Inaugural-Dissertation
zur Erlangung des Grades Doktor der Naturwissenschaften (Dr.rer.nat)
an der Ludwig-Maximilians-Universität München

The isotonic regression framework

Estimating and testing under order restrictions

vorgelegt von

Georgia Salanti

an der Fakultät für Mathematik, Informatik und Statistik, Januar 2003

Referent: Professor Dr. Kurt Ulm Koreferent: Professor Dr. Ludwig Fahrmeir

Rigorosum: 22 April 2003

Ehrenwörtliche Versicherung

Ich versichere Hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst habe. Die aus fremden Quellen direkt oder indirekt übernommen Gedanken sowie mir gegebene Anregungen sind als solche kenntlich gemacht. Die Arbeit wurde bislang keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlich.

München, Januar 2003

Acknowledgements

This work was supported by a grant UL 94/11-1 of the German Research Foundation DFG (project "Threshold value estimation methods") and a second one from the Special Research Areas foundation SFB 386 (project C1 "Semi- and non parametric approaches") which kept me employed at the Institute for Medical Statistics and Epidemiology at the Technical University of Munich since July 2000. My involvement into these two projects provided funding for conference attendance, literature, and software.

My sincere gratitude goes to my advisors Prof. Dr. Kurt Ulm and Prof. Dr. Ludwig Fahrmeir. First and foremost I must thank Prof. Ulm for his unflagging support and for favoring a stress-free working relationship. He gave me academic guidance and endured me when no results and solutions were forthcoming with interesting suggestions. Dr. Tony Morton-Jones together with Prof. Ulm helped me to get an insight and an in-depth appreciation of the concepts as well as the applications of the isotonic framework.

I'm grateful to my parents Fani and Angelos who have listened to my moan, rejoice and complain during the writing of this dissertation. My sister Tasoula, my uncle George and all my friends in Athens who supported me with their everyday e-mails deserve my deepest acknowledgements. Thanks also to Christos and Victoria for linguistic consulting. Ralf 's music and Martina's jokes at work have been a great source of inspiration. Many thanks to my two flat mates Thomas and Markus for taking so much care of me and for being patient with my fickle mood. Finally thanks to these who a year before assured me they would love me even if I wouldn't have finished this theses.

Georgia Salanti

Contents

1 PREFACE **9**

1.1 The dose-response relationship 10

1.2 Categorizing continuous variables 10

1.3 The order restricted statistics 13

1.4 Contents overview 13

1.5 Data set description 15

2 MONOTONIC REGRESSION **18**

2.1 Introduction 18

2.2 Exponential families 19

2.3 The simple order case and isotonic estimation 20

2.3.1 The greatest convex minorant 20

2.3.2 Computational Algorithms 21

2.4 Testing using isotonic estimates 22

2.4.1 Tests for exponential families 22

2.4.2 Binary response 24

2.5 Isotonic regression as a smoother 25

3 TESTS FOR TREND IN A 2xK TABLE: isotonic alternatives **26**

3.1 Introduction 26

3.2 An overview of tests for trend for binary response 28
3.2.1 Tests that treat the response as continuous 31
3.3 Simulation Study 35
3.3.1 Description 35
3.3.2 No dose-response assumption 36
3.3.3 Increasing trend assumption 39
3.4 Case study in tests for trend 40
3.5 Extensions: Adjustment for dose-induced mortality 44
3.5.1 The Poly-3 test 45
3.5.2 The survival adjusted isotonic Likelihood Ratio test 46
3.5.3 Case study: Example for survival adjustment 46
3.6 Further discussion on tests for trend 47

4 REDUCED MONOTONIC REGRESSION 51
4.1 Reducing the number of solution blocks: why and how 51
4.2 Method A: Elimination using a sequence of Fisher tests 52
4.2.1 Estimation 52
4.2.2 Simulation study: Estimation of the corrected significance level in the backward elimination procedure 55
4.3 Method B: Elimination using the closed test procedure 56
4.3.1 The closure principal 56
4.3.2 Implementation of closed testing for eliminating the solution blocks 58
4.4 Selecting between full monotonic and reduced model 60
4.4.1 An approach based on bootstrap 60
4.4.2 Simulation study: Comparison of full isotonic and reduced isotonic regression 61
4.5 Case study in reduced monotonic regression 64

5 MULTIDIMENSIONAL MONOTONIC MODELS 68

5.1 The multiple regression setting 68

5.2 Additive isotonic model 69

5.3 The isotonic-surfaces model 72

5.3.1 Estimation algorithm 73

5.3.2 Advantages and limitations of the isotonic-surfaces model . . . 76

5.4 Multidimensional extension of the isotonic test for trend 77

5.4.1 Test for isotonic matrix: an asymptotic-based proposal 77

5.4.2 Multiple permutations test 78

5.4.3 Comments on multivariate tests for trend 80

5.5 Reducing the isotonic-surfaces model 81

5.6 Extension of the additive isotonic model for interactions 82

5.7 Case study in multivariate analysis 83

6 THRESHOLD VALUE ESTIMATION 90

6.1 Introduction 90

6.2 Methods 92

6.2.1 Formulation of the problem 92

6.2.2 How to approach the threshold value problem 93

6.2.3 How to estimate thresholds using isotonic regression: two new alternative approaches 94

6.3 Simulation study 98

6.3.1 Evaluating the chi-square approximation in the pooling procedure 98

6.3.2 Evaluate the two approaches on estimating thresholds using isotonic regression 99

6.3.3 Results 103

6.3.4 Comments on simulations 109

6.4 Adapting the closed testing procedure for estimating thresholds . . . 112

6.5 Case study in threshold value estimation 114
6.6 Conclusions 116

7 MONOTONIC REGRESSION IN SURVIVAL ANALYSIS 118
7.1 Introduction and background 118
7.2 The time-varying coefficients Cox Model 121
7.3 Detecting PH departures under order restriction 122
7.3.1 Smoothing Schoenfeld residuals scatterplot 122
7.3.2 Grambsch and Therneau test and its isotonic version 124
7.4 Fitting the generalized additive model using isotonic smoothing techniques 125
7.5 Simulation study 127
7.6 Case Study in time-varying Cox model 129
7.7 Extensions 133

8 SUMMARY 136
8.1 Summary 136
8.2 Zusammenfassung 140

A APPENDIX: Software implementation in S+ 144

Bibliography 152

List of Tables 159

List of Figures 162

List of Algorithms 165

1 PREFACE

Classical statistical techniques - models and tests – for explanatory analysis can be classified into two categories. The first one contains approaches which are easy to interpret and to apply but are too "demanding" in terms of requirements for the nature of the data. In this category falls every technique that lies somehow on special assumptions regarding the dose-response shape, for example linear regression. In the second category belong less restrictive techniques as spline models, but the price one has to pay is the loss in simplicity; in many cases these models are very difficult to interpret or to drawn conclusions for the data.

As a moderate solution simple and less restrictive approaches more flexible than generalized additive models have been developed. Recursive partitioning based models [68] offer a reasonable alternative. This model is tree-structured and is fitted by splitting progressively the dataset in subgroups with respect to the maximization of a function. Here, another framework will be studied, ***monotonic regression related methods.*** This approach is simple and the only requirement is monotonicity, which is the minimum of requirements for the first category. Furthermore, the result is easy to interpret.

When analyzing experiments, the selection of statistical method is made regarding the pre-defined goals of the study. It is important to know when to use which

method. Thus monotonic approach is not always desired; only in the frame of a certain sort of aims. This method is especially adequate when *the proof of a dose-response relationship* is of interest and *categorization of the predictor variables* regarding the outcome is important. These two issues are discussed below.

1.1 The dose-response relationship

With the most general sense, the term dose-response is defined as *the shape of the exposure-outcome curve*, whatever that shape may be. A more strict definition places in the dose-response term the restriction of monotonicity. Breslow's definition [10] is *"a relationship in which a change in amount, intensity, or duration of exposure is associated with a change - either an increase or a decrease - in risk of a specific outcome"*. For a thorough discussion on the definition of dose-response see [38].

Under this consideration, establishing a dose-response relationship implies the proof of monotonicity. This is related to a family of tests, the *tests for trend*. In epidemiology the proof of a dose-response relationship is listed as one of the criteria for inferring causality. Thus, dose-response analysis is an important issue on analyzing experiments and the procedure followed can play a determinant role in the result of the study.

1.2 Categorizing continuous variables

Categorizing continuous variables arises as an important task in statistical analysis, especially in studies concerning exposure-effect problems [6]. This practice is used sometimes for the outcome variable of an experiment, but it is more common for explanatory variables [28]. There is a great debate among statisticians about the

strategy that should be followed to categorize a variable given the several advantages and disadvantages for each of the categorization techniques.

Create meaningful groups of the predictor variables regarding the outcome is desirable in many studies. Consider for example a car insurance study: an insurance company wants to test whether the accident risk is related to the age of the driver and to state several groups with increasing fees. Categorization is also routinely observed in medical studies, for example on deciding between several types of treatment for a patient regarding high blood pressure. The fitted values from the chosen model split the data in "risk" groups. To compare between several categorization patterns with multiple cuts for one or more explanatory variables, the area under the Receiver Operating Characteristics (ROC) curve can be obtained: the bigger, the better.

While creating categories is popular and attractive, it creates many problems. Grouping is equivalent to introducing a sort of measurement error and leads to an inevitable loss of power. The reduction in efficiency along with the introduction of bias are the most common drawbacks when a continuous variable is summarized to an ordinal one. An intuitive functional representation for the categorization of a continuous variable to ordinal is a step function with respect to the outcome. However, whether a step function is biologically plausible for the effect of exposure, is questionable in many studies. The researcher has to decide a priori whether he believes that a function with abrupt jumps fits adequately the data or not. Grouping may hide important complexities of the exposure-effect relationship. For example, in cases where the effect of a the exposure is observed to - let say - the upper 10% of the exposure, grouping the data into categories will dilute and even conceal the effect.

On the other hand, categorization is attractive for many reasons. An important one, although the least mathematically justified, is simplicity: the analysis is often easier to perform and it is understandable by non-statisticians. Some other reasons

for this preference are statistical. The researcher has no longer to consider some issues as it is the case when modeling numerical variables: assumptions about the shape of the dose-response curve and problems related to influential observations (for example outliers) are sometimes difficult to deal with. Finally, if the analysis based on categories indicates that the relationship between explanatory variable and outcome has a simple form, the investigator has always the option to model using the continuous form. The analysis using variables in categories is so conceptually simple and easy to interpret, that explains its popularity despite the objections described above.

After having decided to categorize a predictor, one has to define the cutpoints. One common approach is based on binary splits and the corresponding cutpoint is determined by the maximization of a test statistic [17, 20, 36, 43]. A binary split is simple to interpret, but increases the impact of the disadvantages of categorization described above. The simplicity is gained at the expense of throwing away a lot of information. An optimal stratification of the predictor into more than two groups, can often be more informative, especially if the shape of the dose-response relationship is of interest. Multiple cut-offs lead to less biased estimates and the small loss in power is offset by gain in simplicity.

It is not always obvious how many groups should be build and where the cutpoints should be placed. The pattern of the response, the underlying biological mechanism and the sample size should be taken into account. However, it is commonly observed that arbitrary, equally spaced or equally sized cutpoits, suggested by the sample size are used in practice. Nevertheless, rather than grouping according to the distribution of the explanatory variable, a better strategy is to base the selection of the cutpoints on the outcome. If there is more than one explanatory variable, the application of a statistical model is necessary [28].

This theses deals with situations where categorization of numerical predictor vari-

ables results as effect of the dose-response relationship. Moreover, as it actually occurs in practice, more than one explanatory variable has to be included in the analysis, and therefore one has to apply an appropriate statistical model. The variables may interact, and in this case the categorization can be seen as a combination of the predictor variables in homogenous subgroups.

1.3 The order restricted statistics

The origins of order restricted statistical inference are dated back to the 50s. David Bartholomew was one of the first researchers who started working on this topic. Monotonic regression in its current form appeared first in the book *Statistical Inference under order restrictions* by R. Barlow, D. Bartholomew, J. Bremner, and D. Brunk [3]. Monotonic regression became very popular in applications much later; its utility in testing [21, 37, 39, 53] in modeling [2, 44, 55, 62] and estimating thresholds [63] has gained a lot of attention recently. Monotonic regression has two principal characteristics:

- The order restricted nature in estimating and testing
- The step function shape

Automatically the reader can imagine the utility of this method in modeling and stratifying in the context outlined in the preview paragraphs: in dose-response analysis and in categorizing continuous predictor variables.

1.4 Contents overview

While quite simple as concept, monotonic regression presents some difficulties in practice. The most important one is related to the outcome variable; many of

the monotonic-related methods used so far for continuous response, are either not developed or misapplied when the response is binary. The following chapters will be strictly concerned on developing methods and "filling gaps" regarding binary responses. Another problem is the poor performance of asymptotics in the majority of the monotonic-related models. In this report different uses of monotonic regression (as test for trend, as modeling alternative, as threshold value estimation method) are put together. Old approaches are evaluated and new developments are proposed.

In **CHAPTER 2** some basics about monotonic regression will be shortly presented. In **CHAPTER 3** the monotonic test for trend and its importance in establishing dose-response relationship are discussed. It will be proven that the asymptotic approximation for its distribution proposed by Bartholomew does not always hold and an alternative based on permutations is presented. Within a simulation study, several tests for trend will be compared to the monotonic test, and their performance will be evaluated. In **CHAPTER 4** monotonic regression will be studied as modeling proposal and categorization method. Two new methods for backward elimination of the model in order to improve parsimony will be introduced and evaluated. Chapter (**CHAPTER 5**) deals with multidimensional extensions of monotonic regression and their utility in two-dimensional classification. A new multidimensional version of the monotonic test for trend for overall a partial significance will be presented. **CHAPTER 6** outlines the use of monotonic regression in threshold value estimation problems. In this regard, two new proposals will be introduced. Finally **CHAPTER 7** gives a taste about how one can introduce monotonic regression in Cox models to correct for time-varying effects. A help on the isotonic library in S+ is added in the **Appendix**.

1.5 Data set description

MAK (Maximale Arbeitsplatz Konzentration) study.

The data are taken from the DFG study "Chronic Bronchitis" [15]. A detailed description of the data can be found in the monograph of the study (DFG-report, 1978 and 1981) or in [27]. Data from 5578 workers of three different plants (in Moers, Munich and Saarbruecken) are available. The three plants had a mixture of dust, mainly from iron, steel, foundry and engineering. Here, the data from the plant "Munich" are analyzed (see table 1.1).

Table 1.1: The MAK study data (sample from Munich).

Variables	With CBR	Without CBR	**Total**
	Non and ex- smokers: median(min-max)		
n (%)	51(15.6%)	275(84.4%)	326
Time since first exposure	33(8-66)	23(1-55)	24(1-66)
Inhalable dust (mg/m^3)	1.5(0.4-8)	1.4(0-15)	1.4(0-15)
	Non and ex- smokers: median(min-max)		
n (%)	241(26.2%)	679(73.8%)	920
Time since first exposure	28(6-49)	24(3-51)	25(3-51)
Inhalable dust (mg/m^3)	4.6(0.3-12.1)	1.07(0.2-15)	1.4(0-15)

The goal of the statistical analysis has been to test whether the inhalable dust concentration in workplace has adverse effects on the health of the workers. Apart inhalable dust, additional factors as the time since first exposure – highly correlated

with age – and the smoking habits need to be taken into account. The endpoint of the study has been the chronic bronchitic reaction (CBR). In the statistical analysis of this data, the proof of a dose-response relationship i.e. increasing risk with increasing dust concentration was an important task in order to establish causality. In case of evidence, the stratification of dust concentration into certain risk groups was of great interest: according to the established risk categories the MAK commission could take decisions to protect workers against the dust effects and to assess a threshold limit value for dust concentration in workplace.

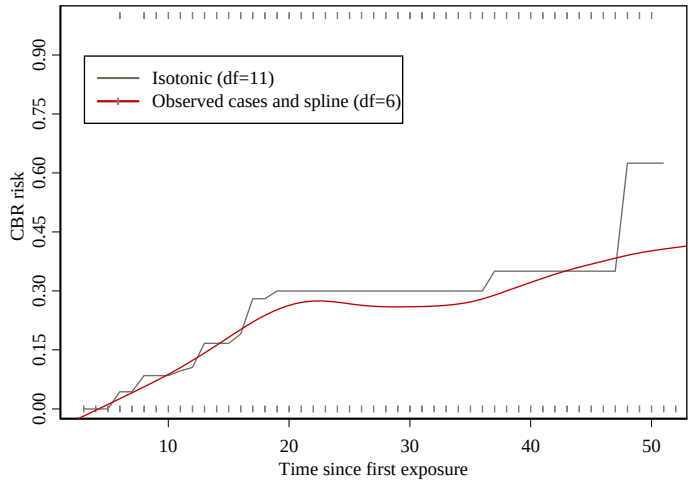

Figure 1.1: Example from MAK study. Isotonic regression together with smoothing spline.

Just to get a taste about isotonic regression, an example is presented (figure 1.1). The chronic bronchitic reaction probability is a function of time since first exposure. A smooth curve (spline with 6 degrees of freedom) can be fitted to summarize the relationship, whereas a step function (isotonic regression) is also possible. The categorization of the variable TIME occurs as the result of this relationship which segments the y-axis in 11 categories.

In most of the chapters, the MAK study is used to demonstrate the methods. The Para-Aramid study or the Acute Leukemia study are used in addition, and they are described in the corresponding chapters.

2 MONOTONIC REGRESSION

2.1 Introduction

This chapter surveys the most important issues about monotonic regression as introduced by Robertson et al. [49]. It concentrates only on the case of one explanatory variable - *dose*. Definitions, estimation procedures, tests and their asymptotic distribution are summarized, and a view of the isotonic procedure as a smoother is given.

Although we are mainly interested in the case of binary response, for the sake of completeness the monotonic framework for the parameter θ of a distribution belonging to the exponential family is presented. Note that under the term monotonic regression either an increasing (isotonic) or decreasing (antitonic) trend is included.[1] Without loss of generalization, an increasing trend is assumed (isotonic regression) throughout this chapter.

[1] The terms "isotonic" and "monotonic" are often used somewhat interchangeably.

2.2 Exponential families

Consider the case of K dose groups $\mathcal{D} : \{d = (d_1, d_2, ..., d_K)\}$ where the dose levels are in increasing order and the outcome of an experiment is Y_{di}, $i = 1, ..., n_d$. The response $y = g(d) = (Y_{dj})$ does not have to be specified and it can be binary, Poisson, continuous or survival time. The distribution of Y_{di} can be written in the following form

$$f(y; \theta_d, t) = exp(p_1(\theta_d)p_2(t_d)K(y; t_d) + S(y; t_d) + q(\theta_d, t_d)) \tag{2.1}$$

where θ_d is the parameter of the distribution of Y_{di} for a given dose. For example θ_d can be the mean of the continuous distribution, the Poisson parameter, or the positive outcome probability of the binary response. As p_1, p_2, q are denoted some functions which have continuous second derivatives and they satisfy

$$p_1(\theta_d) > 0, p_2'(t_d) > 0 \text{ and } q'(\theta_d, t_d) = -\theta_d p_1'(\theta_d)p_2(t_d) \tag{2.2}$$

with respect to θ_d.

Considering the independent dose-sample, it is $E[K(y, t_d)] = \theta_d$ and $V[K(Y, t_d)] = 1/[p_1'(\theta_d)p_2(t_d)]$. The maximum likelihood estimator of the parameter θ_d is

$$\hat{\theta}_d = \frac{1}{n_d} \sum_{j=1}^{n_d} K(Y_{ij}; t_d) \tag{2.3}$$

Consider now K independent dose samples from K populations belonging to exponential families with densities $f(\bullet, \theta_d, t_d)$ for every $d \in \mathcal{D}$. Then without any *a priori* assumption about the shape of the parameter θ, the corresponding maximum likelihood estimator is $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K)$. Given that the dose is of increasing order, we wish to have $\hat{\theta}$ of non decreasing order. This is true only if all the estimates

fulfill the isotonic relationship $\hat{\theta}_d \leq \hat{\theta}_{d+1}$. Then, if $\hat{\theta}^*$ is the isotonic estimator of $\hat{\theta}$, it is trivial that $\hat{\theta}^* = \hat{\theta}$. If somewhere is a violator such that $\hat{\theta}_d > \hat{\theta}_{d+1}$ for some d, then an isotonic estimator of θ needs to be found.

