons met van alles te helpen en om advies te geven. Of je het er nou helemaal mee eens bent of niet, jij steunt ons in wat wij ook bedenken.

CURRICULUM VITAE

Sanne Willems was born in Ravenstein, the Netherlands, on February 19th 1989. Upon completion of her pre-university education at the Maaslandcollege in Oss in 2007 she continued her studies at Utrecht University. In addition to the curriculum of the Mathematics & Applications bachelor program she took extra courses to broaden her knowledge in economics, computer science, biology, and psychology. A few months into the Mathematical Science master in Utrecht Sanne switched to the Statistical Science master program at Leiden University to pursue her growing interest in applied statistics. She obtained her master's degree with a thesis project on survival analysis supervised by Dr. M. Fiocco.

Her master project led to the idea to combine the topics of survival analysis and optimal scaling transformation into a PhD research project supervised by Prof.dr. J. J. Meulman and Dr. M. Fiocco at the Mathematical Institute at Leiden University. During her PhD research she was appointed part-time lecturer in the Statistical Science master program.

In anticipation of her appointment as an assistant professor at the Mathematical Institute after her graduation she intensified her involvement in the master program by becoming a study advisor.

During her studies in Utrecht Sanne enjoyed her duties as a board member of her student badminton association and she continued to take up responsibilities within several organizations and societies. For example, she was a board member of the Dutch chapter of European Women in Mathematics and organized the conference Survival Analysis for Junior Researchers in 2018. She has also been an active member of the Netherlands Society for Statistics and Operations Research as a board member of the Young Statisticians, editor of the society's magazine STAtOR, and founding chair of the Statistics Communication section. These last two activities reflect her strong interest in the communication of statistics.