2.3 The simple order case and isotonic estimation

Definition 2.1 *Consider a given function g on $\mathcal{D}$. A function g^* on $\mathcal{D}$ is an **isotonic regression** of g with weights w if and only if g^* is isotonic and g^* minimizes*

$$\sum_{d \in \mathcal{D}} [g(d) - f(d)]^2 w_d \tag{2.4}$$

over the class of all isotonic functions f on $\mathcal{D}$.

For the case of an exponential family it can be shown that $w_d = n_d p_2(t_d)$. Before presenting an algorithm that computes the function that minimizes equation 2.4, a graphical representation of the problem will be discussed.

2.3.1 The greatest convex minorant

Let $U_d = (w_d, g(d))$, $W_d = \sum_{i=1}^{d} w_i$ where $w_i, i \in \mathcal{D}$ defined as previously and $G_d = \sum_{i=1}^{d} g(i) w_i$ with $d \in \mathcal{D}$. The plot of U_d is called *cumulative sum diagram CSD* and depends on function g. The slope of the segment joining points U_{d-1}, U_d equals $g(d)$. The supremum of all convex functions lying below the cumulative sum diagram is the *greatest convex minorant GCM* of CSD. Since GCM is convex on $[0, W_K]$, it is left differentiable at each point U. Denote now g^* the left derivative of GCM. If for some d, GCM(d) lies below CSD(d) then the slope on the left and the right of d are the same[2] i.e $GCM(d) < CSD(d) \rightarrow g^*(d) = g^*(d+1)$.

[2]For simplicity two successive dose groups are denoted as d and $d+1$ and not d_i and d_{i+1}

Theorem 2.1 *If $\mathcal{D}$ is simply ordered, the left derivative g^* of the greatest convex minorant yields the isotonic regression of g.*

Indeed, if f is isotonic then the following holds

$$\sum_{d\in\mathcal{D}}[g(d)-f(d)]^2 w_d >= \sum_{d\in\mathcal{D}}[g(d)-g^*(d)]^2 w_d + \sum_{d\in\mathcal{D}}[g^*(d)-f(d)]^2 w_d$$

2.3.2 Computational Algorithms

While many computational algorithms are proposed, the most widely used one is the *Pooled Adjacent Violators Algorithm* or PAVA. The background and justification for this algorithm is related to the greatest convex minorant. Recall that a graph of observations is isotonic if the CSD is convex, since in this case it is equal to GCM. A violator of the isotonic assumption occurs when the slope of a segment between $g(d-1)$ and $g(d)$ is smaller than the slope in the previous segment $g(d-2), g(d-1)$. If these two segments are replaced by a segment $g(d-2), g(d)$ then the GCM of the new graph is the same as the GCM of the old graph. Following this idea, the GCM can be constructed by a sequence of estimations where adjacent line segments are replaced by a single line segment to correct for violators of monotonicity. Alternatively to PAVA, the Minimum Lower Set algorithm [49] can be used. This algorithm involves averaging g over suitable selected subsets of $\mathcal{D}$ and it is more general than PAVA. It can be also applied for several shape restrictions, as the U-shape.

Consider again the situation of a set $\mathcal{D}: d_1, ..., d_K$ of dose groups where the dose is in increasing order and the outcome $g(d)$. To estimate $g^*(d)$ the isotonic regression of $g(d)$, the PAVA as described in algorithm 1 can be applied. The algorithm assuming a decreasing trend is similar.

We revisit the situation described in section 2.2. The isotonic maximum likelihood estimator is the isotonic regression of $\hat{\theta}_d$ estimated through PAVA with weights $w_d = n_d \cdot p_2(\hat{\theta}_d)$.

Algorithm 1 ***The Pooled Adjacent Violators Algorithm***

1. *If $g(d)$ is in non-decreasing order then $g^*(d) = g(d)$.*
2. *Otherwise there is somewhere a violator such that $g(d) > g(d+1)$ for some d. Replace these two values by their weighted average $Av\{g(d), g(d+1)\} = [w_d g(d) + w_{d+1} g(d+1)]/(w_d + w_{d+1})$*
3. *Now the elements $d, d+1$ form a block called level set (LS) or solution block. If the new set of $K-1$ values is isotonic, then $g^*(d) = g^*(d+1) = Av\{g(d), g(d+1)\}$ for the violator and $g^*(d) = g(d)$ for all other observations.*
4. *If the set is not isotonic repeat the procedure using the new set of values*

Example 1: The normal means

The distribution function has the exponential form:
$f(y;\theta,t) = exp\{\theta(1/t)y - (\theta^2/2t) - (\theta^2/2t) - (ln(2\pi t)/2)\}$ where θ is the mean of the distribution and t the variance. Consequently $S(y;t) = -y^2/2t$, $q(\theta,t) = -\theta^2/2t - ln(2\pi t)/2$, $K(y;t) = y$, $p_2(t) = 1/t$ and $P_1(\theta) = \theta$. The isotonic estimator of the mean $\hat{\theta}$ is obtained by applying PAVA with weights $w_d = n_d/t_d$.
Equivalently for binary response the weights are equal to n_d.

2.4 Testing using isotonic estimates

2.4.1 Tests for exponential families

The following hypotheses are defined:

$$H_0 : \theta_1 = \theta_2 = ... = \theta_K = \theta_0$$

$$H_1 : \theta_1 \leq \theta_2 \leq ... \leq \theta_K \text{ with at least one strict inequality}$$
$$H_2 : \text{No restrictions for } \theta_1, \theta_2, ..., \theta_K$$

where $\theta_0 = \dfrac{\sum_{d=1}^{D} w_d \hat{\theta}_d}{\sum_{d=1}^{D} w_d}$ is the weighted mean and $\hat{\theta}_d$ the mean in group d. Denote as T_{01} the test statistic that tests H_0 against H_1 and T_{12} the test statistic that tests H_1 against H_2. These two tests have the following form:

$$T_{01} = 2\sum_{d=1}^{K} w_d \hat{\theta}_d [p(\hat{\theta}_d^*) - p(\theta_0)] + 2\sum_{d=1}^{K} w_d [q(\hat{\theta}_d^*, t_d) - q(\theta_0, t_d)] \tag{2.5}$$

$$T_{12} = 2\sum_{d=1}^{K} w_d \hat{\theta}_d [p(\hat{\theta}_d) - p(\hat{\theta}_d^*)] + 2\sum_{d=1}^{K} w_d [q(\hat{\theta}_d, t_d) - q(\hat{\theta}_d^*, t_d)] \tag{2.6}$$

where $\hat{\theta}_d^*$ is the isotonic estimator of $\hat{\theta}_d$ assessed using PAVA. These tests follow approximately a weighted chi-square distribution:

$$P(T_{01} \geq c) \sim \sum_{l=2}^{K} P(l, K, w) P[X_{l-1}^2 \geq c] \tag{2.7}$$

and

$$P(T_{12} \geq c) \sim \sum_{l=2}^{K} P(l, K, w) P[X_{K-l}^2 \geq c] \tag{2.8}$$

where $P(l, K, w)$ denote the probabilities that under H_0 and given K distinct dose-levels, the isotonic regression will build l level sets and $\sum_l P(l, K, w) = 1$. For a more detailed description of the weights $P(l, K, w)$ see 2.4 in [49].

2.4.2 Binary response

Consider that $Y_d \sim B(p_d, n_d)$. There are several modifications of the test in equation (2.5) for this case. In their majority they assume that the proportions approximately follow a normal distribution with $\bar{Y}_d = p_d$ and $\sigma^2(Y) = p_0(1-p_0)/n_d$. Then the test takes the form $\bar{X}^2 = \dfrac{\sum n_d(\bar{p}_d - p_0)^2}{p_0(1-p_0)}$ and under H_0 it is distributed as in equation 2.7. Alternatively, the test

$$\bar{E}^2 = \frac{\sum n_d(\bar{p}_d^* - p_0)^2}{\sum(p_d - p_0)^2} \tag{2.9}$$

can be used. $\bar{E}^2$ follows a weighted average of beta random variables with level probabilities $P(l, K, w)$ defined as before.

However, it is well known that the normal approximation for proportions performs poorly and thus these two tests are not always adequate. Approximation for proportions using sine and cosine transformations lead to *Arcsine Isotonic Test* that behaves very unpredictably [13].

One of the most popular alternatives is the *Isotonic Likelihood Ratio test.*

$$R = D(\hat{p}_{H_0}) - D(\hat{p}_{H_1}) = 2\sum_{d=1}^{K}[n_d\hat{p}_d ln(\frac{\hat{p}_d^*}{\hat{p}_0}) + n_d(1-\hat{p}_d)ln(\frac{1-\hat{p}_d^*}{1-\hat{p}_0})] \tag{2.10}$$

where the deviance $D(\hat{p}_{H_d})$ is the function -2log(Likelihood) under the hypothesis H_d, $d = 0, 1, 2$. This test, very popular so far, has been used in connection with equation 2.7 following the proposal in [49]. This test will be evaluated in the next chapter. The assumed asymptotic distribution does not hold for this test as will be shown later.

2.5 Isotonic regression as a smoother

The process of isotonic regression can be thought of as a smoothing technique. The term smoothing is here somewhat excessive, since the result is far from being smooth because of the presence of "flat spots" in an increasing regression. With smoothing here we refer more to the fact that isotonic regression is connected to conditional expectation and this conditioning is referred to as a smoothing process: the values of the variables are regressed and replaced in the conditioning process by constant values, which is a smoothing operation. Under the light of this consideration, isotonic smoother has global nature but results in locally flat averaging.

Following the proposal of Hastie and Tibshirani [29] the degrees of freedom of a smoother are defined as the trace of the smoother matrix S. Let B_k be the subset of successive indexes from $i = 1, ..., K$ corresponding to observations $g(d_i)$ that are estimated through $g^*(d_k)$ i..e a level set and $k \in (1, l)$ with l denoting the final number of level sets. The smoother matrix will have the following form:

$$S_{ij} = \begin{cases} \dfrac{w_d}{\sum\limits_{s \in B_k} w_s} & \text{if } d \text{ and } j \in B_k \text{ for some } k \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

Extensions of isotonic regression to be smooth were proposed by Friedman and Tibshirani [19]. The idea is based on combining moving average and isotonic regression. First the scatterplot of the observations $(g(d), d)$ is "smoothed" by replacing $g(d_i)$ by an average of the $g(d_j)$ over a window of values d_j around $g(d_i)$. At these values, isotonic regression is applied. It is also possible to start from isotonic regression and to smoothing then. Combining isotonic regression with other smoothing techniques as splines is also an alternative.

3 TESTS FOR TREND IN A 2xK TABLE: isotonic alternatives

3.1 Introduction

Many important decisions about the classification of substances at the workplace are based on animal experiments. Among the most important ones are those concerning carcinogenic agents. As outlined in chapter 1, a main criterion in order to establish causality is the proof of a dose-response relationship. This is accomplished by a "test for trend" [25, 26].

One of the most complete guidelines on selecting the appropriate test for trend is the tutorial by Chuang-Stein and Agresti [12]. Among the tests presented in this paper, the *Cochran-Armitage* test (*CA*-test) [1] appears as the most widely used. However, it is well-known that the result of this test is associated with the score assigned in the dose levels.

Recently the isotonic approach has gained a lot of attention on testing dose-response relationships [21, 37, 39]. The most important reason is that applying isotonic-based

tests, the result is independent of any score assignment. The previous chapter discussed this test (equation 2.5) and the different forms it can take when the response is binary ($\bar{X}^2$, $\bar{E}^2$ in section 2.7), as well as the Isotonic Likelihood Ratio R test (equation 2.10) and the corresponding large sample approximations. Other approaches with connection to isotonic procedure have been proposed by Mancuso et al. [39]; they proposed a mixture of CA test and isotonic estimation. Peddada et al. [48] introduced a test based on the width of the interval of the isotonic estimators, and Gautam et al. [21] proposed an isotonic modification of the usual Pearson's X^2 for ordered contingency table.

Table 3.1: Notation used for the various test-statistics.

Dose-levels	d_1	d_2	...	d_K	$\sum$
No of response	e_1	e_2	...	e_K	E
Total	n_1	n_2	...	n_K	N
Proportion	p_1	p_2	...	p_K	$p_0 = \frac{E}{N}$
Score	s_1	s_2	...	s_K	$s_0 = \frac{\sum s_i n_i}{N}$

This chapter presents a review of tests for trend, focusing on isotonic regression. The Isotonic Likelihood Ratio test R will be discussed in detail and I will show that the large sample approximation presented in the previous section and used in several papers does not always hold. Furthermore, a comparative study among the most popular tests for trend will outline the advantage on using the PAVA transformation when testing monotonic shapes. The basis of this comparative study is a paper by Ulm et al. [60]. They compared the Cochran-Armitage test to isotonic likelihood ratio test. Starting from this aim, I will include in the study a rank test, the isotonic Pearson's X^2 test by Gautam [21] and a skewness correction for the Cochran-Armitage test. Additionally, the results will be based on 10 000 permutations instead of 1000 used in the original paper.

The type of experiments which are considered throughout the paper can be displayed in the form of table 3.1.

3.2 An overview of tests for trend for binary response

The Cochran-Armitage test

This test is designed to detect a linear trend in a response proportion. It is assumed that

$$P(X = 1) = p_i = a + \beta s_i + \epsilon,$$

where β is the slope of the regression line and s_i the score assigned to dose d_i. The slope is estimated by the usual weighted least squares-method [18]

$$\hat{\beta} = \frac{\sum n_i(\hat{p}_i - p_0)(s_i - s_0)}{\sum n_i(s_i - s_0)^2}.$$

The test due to Cochran and Armitage [1] is applied to investigate whether $\hat{\beta} = 0$. The usual chi-square test

$$X^2 = \sum_{i=1}^{k} \left\{ (e_i - E(e_i))^2 \left[\frac{1}{E(e_i)} + \frac{1}{n_i - E(e_i)} \right] \right\}$$

(with $E(e_i) = n_i \frac{E}{N}$) for investigating an association between the dose and the response

rate can be decomposed into two parts:

$$X^2 = X^2_{linearity} + X^2_{slope}.$$

The statistic

$$\chi^2_{slope} = \hat{\beta}^2 \frac{\sum n_i(s_i - s_0)^2}{p_0(1-p_0)} \tag{3.1}$$

is the *Cochran-Armitage test* or *CA* test, which is under ($\hat{\beta} = 0$) chi-square distributed with one degree of freedom.

Tarone and Gart [58] showed that this test is a *C(a)* test i.e. a statistic derived from the partial derivatives of the log-likelihood function for testing H_0 against the alternative

$$H_1 : p_i = h(a + \beta d_i).$$

The function h is an arbitrary monotone function, twice differentiable and therefore the *CA* test is a test that can be applied under any monotone shape alternative and not only for a linear one. Whereas the *CA* test may be robust against departures from the linearity assumption, it is believed that in case of ordinal data this test would be an inappropriate approach [48].

Score selection

Obviously the disadvantage of this approach is the necessity to assign scores at the dose groups. The choice of the scores can have a substantial effect on the result of the test, especially when the data are unbalanced i.e. n_i are unequal. A popular solution is to use *midranks*, the mean values of ranks if one could have made a complete enumeration of the sample. However this approach is inadequate for unbalanced data, since it can conceal important differences between dose levels. Chuang-Stein and Agresti argue that the best way is to use scores that reflect the perceived distances between doses. However, the actual dose may not be available or not appropriate [48]. A possible guideline on selecting scores is to perform a *sensitivity analysis*: to use several scores and checking if the conclusions are similar.

To demonstrate the impact of the score selection, consider $K = 4$ dose groups with $d_i = (0, 2.5, 25, 250)$ and 50 animals per dose group. The responders are $e_i = (0, 4, 5, 6)$. Assigning the index as score ($s_i = (1, 2, 3, 4)$) the test statistic leads to a value of $CA = 5.204$ which gives a p-value of 0.023. If the actual dose levels are used as scores, the associated test-statistic $CA = 2.321$ indicates a non significant slope (p-value =0.128). Using the log(dose + 0.01) as scores, the corresponding test-statistic is 5.878, which is statistically significant (p-value = 0.015).

The overall X^2 test is independent of any score assignment and gives a value of 5.978. This indicates that $X^2_{linearity}$ is close to zero for the index and the log dose score assignment. However if the doses are used as scores, the value for $X^2_{linearity}$ of 3.661 is too high to reject the hypothesis of linearity.

Correction for skewness

Tarone [57] introduced a correction for the CA test in case of skewness. Additionally to scores, the result depends on the experimental design (combination of sample size per dose-group and scores). The type I error rate can be lower or higher depending on the design. The coefficient of skewness is

$$\gamma = \frac{(N-2E)\sqrt{N(N-1)}\sum n_i(s_i - s_0)^3}{(N-2)\sqrt{E(N-E)}(\sum n_i(s_i - s_0)^2)^{3/2}}. \tag{3.2}$$

The sign of this coefficient is determined by $m_3 = \sum n_1(s_i - s_0)^3/N$ and depends only on the experimental design. Tarone showed that when $\gamma > 0$ the CA test is liberal i.e. when H_0 is true, the test will reject it with probability higher that 5% whereas for $\gamma < 0$ the test is conservative. For correction, he proposed to replace the actual significance level z_α by

$$z_\alpha^{corr} = z_\alpha + \gamma(z_\alpha^2 - 1)/6. \tag{3.3}$$

In the aforementioned example the coefficient of skewness for the different score assignments is: for index $\gamma = 0$, for dose as score $\gamma = 0.259$ and for log(dose+0.01) $\gamma = -0.127$. The corrected p-values are 0.023 for index (no correction), 0.142 for dose and 0.014 for log(dose+0.01).

The Cochran-Mantel-Haenszel test

This is a test for conditional independence between $(K-1) \times 2$ contingency tables. It is a chi-square statistic with one degree of freedom. It pools information from comparing adjacent doses against one-sided alternative.

3.2.1 Tests that treat the response as continuous

The test $\bar{E}^2$ test (equation 2.9) and modifications of the t-test fall in this category. Although these tests may work quite satisfactory for ordinal response, they are inadequate for binary response; in most cases one has to assume excessively that p_i follows a normal distribution.

Rank tests

The most popular rank tests are the *Jonkeere-Terpstra* and the *Wilcoxon* sum rank test. One does not need to assign scores, but using rank tests this problem is not bypassed, since this family of tests uses midranks as preassigned scores.

Graubart and Korn [24] discussed methods that need score pre-assignment (like the *CA* test) and rank statistics concluding that rank test are not necessarily preferable.

They argue against the "illusory" advantages of the rank statistics and they state that they perform poorly in case that the data are not uniformly distributed in the dose-categories, which is true in many settings.

Consider for example the case where the two higher dose-categories differ a lot in terms of dose concentration but they correspond to a small proportion of the sample. The midranks will be similar for these two dose-categories and this closeness of scores is inappropriate. However, in this paper [24] no simulation study has been performed; they based their statements on an example. Note that isotonic regression is not a rank test, so the advantage of tests with pre-assigned scores that Graubart and Korn support, may not hold when compared to R test. Such a situation has not been considered in the paper.

Tests based on a model

The most popular approach is the proportional odds model. The well-known logistic regression model

$$logit(p_i) = \alpha + \beta s_i$$

is assumed. One can test for $\hat{\beta} = 0$ applying the logistic Likelihood Ratio test or the Wald test, based on the square of the ratio of the maximum Likelihood Ratio estimator of $\hat{\beta}$. Last, the score test based on the derivative of the log-likelihood at $\hat{\beta} = 0$ can be used.

The logistic regression approach is quite robust against departures from linearity and adjustment for other predictors can be made. However, there are important restrictions about the sample size N and n_i. Further, the score assignment is still present: the dose can enter the model either in its actual form or using scores. For the same example as before, the Likelihood Ratio test for logistic regression results is a p-value=0.147 when the actual dose levels are used as scores.

The isotonic Likelihood Ratio test (R)

This test – equation (2.10) – follows a weighted Chi-square distribution (equation 2.7). The main advantage of this test in the context of analyzing animal experiments is that **no scores need to be assigned**, and therefore the result is stable. Regarding the same example analyzed so far the Isotonic Likelihood Ratio test R=9.48 is derived, which is statistically significant with p-value < 0.001.

Algorithm 2 (*Univariate permutation test*)

- *Each animal is characterized by a pair of data* $(d_i, Y_i), i = 1, .., N$ *with* d_i *denoting the dose-group and* Y *is the status* ($Y = 0$ *without event and* $Y = 1$ *with event).*
- *This pair is broken up. Dose-level and status are combined per random allocation.*
- *Within each permutation* H_0 *is considered (equal risk in all dose-groups) and it is analyzed by the Likelihood Ratio test statistic* R. *That results in a set of values* R_{perm}.
- *The p-value is the probability that the result of a permutation is equal to the observed value of* R *estimated for the data set in hand* R_{obs}, *or to exceeds it:*

$$p - value = Pr\{R_{perm} \geq R_{obs}\} \tag{3.4}$$

If this p-value is less than the predefined significance-level, H_0 *is to be rejected and a dose-response relationship can be assumed.*

A problem associated with this test is that the theoretical distribution does not hold, as I will show later, in cases where the response probability is low, a situation

not rare in carcinogenic studies. The critical values assessed by equation (2.7) are lower that the critical values estimated by Monte Carlo methods. In this situation the researcher should apply an exact method.

One way to give the correct p-value is to perform a permutation test. Based on the observed margins (number of animals per dose-group and total number of events) a large number of permutations (e.g. perm = 10 000) is analyzed. The alternative to the permutation test is to analyze all possible combinations, applying the test to all of them, calculating the probabilities for each combination with a test-value equal to or greater than the observed one ($=R_{obs}$) and adding all these probabilities up. If the sum is less than the predefined significance level, H_0 is to be rejected. The probability to observe the combination $(e_1, e_2, e_3, e_4, ..., e_K)$ with $\sum e_i = E$ is

$$p(e_1, .., e_K) = \frac{\binom{n_1}{e_1}\binom{n_2}{e_2}...\binom{n_K}{e_K}}{\binom{K}{E}} \tag{3.5}$$

The number of all possible combinations depends on the total number of different dose-groups K. For example with $E = 10$ events and $K = 5$ dose-groups (the para-aramid data) altogether 1001 combinations are possible. I strongly recommend the use of exact critical values for small response probabilities.

Iso-chi-squared test (W)

Gautam et al. [21] proposed a modification of the usual Pearson's chi-squared statistic for a $2 \times K$ table based on isotonic regression. Pearson's X^2 test does not take into account the order categories of the different doses. Denoting the Pearson's correlation coefficient with r, one can write the X^2 statistic as

$$X^2 = max_{\forall\{s_1,...,s_K\}} Nr^2(s_1, ..., s_K)$$

Gautam proposed the following modification that restricted the test to the case of ordered proportions:

$$W = max_{\forall \text{increasing}\{s_1,...,s_K\}} Nr^2(s_1, ..., s_K) = max\{W_1, W_2\} \tag{3.6}$$

where W_1 is the Pearson's X^2 statistic calculated for table 3.1 after isotonic regression for e_i/n_i and W_2 is the Pearson's X^2 statistic calculated for table 3.1 after isotonic regression for $(n_i - e_i)/n_i$. For the distribution of the test statistic under H_0 they derived tabulated critical values for $3 \leq K \leq 10$ (see Appendix). However they suggest to use simulations for every special case and report the failure of the approximation. Gautam et al. propose to generate a large number of 2xK contingency tables keeping both margins constant, and to estimate the exact p-value as the proportion of the W_{perm} that exceed the observed value.

3.3 Simulation Study

3.3.1 Description

Three tests are selected to be compared to the Isotonic Likelihood Ratio R test:

- A score-based test: *CA* test
- A rank test: *Wilcoxon* rank sum test
- The iso-chi-squared test W by Gautam

This selection is based on the fact that the *CA* test and the *Wilcoxon* are the most popular among score-based and rank tests, whereas the W-test is a good established test among the PAVA-based tests, and the critical values are already computed [21].

Table 3.2: The coefficient for skewness γ if $K = 5$ and $n_i = 50$.

score	γ	Correction
index	0	No
dose	0.223	Upwards
log(dose+1)	0.002	No

With respect to the example presented later, 5 dose-groups with 50 animals per group are considered. The *CA* test was applied using three different scores i) the index s_i =order(d_i) ii) the dose $s_i = d_i$ iii) the logarithmic transformation of dose, s_i =log($d_i + 1$). All results are based on 10 000 replications. Regarding the *CA* test, the coefficient for skewness has been found different from zero for the situation in which the dose is used as score, otherwise *gamma* is zero or close to zero (table 3.2).

Recall that the tests used here are two-sided, to be consistent with the iso-chi-squared test *W*. Further, Cochran-Armitage trend test is generally reported with two sided p-value, in most of the publications. Isotonic regression (R test) assumes however only one trend at a time, thus can be thought as one-sided test. Therefore the following procedure is followed: without any a priori information about the trend, both isotonic and antitonic regression are fitted and the maximum value of the two Likelihood Ratio tests is taken. Using an one-sided test, the power of *CA* and *Wilcoxon* tests are potentially improved, but the R test has still greater power.

3.3.2 No dose-response assumption

a) Response probability=10%

The proportion of events is assumed to be 10%. On average 5 out of 50 animals in each group will give a positive response. The agreement between theoretical and

empirical distribution was good for both the *CA* test and the *W* test. All three assignments for *CA* where in agreement with the chi-square distribution (one degree of freedom).

The critical values proposed by Gautam (6.060) was not far from the result obtain from the simulation result (6.086). In contrast, the estimated 95% critical values for *R* was slightly higher than the theoretical. However in general, the empirical distribution of the test statistics follows the theoretical one (table 3.3).

Table 3.3: Simulations under H_0 (constant risk). Critical values for comparing $K = 5$ dose groups with $n_i = 50$ observations in each dose group, and response probability 10% and 4%. In first column the result from the theoretical distribution is depicted and in the second column the critical value estimated from 10 000 simulations. For the *W* iso-chi-squared test, the theoretical value corresponds to the approximation derived by Gautam.

Test	Critical value for significance level 5%		
	Theoretical	Estimated	
		10%	4%
CA(index)		3.833	3.851
CA(dose)	3.841	3.829	3.528
CA(log(dose+1))		3.837	3.755
R	5.048	5.248	5.730
W	6.060	6.086	6.277

b) Response probability=4%

The same situation but with less events was considered. The response rate was 4% i.e. 2 out of 50 animals are expected to develop the disease. Regarding *CA* there

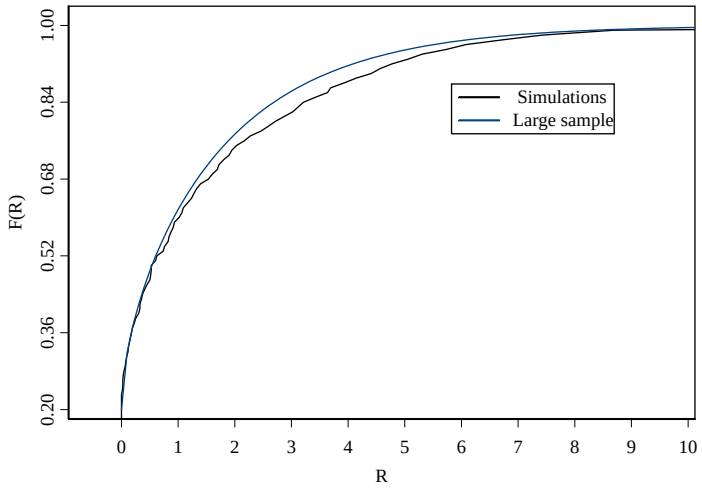

Figure 3.1: Theoretical (estimation from formula) and empirical (estimation from permutations) cumulative distribution for the Isotonic Likelihood Ratio test R (response probability 4%, $n_i = 50$ observations in each dose and $K = 5$ dose groups).

Table 3.4: Simulations under H_0 assumption (constant risk). Isotonic Likelihood Ratio (R) test: Critical values for comparing $K = 8$ dose groups and sample size 400 (50 in each dose group) when the response rate is 5%, 10% and 25%. The estimation has been accomplished using 10 000 simulations.

Significance level	**Critical values**			
	Theoretical	**Estimated**		
		5%	10%	25%
0.05	6.088	6.526	6.355	6.200
0.01	9.640	10.142	9.963	9.711

is again a good agreement: the cumulative distribution curves (the empirical and the chi-square) are shown together. In contrast, the estimated value (6.277) for W differs from the tabulated one. The distribution of R is far lower than the theoretical one as shown in figure 3.1. The estimated critical value for a significance level of 5% is 5.730 whereas the theoretical one is 5.048 (see table 3.3). If the response probability is even lower than 4%, this difference is increasing.

The same scenario was repeated with now $K = 8$ groups and response probabilities 5%, 10%, 25%. The results are presented in table 3.4. Again, a strong disagreement between theoretical and empirical distribution is observed. Note that in this simulation study it is assumed that the number of observations is equal in each group, situation which except for animal experiments is unlikely to occur in practice. The calculation of the level probabilities $P(l, K, w)$ becomes very cumbersome when the weights in each dose level are unequal. Moreover, when more than one explanatory variable is taken into account the Likelihood Ratio test does not follow any known distribution.

The conclusion from this analysis is: if the number of events is small, the investigator should better rely on the p-value estimation from the permutation's test, as described in algorithm 2. If the proposal of Robertson et al. [49] is followed, the test will be too liberal for small response rate since the true critical value is higher than the tabulated one. This gives a p-value far too small.

3.3.3 Increasing trend assumption

Under the assumption of increasing response probability, several different situations were investigated (table 3.5). The overall response rate was either 10% or 4%. The regression line follows sometimes steep and sometimes flat increasing pattern.

For the *CA* test, one of the score assignment leads to a perfect linear shape, and

shows the best power between the three *CA* representations. Although many researchers believe that departure from linearity do not effect the performance of the *CA* test, these results will show the opposite.

The Isotonic Likelihood Ratio test presents in all situations considered the highest power. In each case it has a power higher than the *CA* test even when a perfect linear shape is established. The critical values for R used to reject the H_0 are those estimated from the simulation study (5.248 for 10% rate and 5.730 for 4%).

The W test presents a very "dispersed" power: when the regression line is steep, this test yields a power value among the highest, and when the line is flat the power is one of the lowest. Similar properties are observed for the *Wilcoxon* test. Note that in this simulation study there are not objections against the use of the *Wilcoxon* test, since all n_i are equal.

In order to make a more general recommendation, the mean over all nine situations may be considered. The overall power is 60.59% for the isotonic regression, followed by *Wilcoxon* test using (51.17%). The log dose assignment leads to average power of 45.35%. The lowest power can be observed using the actual dose as score. From this analysis the use of the isotonic regression can be recommended.

3.4 Case study in tests for trend

The classification of man-made mineral fibres as carcinogenic is still under discussion for some type of fibres. Para-aramid, constitutes a particular interesting case and there is some controversy about its classification as carcinogenic or not [32]. Only one animal experiment is available, and this data set is presented in table 3.6.

To establish a causal relationship between exposure to para-aramid and tumor, the proof of a dose-response relationship (increasing effect with increasing dose) is

Table 3.5: Simulations under H_1 assumption (increasing risk). Comparison of the power under various situations considered among the *CA* test, the Isotonic Likelihood Ratio test R, the *Wilcoxon* and the iso-chi-squared W test. Five groups with 50 observations per group are assumed. The critical values used for R and W are estimated through permutations.

p_0	P(Y=1)	Linear	**Power %**					
%	%	relation	*CA* test with score			R	Wil-	W
		with	index	d_i	$\log(d_i+1)$		coxon	
10	2,6,10,14,18	index	74.35	66.92	72.87	92.63	87.97	84.98
10	6,8,10,12,14	index	31.84	24.56	30.70	40.49	30.94	28.56
4	2,3,4,5,6	index	19.94	17.74	19.57	28.22	18.66	17.57
10	4.8,4.9,5.8,14.9,19.8	dose	68.74	80.59	69.41	89.61	85.03	85.62
10	7.4,7.4,7.9,12.4,14.9	dose	31.67	31.99	30.99	40.53	32.26	31.51
4	0.4,0.4,1.1,7.4,10.8	dose	78.66	83.09	81.64	95.08	93.10	91.81
10	2.1,8.1,11.1,14.1,14.7	log dose	63.42	35.96	58.36	83.77	66.74	66.17
10	8.4,9.6,10.2,10.8,10.9	log dose	7.21	6.19	7.21	11.25	7.22	7.60
4	0.1,3.1,4.6,6.1,6.3	log dose	39.84	21.44	37.42	63.76	38.62	33.53
		Mean	46.19	40.94	45.35	60.59	51.17	49.71

Table 3.6: Data from the para-aramid study (IARC 1997). Tumor: adenoma, bronchido-alveolar without keratinising squamous-cell carcinoma.

dose(x 10^6 F/m^3)	0	2.5	25	100	400	$\sum$
no of tumors	1	1	1	4	3	10
no of animals	137	133	132	137	92	631

important. Since it is *a priori* known that the trend should be increasing, only one-sided tests are performed.

The decision concerning the acceptance of H_0 depends on the test used as good as on the method with which the p-value is obtained. If the *CA* test is applied, H_0 will be rejected if the indices are used as scores. The test-statistic *CA* of 3.819 is statistically significant based on the X^2 distribution (p = 0.026) using 10 000 permutations (p = 0.039) or analyzing all 1001 possible combinations (p = 0.033). If the other two methods of assigning the scores are used, the p-values are sometimes below and sometimes above 0.05. The result of the CA-test therefore is highly dependent on the way the test is performed.

Applying isotonic regression the hypothesis H_0 cannot be rejected. The p-value obtained from the large sample approximation is slightly above the significance level of 0.05 (p = 0.057). The p-value based on a sample of 10 000 randomly selected permutations and on all possible combinations (exact) are both above 0.05 (p-value = 0.110). The results are presented in table 3.7.

This result is in line with the results from the simulation study. If the event rate is small, the p-values obtained from the large sample approximation are misleading. The exact p-value as well as the p-value from the permutation test are showing no significant dose response relationship.

In order to investigate these differences in more detail some results of the isotonic

Table 3.7: Analysing the data from the para-aramid study: results.

		Dose assignment		
Test		index	dose	log(dose+1)
CA	X^2_{slope}	3.819	3.294	2.533
	p-value			
	X^2 distribution	0.026	0.035	0.056
	Permutations	0.039	0.049	0.047
	exact	0.033	0.049	0.048
Isotonic	R	4.779		
	p-value			
	X^2 distribution	0.057		
	10 000 permutations	0.110		
	exact	0.110		

regression and the *CA* test using the index as scores are considered. In general there is a good agreement between both test statistics. However some of the combinations have totally different outcomes. For example the combination (0, 0, 6, 3, 1) leads in the CA-test based on the index as dose assignment to a non-significant value of 2.210 ($\hat{\beta} = 0.006$). The isotonic regression yields a value of $R = 11.29$ ($p < 0.01$). There is also a difference to the two other methods of assigning scores. If the dose is used, the slope switches signs and turns negative, whereas the log (dose)-method gives a test value of = 3.481.

All other situations where both tests differ are similar. The proportion of events follows more or less an umbrella- or a U-shape. The risk is high in the middle dose groups and low at both ends or vice versa. Isotonic regression amalgamates the highest or lowest dose group together with the dose-groups in the middle which leads to an increased risk in the higher dose groups. Linear regression analysis

however leads to a more or less horizontal line ignoring an increase in the lower and middle dose groups. To summarize the results obtained in connection with the para-aramid example the CA-test seems highly vulnerable. It seems that the use of the dose as scores is not a good idea especially when the dose of the highest group is far from the rest.

3.5 Extensions: Adjustment for dose-induced mortality

In carcinogenicity studies, the size of the sample can change during the study due to mortality. This fact is sometimes of interest and the time to death is also taken into account, which can be thought as a confounding factor. At the end of the study the remaining animals are sacrificed and for each animal the dose level d_i, the status $Y = 0, 1$ and the time to death t_i are obtained. The cause of mortality can either occur from the tumor or as treatment effect (high doses may be more toxic).

Consider the notation of table 3.8 where the data are given in $t = 1, ..., T$ time-strata: Additionally note $E_{ij} = n_{ti}\frac{E_{t.}}{N_{t.}}$ the expected value in each cell. Usually the test

$$Z_G = T_G/V_G \tag{3.7}$$

with $T_G = \sum_{i=1}^{K} s_i(E_{.i} - E_{.i})$ and $V_G^2 = \sum_{t=1}^{T} \frac{E_{.t}}{N_{.t}}\frac{(N_{.t} - E_{.t})}{N_{.t}}\frac{N_{.t}}{N_{.t} - 1}$

is used (or modifications of this test). Asymptotically, it is normal distributed. Procedures based on life time tables and stratified logistic regression are possible. The following sections focus though to modifications for *CA* test and I will present survival adjustment for the isotonic tests.

Table 3.8: Notation used for survival adjustment in 2xK tables.

Dose-levels	d_1	d_2	...	d_K	$\sum$
Score	s_1	s_2	...	s_K	$s_0 = \frac{\sum s_i n_i}{N}$
Time strata 1	e_{11}	e_{12}	...	e_{1K}	$E_{1.}$
animals at risk	n_{11}	n_{12}	...	n_{1K}	$N_{1.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Time strata t	e_{t1}	e_{t2}	...	e_{tK}	$E_{t.}$
animals at risk	n_{t1}	n_{t2}	...	n_{tK}	$N_{t.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
proportions	$E_{.1}/N_{.1}$	$E_{.2}/N_{.2}$	...	$E_{.K}/N_{.K}$	$p_0 = \frac{E}{N}$

3.5.1 The Poly-3 test

In the usual Cochran-Armitage test all time strata collapse in one ($T = 1$) and Z_G can be accordingly calculated [12]. To adjust for survival times the weights

$$\omega_{ik} = \begin{cases} 1 & \text{tumor present at death} \\ (t_{ik}/t_{max})^3 & \text{else} \end{cases} \tag{3.8}$$

are defined.

Weights are assigned in a way which gives more weight to observations with events than to censored ones. For each dose group k the weights are defined as

$$\Omega_k = \sum_{i=1}^{n_k} \omega_{ik}. \tag{3.9}$$

For every dose-group $\sum \Omega_k = \Omega$, $p'_k = E_k/\Omega_k$, $p'_0 = E/\Omega$, $a_k = \Omega_k^2/n_k$, $u_{ik} = e_{ik} - p'_0\Omega_{ik}$ and $u_{.k} = \sum_{i=1} n_k u_{ik}/n_i$. The $Poly-3$ test is:

$$Poly-3 = \frac{\sum_{k=1}^{K} a_k p'_k s_k - (\sum_{k=1}^{K} a_k s_k)(\sum_{k=1}^{K} a_i p'_i)/\sum_{k=1}^{K} a_i}{\sqrt{C[\sum_{k=1}^{K} a_k s_k^2 - (\sum_{k=1}^{K} a_i s_i)^2/\sum_{k=1}^{K} a_i]}} \tag{3.10}$$

with $C = \sum\sum(u_{ik} - u_{.k})^2/(N-K)$.

3.5.2 The survival adjusted isotonic Likelihood Ratio test

The usual likelihood ratio test R can be modified to adjust for survival time. One has to estimate the *survival adjusted isotonic proportions* $p_k^{*'}$ by substituting the weights n_k in PAVA by Ω_k from equation 3.9. The survival adjusted Isotonic Likelihood Ratio test R_S is:

$$R_S = 2\sum_{k=1}^{K}\left[\Omega_k p'_k ln(\frac{\hat{p}_k^{*'}}{p'_0}) + \Omega_i(1-p'_k)ln(\frac{1-p_k^{*'}}{1-p'_0})\right]. \tag{3.11}$$

The critical values are assessed by a permutation procedure similar to those for binary response described in algorithm 2. Following the same idea, the test derived by Gautam will take into account mortality if the Ω_ks are used. For a thorough discussion on the several methods for survival adjustment see [4, 39].

3.5.3 Case study: Example for survival adjustment

The data used to demonstrate the methods are taken from a clinical study (EMIAT). This is a randomized trial aimed to compare active drug with placebo. Since no effect has been found for the drug regarding the survival time, the hole sample has been used to identify prognostic factors. The dataset includes the predictor variable LVEF (left ventricular ejection fraction). The outcome variable was death (all causes

Table 3.9: Data from the EMIAT study. The response is all causes death.

mean LVEF	19.873	28.597	34.222	38.484	46.436	62.933	$\sum$
Deaths	51	37	28	14	14	15	159
Patients	197	216	202	190	172	195	1172
Mortality	0.077	0.081	0.074	0.139	0.171	0.259	p_0=0.136
weight Ω	165.373	177.890	171.613	159.340	154.652	175.519	1004.384
isotonic p	0.077	*0.078*	*0.078*	0.139	0.171	0.259	

Table 3.10: EMIAT: The influence of left ventricular ejection fraction in mortality. Results from isotonic regression and poly-3 test.

Test		Dose assignment		
		index	dose^{-1}	log(dose)$^{-1}$
CA	$Poly-3$	2.238	2.369	4.765
	p-value	0.013	0.009	<0.001
Isotonic	R_S	37.494		
	p-value	<0.001		

mortality). The maximum survival time was 808 days. The LVEF percentage is antitonic related to mortality. A summary of the data set is depicted in table 3.9:

The results regarding *Poly-3* test vary according to the scores (presented in table 3.10). Using R_S test, the H_0 is rejected at 0.001 level.

3.6 Further discussion on tests for trend

The transformation of the dose in the *CA*-test using different score assignments can lead to totally different conclusions. Graubart and Korn [24] also explored

this problem in analyzing a study which investigated the association between some organ malformation (yes/no) and alcohol consumption of the mother. The consumption was categorized in 6 groups depending on the average drinks per day. They compared three scoring systems: midpoints, midranks and equally spaced ranks. The corresponding p-values differ dramatically between 0.017 (midpoints) and 0.286 (midranks). In conclusion the authors recommended assigning reasonable scores whenever possible.

Isotonic regression is also applied to this data: the test statistic $R = 6.102$ gives a p-value less than 0.001. This result is independent of any score assignment. The only assumption for this approach is the monotonicity of the response. But this assumption is also required for many other tests, the *CA*-test included. There is no additional assumption needed about the form of the relationship. Any monotonic transformation of the x-axes, in our example the dose-levels, leads to identical results. The power of this approach is higher than that of the optimal *CA*-test. The p-value for the isotonic regression can be obtained in using the weighted X^2 distribution. In situations with enough events per dose-groups, in the examples considered at least 5 events per group, the empirical distribution of the proposed test statistics follows approximately the theoretical one. Only in cases with a lower event rate or with highly unbalanced data a permutation test or simulations are recommended.

There are two options. First to generate permutations on a random base. About 10 000 replications seem sufficient to give p-values of adequate resolution. The other option is to consider all possible permutations, calculate the probability to observe this permutation and to add all probabilities for combinations with an equal test-statistics or a more extreme one (equivalent to Fisher's exact test).

Many researchers have been interested so far on evaluating tests for trend. Collings et al. [13] performed a comparative study between the *CA* test and the isotonic $\bar{X}^2$ test (section 2.4.2) using its theoretical distribution (equation 2.7) to assess p-

values. They concluded that i) the large sample approximation holds for both tests ii) the isotonic $\bar{X}^2$ test presents little advantage in power compared to *CA* test iii) the power of isotonic $\bar{X}^2$ test is significantly higher than the *CA* test's when the monotonicity assumption is violated by a downturn at a higher dose or eventually an umbrella shape.

Whereas points ii) and iii) are true in general, the aforementioned results should be considered under the following light: i) the conclusion about the adequacy of the large sample approximation was based on a simulation study where the lower response probability was $p_0 = 10\%$. This probability is too high – especially when analyzing carcinogenic agents – and as it is outlined in this chapter problems occur with response probabilities with less that $p_0 = 5\%$. Additionally, throughout the simulation study conducted by Collings et al. [13] the parameter K (number of dose groups) was maximally 5. With $K > 6$ the calculation of the critical value is practically impossible as mentioned in this paper. The authors in [13] argue that the isotonic $\bar{X}^2$ test performs better than *CA* test for small sample size, but they did not present any specific results about this conclusion. Finally, a simulation study was presented for the *Arcsine Isotonic Test* (section 2.4.2), and they show that this is a test which should never be used!

One can find a comprehensive review of the different order restrictions in Pedadda et al. [48]. They developed a nonparametric test based on the width of the interval between highest and lowest proportion. They defined

$$\hat{p}_i^* = max_{k \leq i} \left\{ \frac{\sum_{j=1}^{k} e_j}{\sum_{j=1}^{k} n_j} \right\} \tag{3.12}$$

and the test statistic

$$Z_{SO} = \hat{p}_k^* - \hat{p}_1^* \tag{3.13}$$

where the critical value is estimated through bootstrap. Although the lack of theoretical distribution, this tests is rather easy to apply and it can be extended to test

for *simple tree order*, *downturn at higher dose* and *umbrella shape*. They proved through a simulation study that this test is in general more powerful than *CA* and the gain in power is substantial when the shape presents serious departure from the liner shape (steep shape).

Mancuso et al. [39] proposed an isotonic version of the *CA* test which increases the power while controlling the type I error. This is simply the usual *CA* formula, where the isotonic estimators are used instead of the observed proportions. Bootstrap methods are used to assess the theoretical distribution of this test.

Leuraud and Benichou [37] used the isotonic test $\bar{X}^2 = \sum_{i=1}^{k} \frac{n_i(p_i^* - p)^2}{p_0(1-p_0)}$ and the weighted chi-square distribution to infer. They also considered two-sided test by taking the largest of the test values when increasing and decreasing trend is assumed. They compared this test to a t-test based on contrasts, to the Mantel-Haensel test and to the normal distributed Dosimeci-Benichou DB test. They conclude that the DB test presents the best results regarding power and type I error, they find the isotonic regression to be too powerful and the Mantel extension too erroneous, whereas the t-test presented medium performance.

4 REDUCED MONOTONIC REGRESSION

4.1 Reducing the number of solution blocks: why and how

Chapter 2 described how PAVA detects violators of the monotonicity assumption and builds the solution blocks by amalgamating adjacent observations together until there are no more violators. The changepoints that resulted from PAVA are estimated to eliminate violators and **they don't necessarily correspond to significant change in the response**. Moreover, it has been shown that the use of isotonic regression overfits somewhat the data whereas a model with fewer solution blocks fits better [2].

Thus some of the resulting solution blocks could be pooled together, especially these with few elements or those whose estimated values do not differ a lot from their neighboring blocks. Once the isotonic regression is fitted, the solution blocks that do not improve significantly the fit can be eliminated. Several proposals are

possible regarding the elimination procedure in order to improve the parsimony of the model.

An intuitive approach would be to compare all isotonic proportions – one by one – to identify those that do not differ significantly. Obviously, the researcher who applies this approach is called to control somehow the expense of the type I error which occurs from multiple comparisons [5, 8]. This is called the *Family wise Error* (FWE). For correction, Bonferonni inequalities are usually used [66]. ***In this chapter*** two alternative approaches are introduced. The first is *of my own design* and it is based on a sequence of Fisher tests. In the second method, I use the closure principal to eliminate the model. Both approaches reduce the isotonic model and control the FWE when the response is binary.

4.2 Method A: Elimination using a sequence of Fisher tests

4.2.1 Estimation

In order to compute the eliminated isotonic regression, two steps have to be considered: First, which solution blocks can be pooled together and, second, when the pooling procedure should stop. Several methods can be applied to answer these questions.

M. Schell [55] proposes an F-test when the response is continuous. The idea has been to calculate tests for the pairs of all sequential solution blocks, to select and eliminate the "weakest" change point that corresponds to the highest p-value. This idea cannot be appropriately extended for binary outcome variable, since in order to extract the F-distribution, one has to assume normal distribution for the pro-

portions. When the response is binary, P. Bacchetti [2] reduces the partial fitted functions in the additive isotonic model by comparing the change in the likelihood to a considered but ad hoc amount.

For binary response, I propose a reducing procedure based on Fisher's test for contingency tables: to identify the solution blocks that do not differ, one has to look at all 2×2 tables for the sequential solution blocks. The "pairs" that are not proven to differ significantly, are pooled together. The procedure stops, when all pivotal tables give significant p-values. To clarify the elimination procedure, the backward algorithm used to reduce the degrees of freedom in an one-dimensional isotonic regression is described.

Let the isotonic regression summarize the dose in L risk groups and estimated proportions $p_l^*, i = 1, ..., L$ and n_l observations falling at the lth solution block. The aim of the elimination procedure is to reduce the number of groups to S solution blocks ($S < L$) with respect to the outcome. The algorithm for elimination is described in algorithm 3.

Algorithm 3 (*Backward Elimination's procedure*)

1. *Construct all $L-1$ contingency tables for the adjacent solution blocks and calculate $L-1$ exact Fisher tests and their corresponding p-values*
2. *If all p-values $< \varepsilon^*$, where ε^* is a predefined significance level, then stop. Else, go to step 3*
3. *From the set of block-pairs resulting in a p-value $> \varepsilon^*$ select the one with the greatest p-value and pool it. Now the solution blocks are reduced by one. Go to step 1.*

Obviously the reduced isotonic regression depends on the choice of ε^*. For ε^*=1 the reduced isotonic regression is identical to the isotonic solution blocks whereas for ε^*=0 one gets a single solution block.

The use of ε^*=0.05 in the backward elimination will not yield an overall 0.05-level test as usual i.e. if the H_0 assumption of constant risk holds, the elimination procedure will yield more than a single solution block with probability greater than 5%. This is not surprising since the elimination procedure is based on a maximal selected p-value and we face a multiple comparisons problem [34].

Isotonic framework is poorly supported by asymptotic theory, especially in the case for binary response. The lack of theoretical solution forces us to use simulations to assess the value for ε^* that will yield to a significance level of 5%. We have to simulate random noise data (no association between dose and response) and then to assess in each data set the isotonic estimators and their reduced equivalents using ε^*=0. In each replication the p-value from the last Fisher test when only two solution blocks remain to pool is retained. Then the corrected ε^* is the 5 percent value from the distribution of all those "end" p-values. A similar approach has been used in [55] in order to correct the significance level in the F-test used to reduce a continuous response regression.

The elimination procedure can be extended to more sophisticated isotonic models as the additive one [2, 44]. In [2], an additive model for binary response is fitted and the need to reduce the degrees of freedom in the isotonic partial fitted functions is discussed. But the elimination procedure is accomplished by comparing the loss in the fit to an arbitrary amount. The reduced isotonic regression could be used instead, although that would potentially increase the computational complexity. Another application of reduced isotonic regression can be in isotonic-surfaces models, described in chapter 5.

4.2.2 Simulation study: Estimation of the corrected significance level in the backward elimination procedure

Design

A simulation study is conducted to explore the parameters that can influence ε^*. Different values for sample size ($N = 100, 200, 300, 600, 900$) and positive response probability ($p = 0.02, 0.05, 0.10, 0.15, 0.25$) were studied. The desired significance level was set to the nominal value $a = 0.05$. The algorithm described in the previous section was applied to 5000 samples from random noise data with predictor $X_i \sim U[0,1]$ and response $Y_i \sim B(p_0, N)$. The results are depicted in table 4.1.

Table 4.1: Estimation of ε^* based on 5000 simulations for different sample size and response rate. The overall significance level was 5%.

	N				
$\mathbf{p_0}$	100	200	300	600	900
0.02	0.0398	0.0218	0.0120	0.0117	0.0093
0.05	0.0188	0.0129	0.0097	0.0071	0.0066
0.10	0.0126	0.0089	0.0074	0.0064	0.0053
0.15	0.0103	0.0080	0.0074	0.0059	0.0057
0.25	0.0101	0.0077	0.0077	0.0055	0.0043

Results

The estimated ε^* decreases as long as the sample size and the response rate increase. While the decrease is sharp for small p_0, it flattens out with greater response rate. For this simulation study, it is assumed that the predictor variable X has no duplicates. The corrected significance level ε^* is slightly larger if there are any ties, and it becomes clearly greater if the variable X is categorical. For example, with sample

size 100 and event rate 10% the estimated ε^* is 0.0227 when X is in 4 categories i.e. about the double of the value when X is used as continuous (tabulated value 0.0126 in table 4.1). Moreover, when the iterative algorithm for isotonic matrix data is used instead of PAVA (see chapter 5), the values in table 4.1 no longer hold. Thus, it is not useful to estimate an approximate formula for ε^*, although that could potentially facilitate the elimination procedure, but simulations to estimate it for every specific data set are recommended.

4.3 Method B: Elimination using the closed test procedure

4.3.1 The closure principal

Another method to reduce the number of solution blocks and to control the increase of type I error can be accomplished via "closed testing". The problem of elimination can be formulated as follows: consider that isotonic regression resulted in L solution blocks determined by a set of cutpoints $\mathcal{C} = \{c_l, l = 1, ..., L-1\}$ and let the events $e_l = \sum_{i=1}^{l} Y_i$ out of n_l observations falling in the lth solution block and let p_l^* be the estimated isotonic proportions. With the notation $c[+1], c[+2]$ are denoted the greater cutpoint right next to c, the second greater and so on.

Given that the homogeneity is rejected for the L solution blocks by an omnibus test such as the Isotonic Likelihood Ratio test R, one can conclude that the risk in the last solution block is significantly higher than the risk in the first solution block. In the meanwhile, there can be one or more "jumps" in the risk to identify at some cutpoints in $\mathcal{C}$.

A possible approach is to proceed backward following an "edge to interior" pro-

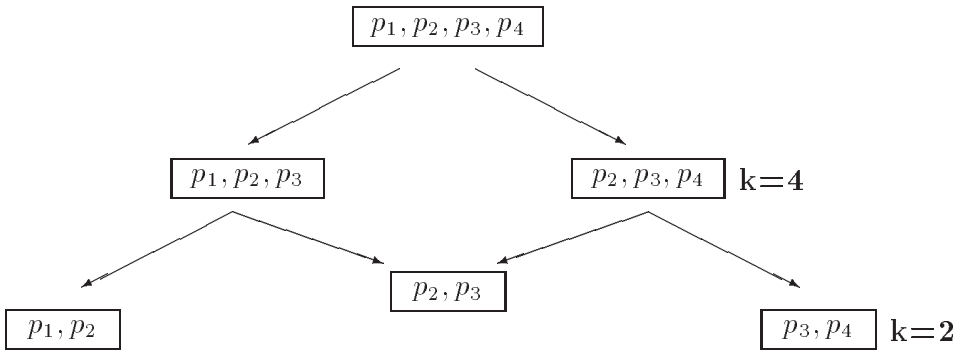

Figure 4.1: Example for closed testing.

cedure: start by considering the two marginal cutpoints c_1 and c_{L-1} and test the subsets that they define. That is equivalent to testing the hypotheses:

$$H_{2,L} : p_2 = p_3 = ... = p_L \tag{4.1}$$

$$H_{1,L-1} : p_1 = p_2 = p_3 = ... = p_{L-1} \tag{4.2}$$

In the next step each of the above two hypotheses is also split into two new hypotheses i.e. $H_{2,L}$ gives $H_{2,L-1} : p_2 = p_3 = ... = p_{L-1}$ and $H_{3,L} : p_3 = p_4 = ... = p_L$. The distinct resulting hypotheses are of course three, and split in their turn produce four new hypotheses. The procedure goes on until all $H_{l,l+1}$, i.e. all hypotheses about pairs of the neighboring solution blocks are generated.

Consider the example in figure 4.1 where $L = 4$. Once the omnibus hypothesis for (p_1, p_2, p_3, p_4) is rejected using the R test, the hypothesis $H_{1,3}$ for homogeneity in $(p_1, p_2, p_3\ )$ and $H_{2,4}$ for (p_2, p_3, p_4) are tested. Then the offsprings of $H_{1,3}$ and $H_{2,4}$

are to be tested. Every time that a non significant result is obtain, the tested set forms a block by pooling the observation.

Clearly the FWE increases with the number of tested hypotheses. A good control to this expense can be accomplished by considering a closed family of hypotheses. In the constructed family of nested hypotheses, not all of these hypotheses are to be tested: **$\mathbf{H_{l,l'}}$ is tested if and only if the hypothesis $\mathbf{H_{l,l'+1} \cap H_{l+1,l'}}$ is rejected**. The arrows in the figure 4.1 note the conditional testing. For later use, in every layer we correspond a value k which equal to the number of parameters tested within every hypothesis. In the layer right after the omnibus hypothesis we assign $k = L$.

In its simple concept, the closed testing procedure can be summarized as follows: start testing $H_{1,L-1}$ and $H_{2,L}$. For each hypothesis that has to be rejected proceed with the test of its offspring hypotheses. If $H_{l,k}$ is not rejected, retain every hypothesis $H_{l',k'}$ where l', k' are all possible combinations of the cutpoints between l and k. Each hypothesis H_l is tested using the isotonic R test in equation 2.10 (the alternative hypothesis is always the isotonic transformation).

The algorithm described above consist the basis of every closed procedure. Modification are imposed by the several ways to test the composite hypothesis. Every different method leads to a new closed testing approach. The resulting approaches differ in terms of power. In the next section I will present an approach appropriate for our goal.

4.3.2 Implementation of closed testing for eliminating the solution blocks

Consider the hypothesis of proportions from cutpoint l to k:

$$H_{l,k} : p_l = p_{l[+1]} = ... = p_{k[-1]} = p_k \tag{4.3}$$

for the homogeneity of *successive risks*. The set containing the original hypothesis ($l = 1, k = L$) and a hypothesis about two or more *disjoint* subsets of *successive risks* is the closure $\mathcal{H}_{\mathcal{L}}$ of the set of all hypotheses H of the form 4.3. The closure $\mathcal{H}_{\mathcal{L}}$ can be formed by taking the intersections of all H. The *closure principal* states that a hypothesis $H_{l,k}$ can be rejected at level a if it has been rejected at level a given that all hypotheses that "contain" $H_{l,k}$ have been rejected at level a. This principal yields powerful procedures.

Algorithm 4 (*Closed elimination procedure*)

- *Form all nested hypotheses of the isotonic proportions.*
- *Test the hypothesis* $H_{l,k} : p_l = p_{l[+1]} = ... = p_{k[-1]} = p_k$ *using the R isotonic test, where the p-value is estimated using permutations.*
- *Retain every hypothesis implied by any other hypothesis that has not be rejected*
- *Reject any hypothesis that is tested and rejected at level* $a_k = \dfrac{(k+1)a}{L}$
- *Reject any hypothesis that is tested and rejected at level a* ***and*** *any of the followings holds:*
  - *the complementary hypothesis has been rejected at the corresponding fraction of the pre-defined overall significance level a*
  - *the complementary hypothesis does not concern successive proportions*

For the case of dose-response relationship, a procedure to implement the closure principal has been proposed by Rom et al [50]. The idea was on the one hand to allocate portions of a to the different horizontal levels of the hypothesis schema and on the other hand to exploit an idea of Marcus et al. [40]: in a closed testing procedure the rejection of a hypothesis about a set of proportions depends on the

rejection of the hypothesis about the complement of this set. The procedure, modified for the current setting is as presented in algorithm 4. The value k takes a value equal to the number of proportions tested and the first two nested hypothesis are tested at $a_k = a$. Condition 3 ensures the important logical property of *coherence* (a hypothesis H_l is accepted if and only if one of the hypothesis that contains H_l is accepted). The second condition controls the error for the multiplicative effect whereas the last condition protects from the additive effect. The procedure will provide the dispensable cutpoints and controls the *family-wise error*. A formal description of the closure and proofs for the *family-wise error* control can be found in [7] and [40]. Modifications based on other approaches as for example the *Bonferroni-Holm min P-value* [30] or the *Westfall -Young Bootstrap* method [65] are also possible.

4.4 Selecting between full monotonic and reduced model

4.4.1 An approach based on bootstrap

Another crucial point in reduced isotonic regression is whether the reduced simpler model or its parent isotonic one should be used. The method presented here is general and can be used independently of the method applied to reduce the model (method A or method B).

Up to now, no distribution theory is available, thus one can use the AIC criterion to choose between simple and more complex models. Alternatively, one can apply a sort of parametric bootstrap [16]. The term "parametric" refers here to the idea that the data set in hand is assumed to be extracted from L populations whose distributions F_l are known, although here the underlying model (the reduced isotonic regression) is not parametric. To be more precise, it is claimed that under

the assumption that the reduced model is the correct one, the reduced $\hat{p}_l^*$s is an estimator for the parameter π of the binomial distribution $F_l \sim B(\pi_l, n_l)$. Following the notation of Efron [16] the measured function of interest for a bootstrap data set x^* is $\theta(x^*) = R(x^*) = D_{reduced}(x^*) - D_{isotonic}(x^*)$ with D denoting the deviance. The procedure is detailed described in algorithm 5.

Algorithm 5 (*Compare reduced to isotonic model*)

1. *Extract B simulated data sets x_j^* from the distribution $F : B(p_l^*, n_l)$.*
2. *In each x_j^* assess the isotonic and the reduced model and the corresponding deviances.*
3. *Assess $\theta(x_j^*) = D^*_{reduced_j} - D^*_{isotonic_j}$ for $j = 1, ..., B$.*
4. *If the 95% interval of $\theta(x_j^*)$s contains the observed value from the original sample $\theta(x_{obs}) = D^{obs}_{reduced} - D^{obs}_{isotonic}$, prefer the reduced model to the isotonic model, since the observed improvement in the likelihood for the isotonic model can be expected by its higher number of solution blocks.*

4.4.2 Simulation study: Comparison of full isotonic and reduced isotonic regression

Design

A limited simulation study is attempted in order to explore the benefits of using reduced isotonic regression instead of full isotonic model. To reduce the model, only the method A based on Fisher's test is applied. Two criteria were used: first, the *number of solution blocks* L as a measure of model complexity (L_{red} and L_{full}). Sec-

ond, as a measure of the model fit, while several criteria are possible, the *coefficient of determination* $\bar{R}^2$ as defined in [28, 45] for binary response models is used here:

$$\bar{R}^2 = \frac{R^2_{LR}}{R^2_{max}} = \frac{1 - e^{-LR/n}}{1 - e^{-D_0/n}} \tag{4.4}$$

where LR is the difference in the deviance between the reduced and the full model and D_0 the deviance for the null model. This measures the "variation explained by the model". The better model is the one with the greater $\bar{R}^2$ and the less complexity i.e. less L. Reduced isotonic regression decreases the model complexity, but it is also expected to reduce $\bar{R}^2$. With this simulation study we investigate if the decrease in the complexity is worth the loss in fit.

Three parameters have been studied: regression shape, sample size and R^2_{max}. Four regression lines have been analyzed:

a) linear LIN: $logit(p) = aX$
b) quadratic QUA: $logit(p) = aX^2$
c) hockey-stick HOK: $logit(p) = c + aX \cdot I_{\{X > median(X)\}}$ and
d) step function STE: $logit(p) = c + a \cdot I_{\{X > median(X)\}}$

where $I_{\{condition\}}$ is an index that takes the value 1 if condition is satisfied and 0 otherwise. I simulate these functions under sample size N=100, 300, 500. In each shape the parameter a has been determined such that the maximum coefficient of determination would be R^2_{max}= 0.3, 0.5, 0.7. That is actually equivalent to different assumptions about the positive response probability (about p_0 = 4%, 11%, 29% respectively).

Results

Regarding the complexity of the model, the number of L_{full} increased with sample size and R^2_{max} (range of mean value: 3.23-14.34). The same trend was observed for the number of the reduced L_{red}, but the variation was not very important (range of mean value 1.23-3.88). The elimination procedure reduces the number of solution blocks at about one third of the starting isotonic solution blocks. It is important to note that the fraction L_{red}/L_{full} becomes smaller with increasing sample size.

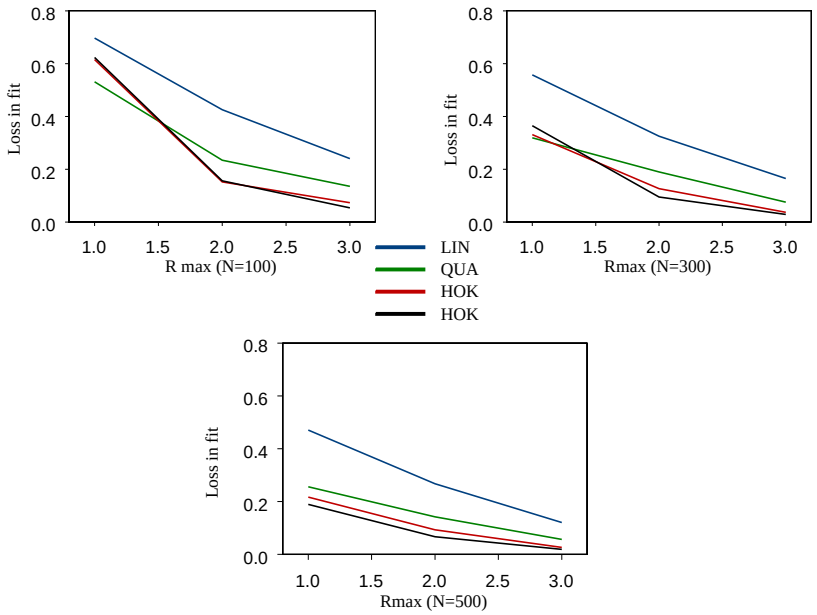

Figure 4.2: Results from simulation study: comparing monotonic regression and reduced monotonic regression regarding the relative loss in the fit as function of the sample size.

Figure 4.2 presents the results regarding the change in $\bar{R}^2$. On the x-axis is R_{max} and on y-axis is the relative "loss" in the fit ($= \dfrac{\bar{R}^2_{full} - \bar{R}^2_{red}}{\bar{R}^2_{full}}$) when the reduced model is used. Recall that it is optimal to have $\bar{R}^2_{full}$ and $\bar{R}^2_{red}$ as similar as possible,

i.e. small loss in the fit.

The difference between the coefficients of determination for the two models become smaller with increasing sample size. However, the influence of the maximum value of the coefficient is very important, or the response probability. While for smaller R^2_{max} the isotonic model has considerably better fit than the reduced model, its advantage is not important when R^2_{max}=0.7 (the reduced model reduces the coefficient of determination $\bar{R}^2$ only by 7%). Regarding the different underlying shapes, the linear regression presents clearly the worst tolerance on reducing the model, whereas the results of HOK and STE where the best for every R^2_{max}.

These findings, together with the results about the reduction of the model complexity, enables us to conclude that *when R^2_{max} is at least 0.5 and the investigator believes that the regression line is segmented,* ***reduced isotonic regression controls quite successfully the trade off between model complexity and fit.***

4.5 Case study in reduced monotonic regression

In the MAK study (table 1.1) instead of handling the variables dust and time separately, a "cumulative exposure" variable has been defined as their product. This new variable has been used as the unique predictor and the subsample for smokers have been analyzed. Modeling separately the effects of dust exposure and time requires either additive modeling or high dimensional smoothing leading to a more sophisticated model that will be described in the next section.

The isotonic regression for CBR probability depending on cumulative exposure (in mg/m^{3}years) and the reduced isotonic regression is presented in figure 4.3 together with pointwise confidence bands which are constructed by applying bootstrap under the assumption that the reduced isotonic regression is true. A smoothing spline with 6 degrees of freedom is also plotted. Its shape is in line with the reduced isotonic

regression. The ε^* value for the elimination procedure was estimated to be 0.0058, according to 10 000 permutations as described in section 4.2.1. The deviances for the fitted models are presented in table 4.2.

First the significance of the predictor was assessed by comparing the deviancies, which results to the Likelihood Ratio test R. The large sample approximation of the test proposed in [49] returns a p-value of $3.1\text{x}10^{-4}$. Additionally, a permutation test has been applied to assess significance. For this purpose 10 000 data sets have been extracted from the original one by permuting the endpoint variable. In each permuted data set the isotonic regression has been fitted and the achieved change in the deviance has been estimated. Then, the value of the observed Likelihood Ratio is compared to the 95th quantile of the empirical distribution, as estimated from the permutations. The permutation test results to a p-value of less than 0.001.

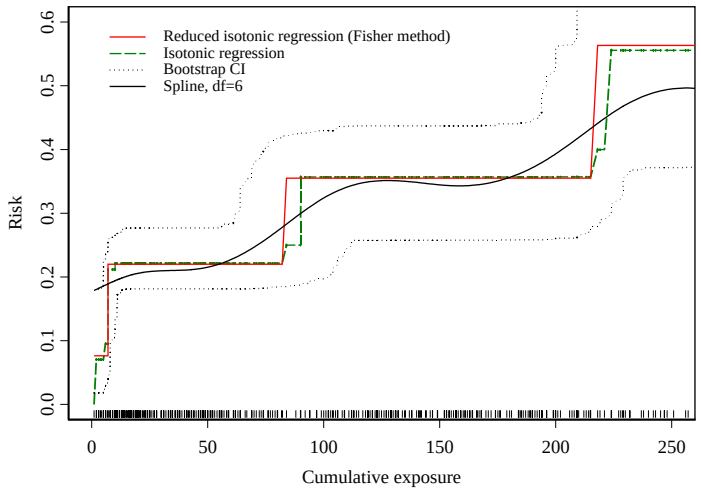

Figure 4.3: Full isotonic and reduced regression using a sequence of Fisher's test for the sample from Munich (smokers). The 95% confidence bands corresponds to the reduced isotonic regression.

The next step in the analysis is to choose between the isotonic model and its re-

Table 4.2: Deviancies and degrees of freedom of models applied in data set from Munich (smokers).

Model	**df**	**Deviance**	AIC
H_0	1	1058.17	1059.17
Full isotonic	11	983.06	1005.06
Reduced	4	988.33	996.33

duced version. First, the Akaike's information criterion is used. According to it, the reduced model fits better than the isotonic one (AIC reduced= 996.33, AIC isotonic= 1005.06). For the same purpose, a simulation study is conducted. Under the assumption that the reduced model is the true one, 10 000 simulated data sets are extracted. Then, one has to check if the observed difference in the deviance between reduced and isotonic model lies within the 95% confidence interval estimated from the simulations. There were 5532 data sets out of 10 000 resulting in a larger difference in the deviance than the observed one (988.33-983.06=5.27). That means that even if the reduced model is the correct model, such a large improvement in fit is likely to result from the greater complexity of the isotonic model.

The isotonic model is reduced using algorithm 4. After reducing the small solution blocks (those containing less than 1% of the sample), the procedure starts with the proportions $p_l = (0.066, 0.095, 0.212, 0.222, 0.355, 0.357, 0.563)$ and frequencies $n_l = (76, 42, 85, 415, 203, 28, 71)$. Testing $H_{1,6}$ and $H_{2,7}$ with R gives p-values < 0.001. The adjusted significance level is 0.05, thus the procedure goes on with testing $H_{1,5}$, $H_{2,6}$ and $H_{3,7}$. For this layer $a_k = 0.036$ and all three hypotheses are rejected. It is interesting to state the test for hypothesis $H_{2,4}$. The significance level is a_k=0.021 and the p-value 0.049. The hypothesis is rejected at a but not at a_k. Thus its rejection depend on the rejection of the complementary hypothesis $H^C_{2,4}$. However, $H^C_{2,4}$ does not concern successive means, thus $H_{2,4}$ is rejected.

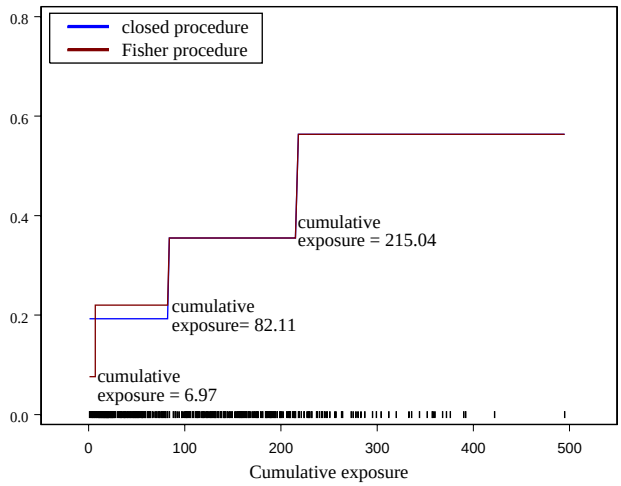

Figure 4.4: Reducing the isotonic model for Munich (smokers) using two different approaches.

The procedure goes on by rejecting every hypothesis at the corresponding significance levels until the fifth (and last) layer by testing all proportions two by two. The significance level is 0.0143. Hypothesis $H_{1,2}$ gives p-value 0.393 and can not be rejected, thus p_1 and p_2 are pooled together. The p-values for $H_{2,3}$ and $H_{3,4}$ were 0.072 and 0.514, thus these hypothesis are retained. All proportions from p_1 to p_4 form a block. $H_{4,5}$ is rejected (p-value<0.001) and $H_{5,6}$ is retained. Testing $H_{6,7}$ gives a p-value 0.028. The complementary $H^C_{6,7} = H_{1,5}$ has been rejected at level $a_k = 0.036$, and so is $H_{6,7}$. Finally the following solution blocks and frequencies are obtained: 0.193 (618), 0.355 (231), 0.563 (71). Reducing the model using the two different methods described in this chapter gives similar result as outlined in figure 4.4. I believe that these two methods are equivalent and would give similar results, as it is the case in this application. However a comparative simulation study remains to be done in order to conclude about the equivalence or not between these two reducing approaches.

5 MULTIDIMENSIONAL MONOTONIC MODELS

5.1 The multiple regression setting

By now, only problems related to one explanatory variable have been discussed. In this chapter isotonic regression for multiple regression data will be introduced. Traditionally two approaches are used when more than one predictor variable is taken into account: either to assume that the predictors contribute additively to the outcome (additive model), or to fit a general regression surface. This second option is not very popular among biostatisticians, since using standard techniques to fit the surface (as for example a bivariate kernel smoother) causes problems either with the estimation or the interpretation of the result [29].

Assume order restrictions regarding the effect of more than one explanatory variable or of a subset of them. Without loss of generality all predictor variables are expected to have an increasing effect to the outcome. The methodology described in this chapter can be easily modified to handle situations where different trends are assumed among the variables.

Two multivariate approaches are possible: the *additive isotonic model* and the *isotonic-surfaces model*. First, the *additive isotonic model*, as established by Bacchetti [2] and Morton-Joneset al. [44] is presented. This model, unlike *isotonic-surfaces model*, has been extensively described in terms of its statistical properties.

This chapter focuses on the alternative model, the *isotonic-surfaces model*. This is easy to apply and interpret, unlike the majority of surface models. This approach is based on a proposal by Robertson [49]. Starting from his idea I will outline the limitations and advantages of this approach and I will discuss some special properties of the estimation algorithm. Further I will explain how to implement the reducing procedure highlighted in the previous section to improve parsimony of the isotonic-surfaces model. Finally, I will introduce a multivariate version of the isotonic test for trend presented in chapter 3 for overall and partial significance based on this model.

Setting: Consider N observations on an outcome (or dependent variable) Y denoted by $\mathbf{y} = (y_1, y_2, ..., y_N)^T$ measured at N designed vectors $\mathbf{x}^i = (x_{i1}, x_{i2}, ..., x_{iP})$, assuming P predictor (or independent) variables X_j, $j = 1, ..., P$. Modeling the dependence of Y on $X_1, ..., X_N$ has three principal goals:

- to *describe* for learning more about the process that produces the outcome,
- to *infer* i.e. assess the relative contribution of each variable,
- to *classify* for binary response problems.

5.2 Additive isotonic model

The additive isotonic model starts from the assumption that the risk (response) does not decrease as long as any of the predictors increases, and extends GAM (Generalized Additive Models) [29] by letting isotonic transformation act like a

"smoother". The local scoring algorithm usually used in GAM is replaced here by PAVA and the contribution to the risk of each isotonic variable is a non decreasing step function. The additive isotonic model with P isotonic predictors takes the following form:

$$h(p_i) = \sum_{j=1}^{P} \tilde{\phi}_j^*(x_{ij}) \tag{5.1}$$

where h is a link function and $\tilde{\phi}^*$ denotes a P-dimensional isotonic function $\tilde{\phi}^* = (\phi_1^*, \phi_2^*, ..., \phi_P^*)$. The estimation proceeds via the backfitting algorithm [29]. To clarify that, the procedure is shortly presented in algorithm 6:

Algorithm 6 (*Backfitting for additive isotonic models*)

1. *Initialize* $\tilde{\phi}^* = (\hat{\phi}_p^*)$ *to be* y_i/w_i *for all* $p = 1, ..., P$, *where* w_i *are weights.*
2. *Set* $p = 1$ *and estimate* $\phi_p^*(x_{ip})$ *holding* ϕ_k^* *fixed for* $k \neq p$ *using PAVA.*
3. *Repeat step 2 for all* $p = 1, ..., P$ *and at the end assess the deviance for the first loop*
4. *Repeat 2 and 3 until the change in the deviance for two successive loops becomes very small.*

The estimation of ϕ_p^* in step 2 uses PAVA as follows: The observations are pooled together to "correct" for violators as described in algorithm 1 and builds L solution blocks denoted by l_k, $k = 1, ..., L$. Then $\hat{\phi}_p^*$ can be found from the equation:

$$\sum_{i \in l_k} y_i = \sum_{i \in l_k} w_i h(\hat{\phi}_p^* + \sum_{t \neq p} \hat{\phi}_t^*). \tag{5.2}$$

The binary response problem revisited

In this case it is $y_i = p_i$ the response probability, the link function h is the logit (or more seldom the probit) and $w_i = n_i$ the number of observations. Equation (5.2) takes the form

$$\sum_{i \in l_k} n_i p_i = \frac{exp(\sum_{i \in l_k} \hat{\phi}_p^* + \sum_{t \neq p} \hat{\phi}_t^*)}{1 + exp(\sum_{i \in l_k} \hat{\phi}_p^* + \sum_{t \neq p} \hat{\phi}_t^*)}. \tag{5.3}$$

Semi-parametric model

The model described in equation 5.1 is non-parametric in nature. However, it is often the case that some of the covariates need to enter the model linearly. The additive isotonic model can be transformed to a semi-parametric model of the form of equation 5.4 where k isotonic predictor variables and s linear predictor variables are assumed:

$$h(p_i) = \sum_{j=1}^{k} \tilde{\phi}_j^*(x_{ij}) + \sum_{j=1}^{s} \tilde{\beta}_j x_{ij}. \tag{5.4}$$

Recall that **the degrees of freedom of each isotonic term are equal to the number of solution blocks** i.e. the "steps" in to which the isotonic transformation ends up. Once the model is formulated, the Isotonic Likelihood Ratio test can be used to infer about the explanatory variables and to test the accuracy of the transformation. The large sample approximation described by Robertson et al.[49] does not hold here, so permutations need to be applied again in finding the empirical distribution of the test. The procedure would be similar to this described in section 5.4.2 for the isotonic-surfaces model. Additive isotonic models can prove to be a useful tool for exploratory analysis, since they speed up the checking of variables as possible predictors by rejecting those in whom the best isotonic transformation performs poorly.

5.3 The isotonic-surfaces model

As already discussed, isotonic regression shows characteristics of a scatterplot smoother. This consideration arises the question: how can isotonic regression be extended to "smooth" a plot when one sets restrictions over multiple axes? In other words, following the same logical pattern as in the univariate case, how can isotonic estimates be produced in a three-dimensional (or even higher) plot?

Before addressing this point, let us state some necessary definitions and considerations. Consider for simplicity only two dimensions, i.e. two explanatory variables ($P = 2$) X_1, X_2 and each variable has N_1 and N_2 *ordered* distinct values. The outcome variable is the response probability p_{ij}. Imagine the data in the form of a matrix M (equation 5.5) with dimension $N_1 \times N_2$.

$$M = \underbrace{\left[\begin{array}{cccc} \hat{p}_{11} & \hat{p}_{12} & \cdots & \hat{p}_{1N_1} \\ \hat{p}_{21} & \hat{p}_{22} & \cdots & \hat{p}_{2N_1} \\ \vdots & & & \vdots \\ \hat{p}_{N_21} & \hat{p}_{N_22} & \cdots & \hat{p}_{N_2N_1} \end{array}\right]}_{\text{Variable 2}} \Bigg\} \text{Variable 1} \tag{5.5}$$

Cell (i, j) contains the outcome $\hat{p}_{ij}$ of the individual corresponding to the i-th observation for the first variable and the j-th observation for the second one.

Definition 5.1 (Partial order for matrix) *A matrix M is isotonic with respect to the partial order if and only if the elements $\hat{p}_{ij}$ of M fulfill the restrictions $\hat{p}_{ij} \leq \hat{p}_{kl}$ for every $i \leq k$ and $j \leq l$.*

This setting can be easily extended to more than two dimensions. For three dimensions the data are in the form of an array. The definition of an isotonic array comes in analogy to the isotonic matrix: the estimated proportions should be in non-decreasing order over rows, columns and layers.

In contrast to model 5.1, the predictors here do not contribute additively to the response, but they rather interact. The model in the case of partial order for P explanatory variables takes the form:

$$p_i = \Phi^*(\mathbf{x}^i) = \Phi^*(x_{i1}, x_{i2}, ..., x_{iP}) \tag{5.6}$$

where Φ^* is the isotonic transformation over P dimensions and x_{ip} the i-th observation for the p-th predictor variable.

5.3.1 Estimation algorithm

Let us return to matrix (5.5). To assess the isotonic estimates $\hat{p}^*_{ij}$ an optimization procedure is needed, assuming order restrictions for two predictors. That can be accomplished by applying the *Iterative Algorithm for Partial Order* (IAPO). This is a version of PAVA algorithm which can be roughly thought as follows: The data are iteratively "projected" in every dimension and the PAVA procedure is applied till convergence is reached.

Algorithm 7 (*The Iterative Algorithm for Partial Order*)

1. *Let M^{*1} denote the isotonic regression of M over rows. Let $R^1 = (M^{*1} - M)$ be the first set of row increments.*
2. *Let M^{**1} denote the isotonic regression over columns of $M + R^1$. Call $C^1 = M^{**1} - (M + R^1)$ the first set of column increments.*
3. *At the beginning of the n-th cycle M^{*n} is obtained by isotonizing $M + C^{n-1}$ over rows. The n-th set of row increments is defined by $R^n = M^{*n} - (M + C^{n-1})$.*
   Next, obtain M^{**n} by isotonizing $M + R^n$ over columns.

Theorem 5.1 (Convergence for the Iterative Algorithm for Partial Order) *Both M^{*n} and M^{**n} converge to the isotonic regression M^{***} with respect to the partial order.*

The result of a two dimensional isotonic regression can be *visualized* as a surface that is non decreasing as long as any of the predictors increases. The algorithm combines both explanatory variables in L constant risk groups (the level sets or solution blocks), and therefore each step in the response variable corresponds to a specified bivariate group for the predictor variables.

Computational problems related to the Iterative Algorithm for Partial Order

Theorem 5.1, that assures convergence, would not necessarily hold if the data array has a lot of zero-weighted cells, since in this case the individual row and column isotonic regressions would not be uniquely determined. In order to illustrate this, consider the following example taken from the MAK study: The entries in the data matrix are the event probabilities as a result of the exposure time in total inhalable dust and its concentration. An extract of the data matrix (events/number of exposed) is presented in table 5.1.

The column smoothing and the row smoothing do not result in a common isotonic matrix. This problem may be avoided by using a procedure proposed in [49]: (1) remove the zero-weight cells and disregard all orderings which involve these cells (2) carry out the row-column isotonic regression (3) insert any values in the zero-weight cells. However this alternative is not always efficient. Applying this procedure to the matrix there is no way to fill in the cells with the appropriate values (table 5.2). One efficient approach in order to avoid this problem is, before starting the isotonic regression procedure to eliminate the zero weighted cells by grouping.

Table 5.1: Extract of the original data matrix, example of no convergence (original data).

$\frac{mg}{m^3}$	**years**		
	35-40	40-45	45-50
5,5-6	0	0	0
6-6,5	1/1	0	0
6,5-7	0	1/1	0
7-7,5	0	0	0
7,5-8	0	0	0
8-8,5	0	0	0
8,5-9	0	1/2	2/4
9-9,5	0	0	0

Table 5.2: Extract of the isotonic matrix, example of no convergence (isotonic estimation).

$\frac{mg}{m^3}$	**years**		
	35-40	40-45	45-50
5,5-6	na	na	na
6-6,5	1	na	na
6,5-7	na	0.475	na
7-7,5	na	na	na
7,5-8	na	na	na
8-8,5	na	0.5	0.5
8,5-9	na	na	na
9-9,5	na	na	na

An other, more complicated algorithm described by Gebhard [22] can be also used to estimate isotonic surfaces.

5.3.2 Advantages and limitations of the isotonic-surfaces model

The monotonic-surfaces model captures interactions between the explanatory variables, a feature that the additive isotonic model does not provide. Another advantage of model 5.6 compared to model 5.1 is connected to the implementation of the reducing procedures described in the previous chapter, and the procedure for selecting thresholds: they are easier and straightforward to implement in an isotonic surface model.

Theoretically, the algorithm for isotonic-surfaces can be extended to more than two variables. For a third factor in τ-ordered levels the result would be a sequence of τ-isotonic surfaces each of them lying above the previous one or touching each other. However, in practice if more than three isotonic predictors need to be included in the model, the use of this approach is not recommended due to its great computational complexity. As outlined in the previous section, a main problem arising from this algorithm is that the convergence is not guaranteed in case the data contains many zero-weighted cells. Therefore the predictor variables need to be in pre-selected groups. It is expected that their choice can affect somewhat the results because of the decrease in the number of the candidate changepoint locations. However, if the categories are selected at a base of many and thin quantiles, the influence is minimized and it can even vanish.

5.4 Multidimensional extension of the isotonic test for trend

5.4.1 Test for isotonic matrix: an asymptotic-based proposal

The estimation approach described in 5.3.1 is appropriate for every response type y belonging to the exponential family described in 2.2. A test – described in [49] – that can be used to infer for the predictor variables is shortly described here. The approach used in the univariate analysis can be extended to take additional possible explanatory variables into account. The following test can be used if one is interested in the effect of a single covariate in case where two predictor variables are included in the model:

Consider a $R \times C$ matrix whose entries are estimated parameters (one parameter) for the exponential family. The usual model with interaction, is

$$y_{ijk} = \theta + a_i + b_j + c_{ij} + e_{ijk}$$

where $i = 1, 2, ..., R$, $j = 1, 2, ..., C$, $k = 1, 2, ..., w_{ij}$, and w_{ij} the weight of each cell. The distribution parameter is denoted with θ, $a_i = \theta_{i..} - \theta_{...}$ the residual of the i-th row mean, $b_j = \theta_{.j.} - \theta_{...}$ the residuals of the j-th column mean, and c_{ij} the difference between the parameter estimated at cells and the overall parameter. Let $\theta^*_{i..}$ be the isotonic estimation of $\theta_{i..}$, assessed using the PAV algorithm. Suppose that the column variable has a given isotonic effect to the response.

Then the isotonic effect of the row variable is tested. The idea is that if the row variable has no significant increasing trend, then the isotonic estimation of the row margins $\theta^*_{i..}$ will not differ significantly from the overall mean $\theta_{...}$.
This is equivalent to the following hypothesis:

$$H_0 : a_1 = a_2 = ... = a_R = 0$$

versus

H_1: $(a_1, a_2, ..., a_R)$ is isotonic with respect to a quasi order over 1,2,..., R.

Denoting $\hat{a}_{H_1} = (\theta^*_{1..} - \theta_{...}, \theta^*_{2..} - \theta_{...}, ..., \theta^*_{R..} - \theta_{...})$ and $\hat{a}_{H_0} = (0, 0, ..., 0)$ then analogously to the test in the univariable approach the following test is defined:

$$T_{01} = \frac{C \sum_{i=1}^{R} w_{ij}(\theta^*_{i..} - \theta_{...})^2}{\frac{C \sum_{i=1}^{R} w_{ij}(\theta^*_{i..} - \theta_{...})^2 + \sum_{ijk}(\theta_{ijk} - \theta_{ij.})^2}{N \cdot R \cdot C}}. \tag{5.7}$$

This test statistic should approximately follow the weighted Chi-square distribution in equation 2.7. It can be used when the response is binomial, considering that the proportion follows approximately the normal distribution having mean p_{ij} and variance $\frac{p_0(1-p_0)}{n_{ij}}$. This is a very common approximation but it holds only if the proportions are around 50%. So I propose a more adequate approach: when the response is binary and when the predictor variables are more that two a test can be accomplished via permutations.

5.4.2 Multiple permutations test

The significance of any predictor included in a model $\mathcal{M}_{X_p, p=1,...,P}$ with P predictor variables can be assessed by the Likelihood Ratio (LR) test. To infer for the subset of S out of P variables $X_j, X_{j+1}, ..., X_{j+S}$, the change in the deviance is assessed when the variables $X_{j,...,j+S}$ are excluded from the model:

$$LR = \text{Deviance}(\mathcal{M}_{X_p, p \notin \{j, j+1, ..., j+S\}}) - \text{Deviance}(\mathcal{M}_{X_p, p=1,...,P}). \tag{5.8}$$

The first part of the difference may contain no variables at all (null deviance) and the test assesses the *overall* adequacy of the model. Otherwise, the tested significance is *partial*.

In generalized linear models context, the LR test follows a chi-square distribution with S degrees of freedom. For isotonic-based models, there is no known large sample approximation for the distribution of LR test, so again permutations have to be used to calculate the critical values for both overall and partial significance.

For instance, the partial significance will include only one variable X_j. The critical value of the LR test can then be assessed via *conditional* permutations. The term "conditional" refers – say in a two-dimensional case – to the following: given the events distribution to the solution blocks of one predictor X_1 to be true (the likelihood estimated, say, at the columns), the probability to have the observed distribution at the cells is estimated (overall likelihood for the model $\mathcal{M}_{X_1,X_2}$).

Algorithm 8 (Conditional permutations for partial significance)

Each response $Y_i = 0, 1$ *corresponds to a vector* $\mathbf{x}^i = (x_1^i, x_2^i, ..., x_P^i)$. *To test the effect of the j-th predictor adjusted for the remaining* $P - 1$ *predictors:*

- *Split the vector* $\mathbf{x}^i$ *at the j-th variable*
- *Create a data set* x_b^* *by randomly combining* $(Y_i, x_1^i, ..., x_{j-1}^i, x_{j+1}^i, ..., x_P^i)$ *and* x_j^i
- *Compute isotonic regression and the corresponding Likelihood Ratio test* $LR(x_b^*)$
- *Repeat steps (2) and (3) B times where B is at least 5000. The critical value is estimated as:*

c-value(j,B)=(1-a)% quantile of $(LR(x_1^*), LR(x_2^*), ..., LR(x_B^*))$

Of course one can test all predictors at once if so desired by randomly combining Y_i to $\mathbf{x}^i$ and then following the same procedure as described above. As in the

univariate case, one can construct confidence surfaces for the estimates, applying parametric simulations under H_0 or H_1. The width of these intervals can yield useful information. In cases where a non significant result is obtained, confidence bands can help us to distinguish between statistical and substantial consistency in the risk. In case of significant result, one can simulate under the assumption that the isotonic estimates are true. That would be equivalent with a test T_{12} (see chapter 2): the fit of the isotonic transformation is assessed, compared to any other possible shape.

5.4.3 Comments on multivariate tests for trend

Several proposals have been made so far in the literature regarding multivariate tests for trend. The Mantel-extension test can be modified to handle more than one categorical variable, but it is expected to present the same drawbacks as in the univariate case. The logistic regression offers an alternative, but the linearity assumption remains a constraint. The T-contrast test and Dosemeci-Benichou test described in [37] can be extended to more than one variable, but as mentioned in the paper, more work remains to be done on this area.

Regarding isotonic regression I proposed a test based on the isotonic-surface model to deal with more than one explanatory variable. It is a Likelihood Ratio test where the critical value is computed using permutations, and can be overall (testing all variables included in the model) or partial (assessing the influence of a variable adjusted for the others).

The main problem remains the restriction to maximal three predictors to be used. Another equivalent approach can be accomplished by applying additive isotonic models and thereafter the same overall and partial permutation procedure. A comparative study of these multivariate tests could provide useful information. However, it is believed that the main characteristics of each test as discussed in chapter 3 remain the same, independently of the number of variables taken into account.

5.5 Reducing the isotonic-surfaces model

Although the closed testing procedure could also have been used, within this chapter the approach based on Fisher's sequence is discussed (algorithm 3 in page 53). Each solution block is compared to its neighboring using the Fisher test, and the solution blocks that do not differ a lot are pooled together until the estimated ε^* level is achieved. The estimation of ε^* as well as the procedure applied to choose between the isotonic and reduced model are as described for the univariate case.

At this point some important remarks about reducing multidimensional isotonic regression need to be made. Consider an outcome variable of length N and two scenarios about the explanatory variable: in the first case, (i) one explanatory variable is assumed with K distinct observations whereas in the second case (ii) two explanatory variables in also K but *bivariate* groups are assumed i.e. the dimension of the data matrix is K. Clearly, *the IAPO (algorithm 7) will result in more solution blocks that the univariate PAVA (algorithm 1).*

Recall then that the estimation of ε^* depends on the number of starting comparisons i.e. the number of isotonic level sets: the more pivotal tables are analyzed, the smaller value for ε^* we get. Thus, it is obvious that when algorithm 7 (IAPO) for partial order is used instead of algorithm 1 (PAVA) on the estimation of ε^*, **the obtained values will be much lower that these displayed in table 4.1**, for the same number of starting isotonic groups.

The elimination procedure can be implemented in the additive model as well. In [2], an additive model for binary response is fitted and the need to reduce the degrees of freedom in the isotonic partial fitted functions is discussed. But the elimination is accomplished by comparing the loss in the fit to an arbitrary *adhoc* amount, and the estimation has been accomplished without backfitting. Alternatively, the reduced isotonic regression could be used to estimate ϕ_p^* (step 2 in algorithm 6, page

70), although that would potentially increase the computational complexity. It is intuitively simpler to combine the elimination procedure with an isotonic-surfaces model.

In analogy to the univariate case, reduced isotonic surface can be very useful, leading to a "compromising" choice between goodness of fit and model complexity.

The reduced multidimensional isotonic regression model yields a simple and interpretable classification of the P predictor variables in P-variate groups with respect to the outcome variable by detecting cutpoints under the assumption of monotonicity.

5.6 Extension of the additive isotonic model for interactions

The surfaces-model is useful in cases where the predictors are suspected to interact. For the additive isotonic model, no proposal is available by now about how interactions can be included in the model.

A simple but approximate approach is to examine the residuals for interactions. A natural step is to plot the residuals and to scan them: smoothing the scatterplot on the use of an isotonic-surface can alert for the presence of interactions. That can also be seen as a first step in a backfitting algorithm for fitting the model:

$$h(p_i) = \phi_1^*(X_1) + \phi_2^*(X_2) + \phi_3^*(X_1, X_2)$$

For the case of binary response one can use the deviance residuals

$$\delta(y_i, p_i^*) = \pm 2[y_i log(\frac{y_i}{n_i p_i^*}) + (n_i - y_i) log(\frac{n_i - y_i}{n_i(1 - p_i^*)})]^{1/2}$$

where $\pm$ is the sign of the quantity $y_i - n_i p_i^*$.

Table 5.3: The deviance of isotonic and reduced models.

Model	**Deviance**
H_0	1058.17
Isotonic (Time)	1008.65
Isotonic (Dust)	1025.65
Isotonic (Dust and Time)	959.87
Reduced isotonic	976.45

5.7 Case study in multivariate analysis

On analyzing the MAK study, the time since first exposure (in years) was also taken into account, and a two dimensional model (5.6) was fitted for the smokers. The amount of dust was categorized in 17 quantiles and the time since first exposure in 10 quantiles. As noted in the methodology part, this choice can affect the results. It would be more adequate to construct more that 17 quantiles, but the data are not that precise. The result of isotonic regression is depicted in figure 5.1.

The permutations procedure was applied to perform an overall test for the model and conditional permutations as described in section 5.3 to assess the statistical significance for the effect of dust given the effect of time. For the overall test (dust and time) the p-value for the observed value T_{01}=98.30 (see table 5.3) was less than 0.001 based on 5000 permutations. The conditional test for the improvement in the fit after entering the dust in the model (1008.65-959.87=48.78) resulted in a p-value = 0.002.

Since the result is significant given the effect of time, one has to simulate under the isotonic estimates, to assess the confidence surfaces. The result (figure 5.2) shows the width between the lower and upper surface: it is not very wide except perhaps the dust groups $0.2 - 0.35\frac{mg}{m^3}$ and the last one.

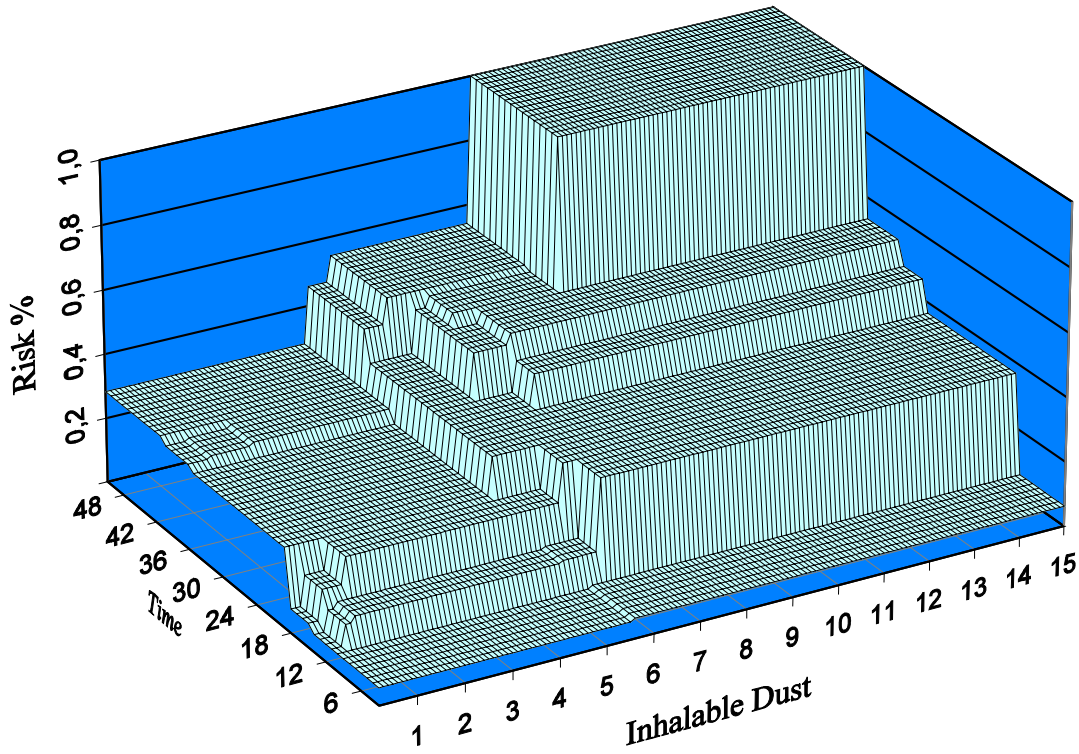

Figure 5.1: The two-dimensional isotonic model (L=40 blocks).

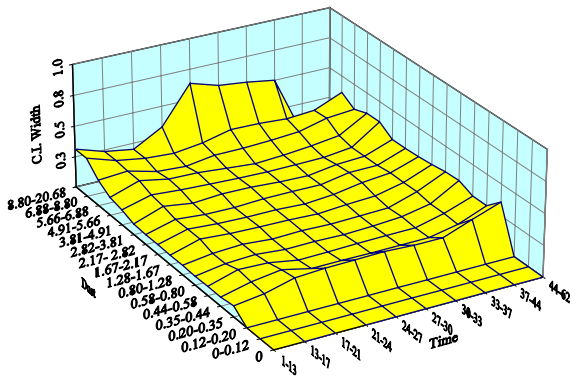

Figure 5.2: The width of the confidence surfaces simulated under the isotonic estimates.

The final number of bivariate solution blocks was 40. In figure 5.1 note that some of the solution blocks contain very few information and while some "jumps" are high, some others are not and it is expected that their contribution to the likelihood may not be important. On fitting the reduced model, simulations were used to assess the ε^*-level. On this purpose 5000 random permutations of the response variable were produced as if it was independent on the explanatory variables ($17 \times 10 = 170$ pairs). In each permutation the isotonic and reduced isotonic regression were fitted. Then to get the ε^* that leads to 0.05-significance test, we picked the 250th smallest p-value when only two solution blocks remain. The estimated value ε^* for 5% singificance level was 0.00038. Figure 5.3 presents the reduced model.

The change in the deviance between isotonic and reduced model is 16.59 (Table 5.3). The number of solution blocks has been reduced from 40 (isotonic) to 3 (reduced isotonic). The cutpoints for dust in the reduced model were at concentrations 0.9, 4.5, 5.5 and 5.8 $\frac{mg}{m^3}$.

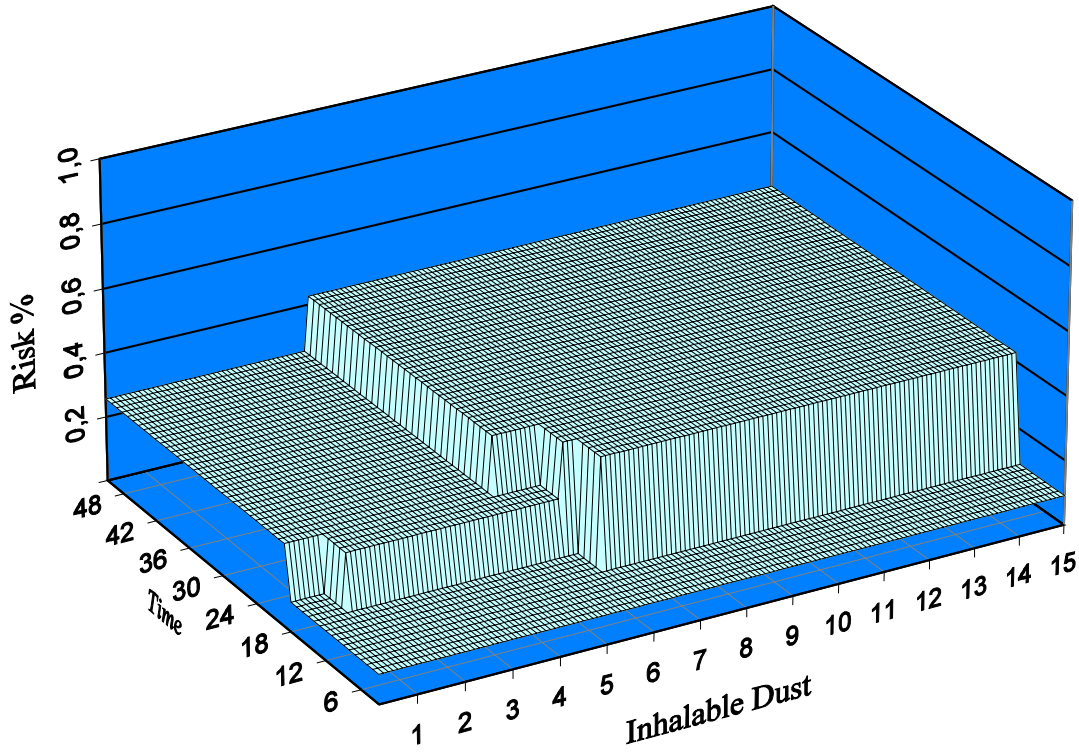

Figure 5.3: The two-dimensional reduced isotonic model (L=3 blocks).

A last step is performed in order to compare these two models. After simulating (5000 simulations) under the assumption that the reduced model is the correct one, the conclusion is that such a large change in the Likelihood as the observed one could have occurred with probability p-value = 0.632 and the more parsimonious model is selected. Thus, applying reduced isotonic regression a useful stratification for both variables time and dust is assessed, by combining them in three groups of higher, intermediate and lower risk.

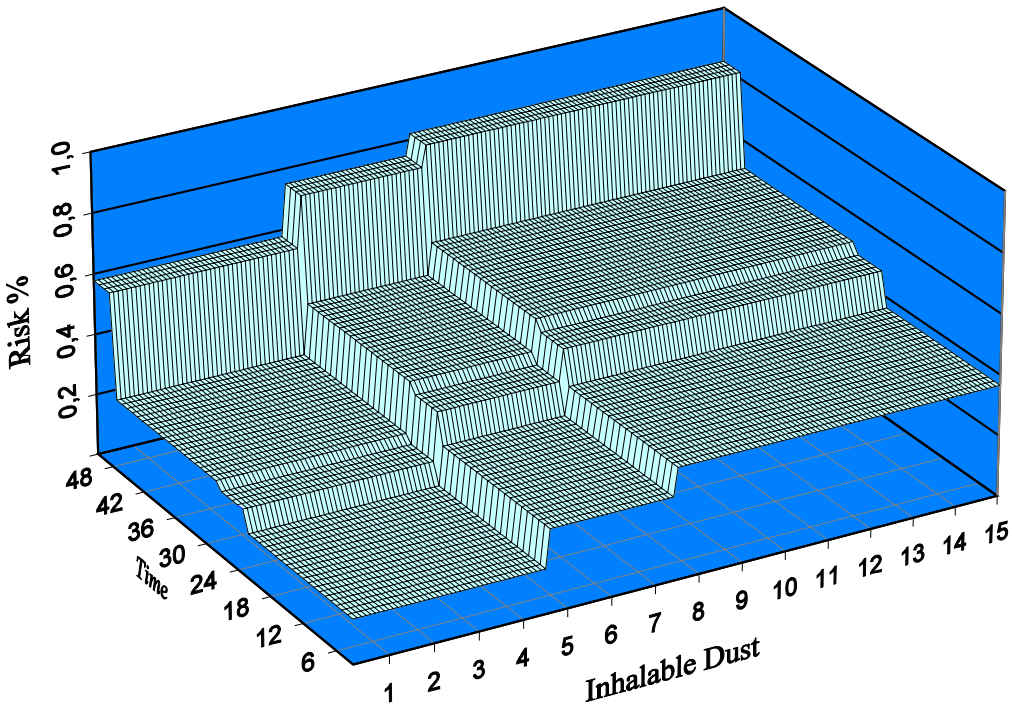

Figure 5.4: The additive isotonic model (L=12 blocks).

The additive isotonic model was also applied to the data. The model has deviance 999.94 and summarizes the dust in three groups (cutpoints: 4.5 and 7.4 $\frac{mg}{m^3}$), and time in 4 groups i.e. 12 solution blocks (figure 5.4) in total.

Table 5.4: Several criteria to compare isotonic-surface, reduced-surface and additive isotonic model.

Model	Deviance	LS	AIC	BIC	$\bar{R}^2$
Isotonic surface	959.87	40	1039.87	1096.36	0.148
Reduced surface	976.45	3	982.45	986.69	0.085
Additive isotonic	999.94	12	1025.94	1040.89	0.061
CART	990.50	3	990.50	1000.74	0.104

Table 5.4 gives a comparison of the three models: isotonic, reduced isotonic and additive isotonic using Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC=Deviance+$\frac{1}{2}p \cdot log(n)$) and the coefficient of determination ($\bar{R}^2$). Reduced isotonic regression controls better the trade off between fit and model complexity. Recall that our proposal is rather adequate for stratification and thus it would be relevant to compare the different models using ROC curves. The area under the curve was 0.658 for additive model, 0.690 for the isotonic surface and 0.677 for the reduced surface.

Under the light of this consideration a classification tree [68] is also applied to the data. The results are presented in table 5.5 and they correspond to a final tree with three terminal nodes, which has been selected after cross-validation. The predictors were combined in 3 groups. The result is roughly similar to the one obtained by applying the reduced-surface model. Note that the group with higher risk in the classification tree is defined by time $\geq$ 16.5 and total dust $\geq$ 4.8. This high risk group is also to be seen in the reduced-surface model (see figure 5.3) where the estimated proportion is 0.421 (0.431 for tree model). The dust cutpoint of about $5\frac{mg}{m^3}$ is also present in the additive model.

Table 5.5: The final classification tree in numbers.

Node	**Covariate**	**Deviance**	**Deviance reduction**	**n**	**proportion**
1(root)	Time < 16.5	1058.17	43.7	920	
$2^\star$	Time ≥ 16.5	169.50		243	0.111
3	Dust < 4.8	844.80	23.8	677	
$4^\star$	T< 16.5 & D< 4.8	481.90		429	0.249
$5^\star$	T< 16.5 & D≥ 4.8	339.10		248	0.431

The final deviance of the tree model is 990.50 with 3 degrees of freedom. According to all three criteria (AIC, BIC, $\bar{R}^2$) the CART model is better than the isotonic surface and the additive model but not better than the reduced surface model.

6 THRESHOLD VALUE ESTIMATION

6.1 Introduction

The estimation of threshold limit values (TLV) is an important task in many medical areas. An obvious example is occupational medicine. If a chemical compound or any other substance in the workplace is known to have adverse effects on the health of the employees the existence of a TVL and its estimation is of great concern. A similar problem is faced in the clinical context with the question who should be treated, - e.g. beyond which blood pressure value one should prescribe antihypertensives. The goal of the MAK study considered within this paper was to assess a TVL for total dust concentration in the working area in order to prevent chronic bronchitic reaction. There is still a great debate about how thresholds should be assessed. The MAK study is a representative example. The estimated values have been first reported in [15], criticized by McLaughlin et al. in [42] and finally justified by the additional contribution of Ulm and Salanti in [63].

The existence of a threshold value corresponds to several regression shapes [14]. The

most popular are the "hockey stick" and the logistic shape. Assuming a threshold value τ, the "hockey stick" shape corresponds to a linear increase beyond the threshold, whereas the logit model assumes an increase in the logit scale of the endpoint. Before τ no effect is assumed. However, it is more plausible to assume a background parameter λ corresponding to the baseline risk rather than of no risk. The model takes the form

$$f(P(x)) = \begin{cases} \lambda, & x \leq \tau \\ \lambda + \beta(x - \tau), & x > \tau \end{cases} \tag{6.1}$$

where the function f is the identity function for the "hockey stick" shape or the logit function for the logit shape. The model can be extended to more than one covariate.

Newer developments are connected to isotonic regression. A Likelihood Ratio test for detecting thresholds within an isotonic-surface regression model has been proposed in [62]. Additionally, in a later paper an additive isotonic model has been applied [63] and compared to the previous approach. The principal motivation in using isotonic regression in assessing a TVL is to relax the linearity assumption in the model (6.1). A step function is used instead, to represent the increase in the risk. This approach has been revised and compared to other methods in [35] where a generalized additive model has been used to fit the data, but no objective statistical strategy has been proposed on estimating the threshold. The maximal allowed concentration for the compound of interest has been estimated as the value that corresponds to an arbitrary amount of increase in the risk.

This chapter discusses three methods based on isotonic regression for estimating thresholds: one introduced by Ulm and two new proposals. The first - referred to as **method 1** – has been proposed in [62]. This method will be compared to **method 2** in a simulation study. This is an alternative approach based on re-

duced isotonic regression introduced in chapter 4. Finally another new method (**method 3**) which makes use of the closed testing principal is outlined. The performance of the proposed methodology will be shown, presenting an application from the MAK study [15].

6.2 Methods

6.2.1 Formulation of the problem

On estimating thresholds in a dose-response relationship between a continuous – or ordinal – explanatory variable and an outcome, there are two main steps. These consist on testing the following hypothesis:

Step 1 H_{ex}: A threshold value exists

Step 2 H_{loc}: The threshold value *is located* at $\tau = x_i$

As complementary hypothesis of the first assumption (H_{ex}^{C}*: no threshold exists*) usually the linear dose-response shape is excessively set. This can be tested by fitting a generalized additive model and applying a non-parametric test for linearity. An alternative is to assume and test whether

$$\tau = min_{i=1,\ldots,N}\{x_i\}.$$

That means that the threshold is located at the first observation. There is some controversy about testing hypothesis H_{ex}. Cox, for example, argues that the question whether a threshold exists or not can not be answered by means of statistical methods [14, 64]. He stated that assuming a threshold is plausible in many toxicological studies, and even in cases where there is no biological justification, this assumption can have practical value. Thus, the point is not to find out if a threshold truly exists and to test this assumption, but rather to estimate it and test its location. The final

decision about threshold problems is usually taken on a biological, ethical and political basis. Statistical methods are only tools that can *eventually* direct the process. This chapter is only concerned with the statistical issues regarding threshold value estimation.

6.2.2 How to approach the threshold value problem

A usual approach in estimating thresholds, is first to select a subset of "suspicious" x_i values (or all x_i if possible), and then to fit a model of equation (6.1) at each of these points. Several criteria can be applied to test H_{ex} and H_{loc}. For example, the AIC criterion can be used: the point x_i which yields the lower AIC is selected as threshold and H_{ex} is not rejected if this AIC is lower than the one estimated from linear regression. Instead of AIC other criteria can be used, as for example the deviance. To determine the set of "suspicious" threshold locations a flexible model is fitted - usually a generalized additive one. Then, one has to screen the GAM graph and take a "neighborhood" about a point where an increase in the risk seems to occur. These approaches, as shortly described here or slightly modified, are widely used. However it is obvious that they are intuitive and at some points arbitrary.

In contrast, two model-based proposals have been published on this topic [47, 61]. Both are well developed in terms of their statistical properties. Pastor and Guallar [47] proposed a general approach based on logistic regression where a changepoint occurs, changing the regression shape from linear to quadratic. The model parameters and the changepoint are estimated iteratively by an algorithm similar to the iterative least squares algorithm described in [41]. To infer for the changepoints and coefficients (testing the observed values i.e. $H_0 : \tau = \hat{\tau}$), they use the chi-square distributed Likelihood Ratio statistic. They justified this approach in a simulation study, even though this approximation holds only under the assumption that a

changepoint truly exists (defined case), as they note in their paper.

To estimate the threshold value Ulm [61] fits changepoint models setting $\tau = x_i$ for all $i = 1, ..., N$ and obtains N values for the Likelihood Ratio test LR_i. The maximum value of all these tests indicates the location of the threshold. Then to test $H_0 : \tau \leq min\{x_i\}$ against $H_1 : \tau > min\{x_i\}$ he compared the models with and without threshold. This LR test follows an one-sided chi-square distribution, as he proved in a simulation study.

It is important to note that in case of estimating thresholds not only one model should be applied but more than one approach has to yield a threshold in order to conclude its existence and position. Different approaches can offer different and contradictory results [56, 64], thus *there is not a unique and formal approach on estimating thresholds, but the decision about H_{ex} and H_{loc} should be taken after applying a variety of approaches.* External validation of the selected models can also provide useful information, since there are cases where the data can be fitted by a variety of models (considering a threshold or not) where all of them may fit well.

6.2.3 How to estimate thresholds using isotonic regression: two new alternative approaches

As already noticed, isotonic regression model is a changepoint model and therefore its use on TVL problems is straightforward. On searching for thresholds the following steps need to be taken.

1. **Prove dose-response relationship**: This is a trivial but important step. A test for trend – for example the isotonic R in equation 2.10 – has to be applied.
2. **Reject the linearity assumption**: The generalized linear model should be compared to the isotonic one (or the reduced) and check out which of

these two models fits better. Given the non parametric nature of the isotonic regression, bootstrap methods will be used to assess the p-value. If the linearity assumption is not rejected, then no threshold can be assumed and H_{ex} is to be rejected. Otherwise, the possibility for threshold value existence is open. However, that does not mean that a threshold truly exists, since there is a variety of shapes that have no threshold without being linear.

3. **Find a set of "suspicious" threshold value locations**: This can be **(i)** the set of the cutpoints estimated by fitting isotonic regression ("iso-cutpoints") or **(ii)** a subset of the "iso-cutpoints" selected by applying the reducing procedure described in algorithm 3 ("red-cutpoints").

4. **Select the threshold**: Among the possible candidates for threshold, the true threshold value location has to be found. This can be **a)** the "iso-cutpoint" that corresponds to an *important* change in the deviance (see paragraph **The one-sided chi-squared pooling procedure**) or **b)** the first "red-cutpoint".

The scheme proposed above is conditional in terms of performance: the power of each step is bounded by the power of the previous steps. Tests for trend and their properties under several circumstances have been studied in chapter 3. This chapter is not concerned with evaluating the linearity assumption. The parametric bootstrap used rejecting H_{ex} will be shortly described in the application. Steps two and three are the subject of this chapter.

The fact that PAVA provides a small set of cutpoints without any a priori information about their location, simplifies the TVL detection procedure in testing the changepoints one by one, and sets the requirements to a minimum of monotonicity. However these changepoints are selected in order to efface the violators and it is not necessary that they correspond to an important increase in the risk. Regarding this, two methods will be used concerning steps 3 and 4:

- **method 1** combining **(i)** isotonic regression and **(a)** the "one-sided chi-squared pooling procedure"
- **method 2** combining **(ii)** reduced isotonic regression and **(b)** selecting as threshold the first reduced cutpoint.

Recall that the reduced version of isotonic regression incorporates an elimination of the cutpoints strictly to those which define a significant increase in the response.

The whole model can be thought of as an extension of the traditional "hockey-stick" threshold model (6.1) where the estimation for the first level set defines the background risk λ and the first cutpoint for the dose defines the threshold τ. After τ the increasing part $f(x-\tau)$ can be a step function or constant.

Given the nonparametric nature of the procedure, to estimate confidence intervals for the threshold, one has to apply bias corrected and accelerated bootstrap confidence intervals [11, 16], where the accelerator ac is estimated applying jackknife. If $\mathbf{x_{(i)}}$ is a subsample of the original data set $\mathbf{x}$ leaving the i-th observation outside, $\tau_{(i)}$ the estimated threshold and $\tau_{(\bullet)} = \sum_{i=1}^{N} \tau_{(i)}/N$, then

$$ac = \frac{\sum_{i=1}^{N}(\tau_{(\bullet)} - \tau_{(i)})^3}{6\{(\tau_{(\bullet)} - \tau_{(i)})^2\}^{3/2}}.$$

If $\mathbf{b}$ is a bootstrap sample of the original data set and $\tau(b)$ the estimated threshold, the bias corrector estimated by B bootstrap samples is

$$z_0 = \Phi^{-1}\left(\frac{\sum_{b} I_{\tau(b)<\tau}}{B}\right).$$

The $(1-a)\%$ confidence interval is

$$CI_a = (\tau(b)_{a_1}, \tau(b)_{a_2})$$

$$a_1 = \Phi\left(z_0 + \frac{z_0 + z_{ac}}{1 - ac(z_0 + z_{ac})}\right),\ a_2 = \Phi\left(z_0 + \frac{z_0 + z_{1-ac}}{1 - ac(z_0 + z_{1-ac})}\right)$$

where z_a is the a-th normal quantile. For an example see table 6.3.

The one-sided chi-squared pooling procedure

This method has been recently proposed [62] in order to estimate thresholds in isotonic regression framework. For the rest of this work I will refer to it as the threshold procedure or **method 1**. In assessing a TVL the constant risk categories corresponding to the isotonic predictor of interest are lumped together starting now from the two lowest groups. The loss in the fit

$$LR_{Thers} = D_l - D_{l-1} \tag{6.2}$$

(where D_l denotes the Deviance corresponding to $l = L, L-1, ..., 1$ the isotonic level sets) is analyzed. As long as the fit does not decrease significantly the categories are pooled together. If not, the cutpoint between the categories is used as a threshold.

This method can be thought of as a variable selection procedure, were one has to choose between two different representations (or different degrees of freedom) of the same predictor to include in the model. Since the degrees of freedom for each term are the number of level sets (and they differ by one), the change in the deviance (6.2) should follow a *one-sided X^2 distribution with one degree of freedom*. Another criterion that can be used here is the Akaike's information criterion. The procedure will be the same as before, but now the AIC will indicate a "gap" in the goodness of fit and consequently the threshold value location.
An important problem rising from **method 1** is related to the appropriate test statistic in order to define more clearly what exactly "a large change in the likelihood" is. Simulations have shown that the Likelihood Ratio test (equation 6.2) is not X_1^2 distributed.

6.3 Simulation study

6.3.1 Evaluating the chi-square approximation in the pooling procedure

First the adequacy of the one-sided chi-square approximation in the pooling procedure in **method 1** will be evaluated.

Grouped data:
Simulated data sets have been generated assuming that the response is linearly dependent on the dose ($p_i = \alpha + \beta x_i$), where p_i denotes the response probability and x_i the dose. Five dose groups are assumed, i.e. the dose is ordinal. **Method 1** is applied for several α, β and number of observations per dose group. Figure 6.1 shows that the X_1^2 approximation is not always consistent. For the first three poolings, the empirical distribution of the change in the deviance is shown together with the one-sided chi-square cumulative distribution. For more than three poolings, the curve is moving downwards.

For small slopes, the approximation holds quite good (figure 6.2). When the slope is high, the disagreement is quite substantial. The distribution was not much influenced by the intercept α. Note that it is quite complicated to find the theoretical distribution of the test in equation (6.2): the main problem above the unequal weights is that the distributions corresponding to test for k and $k-1$ level sets $T_{01,k}$ and $T_{01,k-1}$ are not independent.

Continuous data:
The simulation study is repeated, keeping the sample size constant but now the predictor is not used as continuous. The one-sided chi-square distribution is proved inadequate almost in every situation (figure not shown). For the sake of example, the critical value for slope 0.02 was 7.14 instead of 2.71.

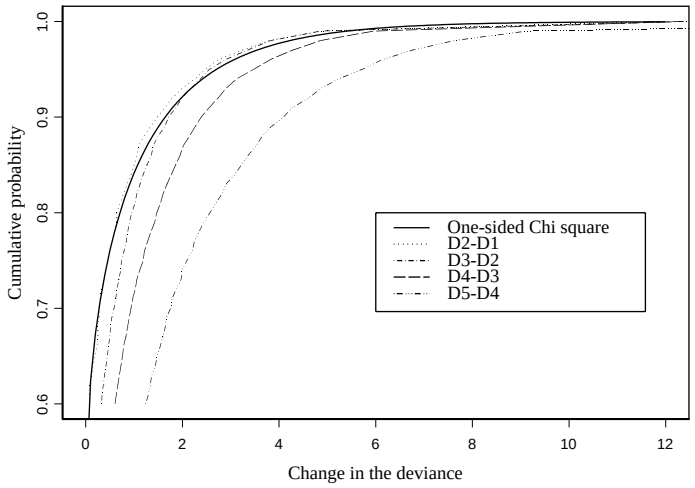

Figure 6.1: Change in the deviance (equation 6.2) while pooling (slope=0.02, 5 dose groups, sample size 250). As Di-Dj is denoted the change in the deviance between the model pooling information until the j-th group and the model pooling information until the i-th group.

6.3.2 Evaluate the two approaches on estimating thresholds using isotonic regression

Simulation study assumptions and set up

Both **method 1** and **method 2** are evaluated within this simulation study. Assume that the dose-response is proven based on a test for trend. However, the elimination procedures for both methods (either point (i) in step 4 for **method 1** or algorithm 3 for **method 2**) can return to a single level set (ideally with probability not greater than 5%) despite the significance of the dose response relationship. In this case no threshold can be estimated even if the test for trend is significant and the linearity assumption is rejected. Thus both procedures will be evaluated as "test for trend"(**a-**

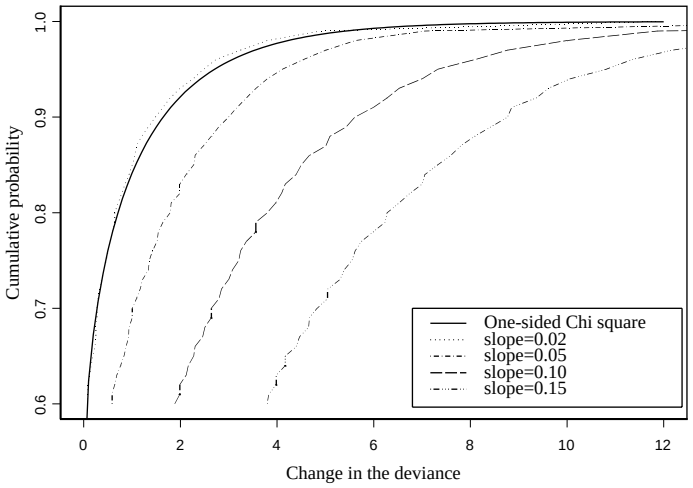

Figure 6.2: The first pooling (D2-D1) distribution assuming different slopes (5 dose groups, sample size 250).

case H_{0a}: constant risk against H_{1a}: increasing trend), where hypothesis H_{0a} is rejected when the algorithm returns at least two level sets. In this context, their behavior will be described by estimating the power and the of type I error.

The capability of the methods to detect thresholds (**b-case** H_{0b}: no threshold against H_{1b}: threshold) will be investigated. Given that both approaches end up with at least two level sets, a threshold value can always be estimated. Hence, the accuracy of both approaches will be evaluated in threshold value existence situations (to measure the power), but also in cases that correspond to dose-response relationship where no threshold is assumed (to measure the error).

Only grouped data are considered. Regarding this decision additional remarks need to be made here. When the assessed threshold is placed in the lowest dose-group, the answer about the existence and location of the threshold is not clear. The threshold may be placed somewhere between the upper and lower dose concentrations of the

first group and H_{ex} can not be rejected unless this dose group corresponds to zero exposure. Exactly this situation is considered in this simulation study.

Four shapes of regression lines have been studied (Figure 6.3): a flat constant function (A), a linear regression line (B), and two types of segmented increasing line (C and D). Each of these regression lines has been considered under the assumption of 5 dose groups. Equal number of observations per dose group that varies between 50 and 250 are assumed. For each combination of regression parameters and sample size, 10 000 simulations have been analyzed. The estimation of the elimination's level ϵ^* for **method 2** has been performed using 1000 simulations. For every regression shape 20 parameter combinations (sample size and slope) have been analyzed.

Shape A:

This regression type corresponds to the absence of a dose response relationship. Under this assumption the error of type I can be estimated, considering the elimination's procedures as a test for trend. That means that **method 2** and **method 1** are expected to result in a single level set with probability about 95%. The event rate of the underlying function was assumed to be 0.05, 0.10, 0.15, 0.20.

Shape B:

This regression type corresponds to a linear increasing dose-response relationship, so the power of each elimination procedure as test for trend can be estimated. Higher power corresponds to greater proportion of non constant estimated regression lines. However, no threshold existence is assumed, so a sort of "error of type I" as TVL detection method (b-case) can be described for both approaches. The regression line starts with event rate 20% and increases with slopes 0.02, 0.05, 0.10 and 0.15.

Shape C:

The third type represents a segmented regression line assuming a threshold. The baseline constant risk is assumed to be 20% and afterwards the risk increases lin-

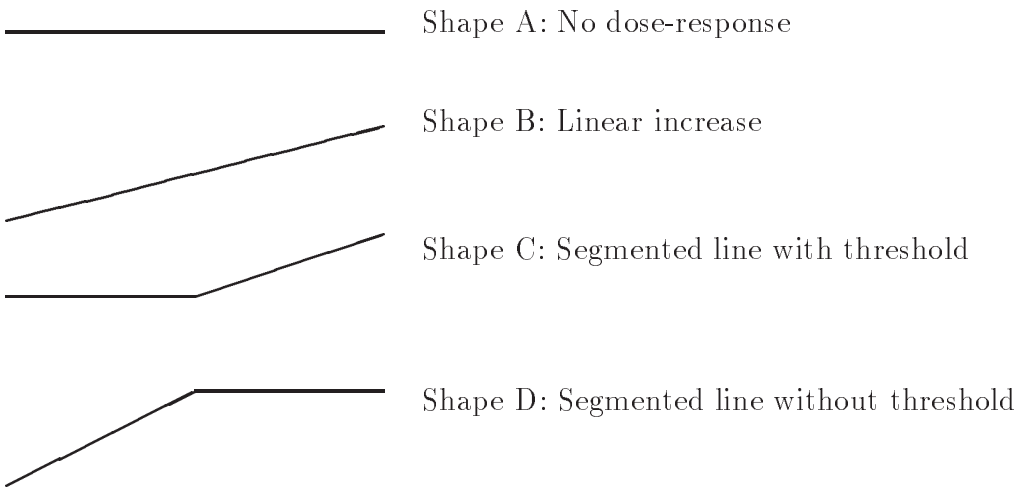

Figure 6.3: Shapes for dose-response studied in simulation study.

early with slopes 0.02, 0.05, 0.10, 0.15. Different cutpoint locations have been also examined. Next to the power of the elimination procedures as test, the TVL detection capability of the methods can be estimated. As a measure, the probability to get the correct threshold given that an increasing trend has been assessed.

Shape D:

The last type of regression is also a segmented line but assumes no threshold: first linear increasing trend and then constant risk. Several considerations have been made again regarding the slope of the increasing part (0.02, 0.05, 0.10, 0.15). The power of the test under that shape and the behavior of the threshold procedure can be estimated.

Notation:

Both **method 1** and **method 2** will be evaluated under two considerations. First, considered as tests for trend, where the non dose-response assumption H_{0a} is rejected when the method results in more than one level set. The power$_a$ (shapes B-D) and

the type I error_a (shape A) will be assessed (second column in table 6.1). Second, both procedures will be considered as threshold value estimation procedures. In this case their capability to reject hypothesis H_{0b}: no threshold existence will assess the power_b when a threshold exists (shape C) and the type I error_b when there is no threshold (shapes B and D). See column 3 in table 6.1.

Table 6.1: Simulation's study for threshold value estimation.

Regression type	Testing dose-response (a-case)	Estimating the threshold (b-case)
A)	Error_a I	-
B)	Power_a	Error_b I
C)	Power_a	Power_b
D)	Power_a	Error_b I

6.3.3 Results

Shape A:

For **method 1**, the type I error_a is about 8% and its small variation regarding sample size and event rate is minimal, except for N=50 and event rate 5% (figure 6.4). Regarding **method 2** the error_a of type I is not to be tested, since the elimination procedure is so designed as to control it. In the following remarks about the power_a of **method 2** as test, the reader should keep that in mind.

Shape B:

As expected, in cases of small slopes the power_a for both procedures is low. For example for slope 0.02 it lies in a range of 30%-70% depending on the sample size. However, the power increases very fast with increasing slope and both **method 1** and **method 2** have a good power for slopes greater that 5%. In figure 6.5 the

power of **method 1** procedure is depicted. Regarding **method 2** the variation of power is similar.

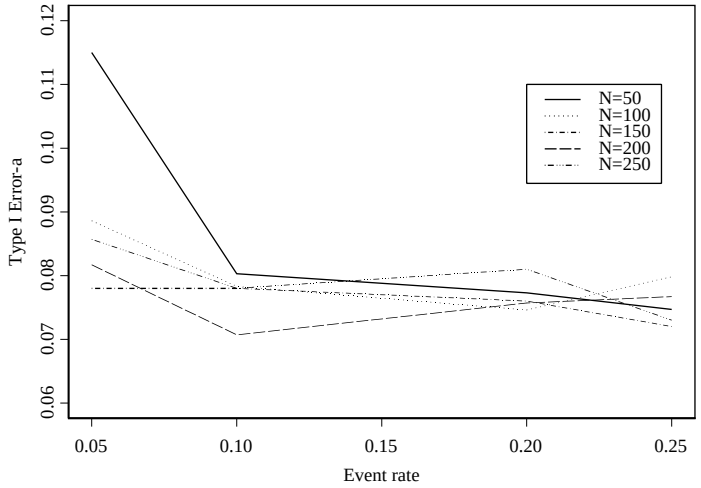

Figure 6.4: Method 1 for shape A corresponding to constant event rate - The type I error$_a$.

Focusing on cases where the power is at least 70%, the assignment of thresholds in the several dose groups has been examined. The **method 1** procedure tends to select as threshold the dose that corresponds to an increase of about 7% from the starting group. For slopes 10% and 15%, the first level set has been estimated as threshold, for every sample size. The thresholds assessed according to **method 2** do not seem to follow any special pattern for small slopes (2%): every cutpoint has more or less the same probabilities to be selected as threshold. However, for slopes higher than 5% or greater sample size, the first groups are more likely to be selected as threshold though less successfully than **method 1** procedure, i.e. **method 1** concludes to no threshold result faster (regarding sample size and slope) than **method 2**.

Shape C:

Power here presents the same characteristics as in shape B. Figure 6.6 presents

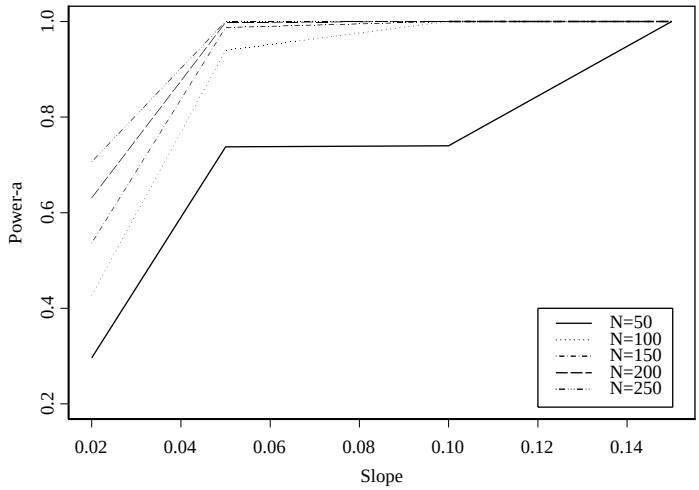

Figure 6.5: Method 1 for shape B corresponding to linear increase - The power_a.

the results for **method 2** only, since the figure for **method 1** does not present remarkable differences. In this case, the success in estimating the threshold is of interest. The correct threshold position lies in the second level set. The probability to assess the correct level set as threshold (power_b) given that more than one level sets have been obtained is presented in figure 6.7 for **method 1** and figure 6.8 for **method 2**. The results about the power_a described in figure 6.6 affects somewhat the capability of both procedures to detect the correct threshold.

For small slope and sample size the success of assessing the correct threshold is not satisfactory but increases rapidly as long as the power_a of the methods as test for trend increases: The power_b of the methods increases sharply with the slope and secondarily with the sample size in an almost linear way. It is remarkable that the probability to assess no threshold (estimated threshold: the first level set) has been very low and almost identical for every sample size and slope for **method 1** whereas for **method 2** it decreases with the sample size. Additionally, figures 6.7 and 6.8

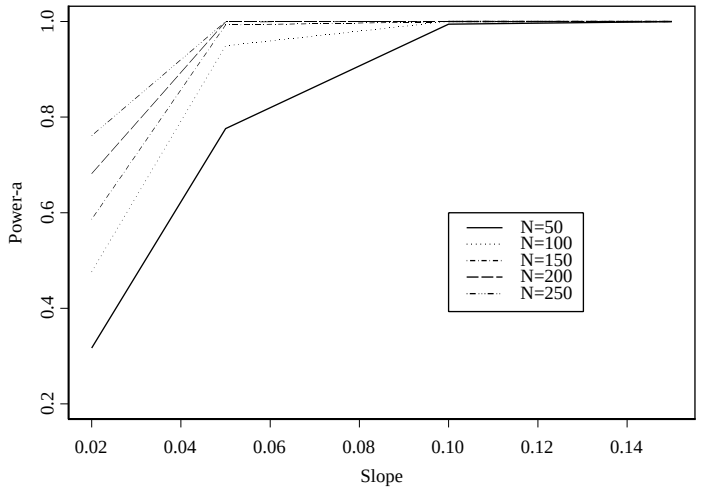

Figure 6.6: Method 2 for shape C corresponding to a segmented line with threshold - The power$_a$.

present the probability to assess the third level set as threshold, since it is observed an important tendency of both methods to assign thresholds to the adjacent group that corresponds to a higher dose value category. This probability, important in cases of low power$_a$, decreases with the increase in slope.

Figure 6.9 compares the two methods in their success to assess the correct threshold. The reduced isotonic regression behaves better with smaller slopes and sample size, but in general the results are similar. The mean power$_b$ for **method 2** is 58.4% and for **method 1** 57.2%. Regarding the probability to get no threshold when one truly exists, the reduced isotonic regression presents the advantage that the increase in the slope can decrease the chances to estimate no threshold (figure 6.10): the error for **method 2** presents variation with sample size whereas for **method 1** remains about 5%. The mean error probabilities were 4.3% for **method 2** and 5.2% for **method 1**.

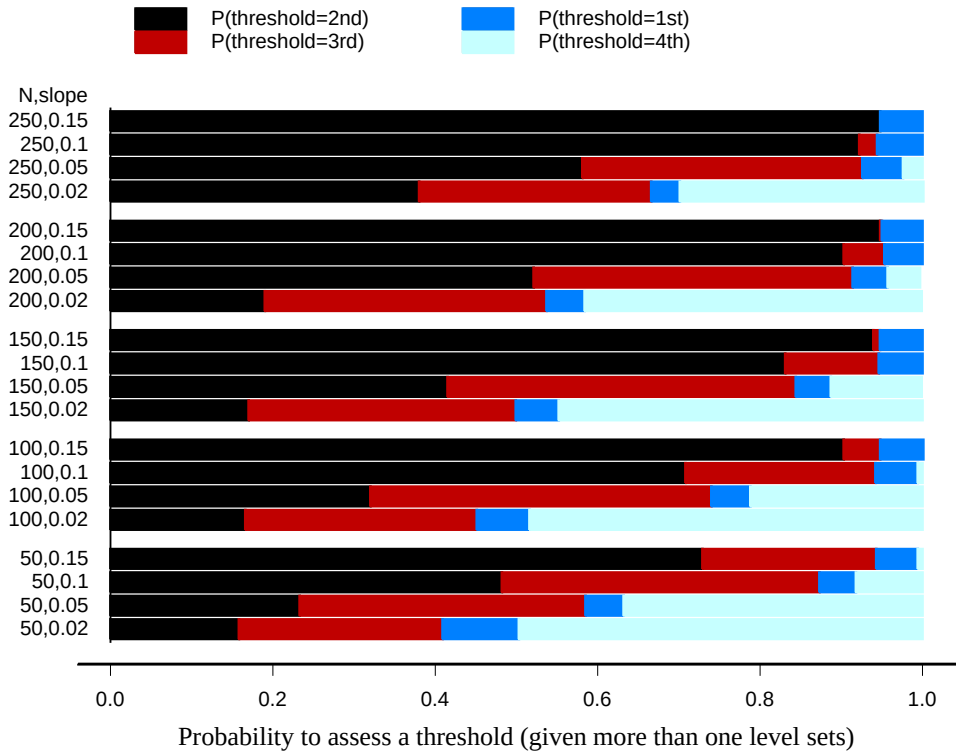

Figure 6.7: Shape C: The probability to assess the correct threshold (power$_b$) using **method 1**. The true threshold location is in the second dose-group.

The last step in our simulation study for shape C consists of setting a different threshold value location (figure 6.11). Now it is after the third dose group that the risk increases. A slope of 10% is assumed for the increasing segment. Both methods improve their behavior for small sample size but no improvement can be seen when the sample size is 150 per dose group or greater. The probability to assess no threshold becomes clearly smaller.

Shape D:

In the third case the power$_a$ seems to increase roughly linearly with "the hight of the step" in risk after the first dose group. Due again to their similarities with

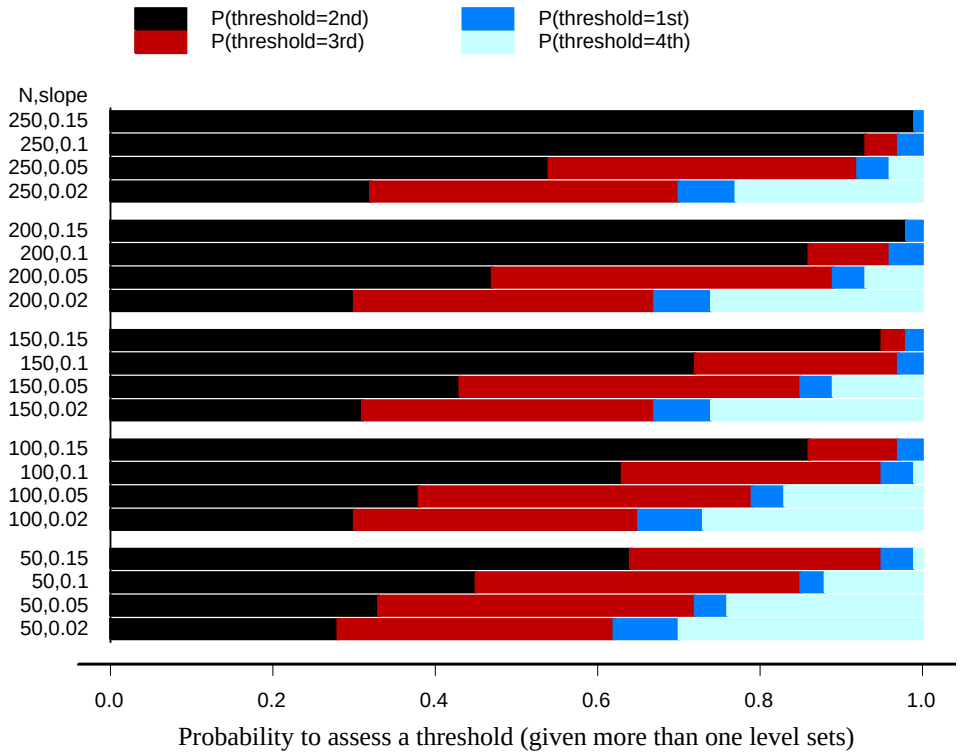

Figure 6.8: Shape C: The probability to assess the correct threshold (power$_b$) using **method 2**. The true threshold location is in the second dose-group.

method 2, I present only the result from **method 1** (figure 6.12). The capability of the procedure to assess thresholds depends again on its power$_a$ as test. The first group has the greatest probability to be selected as threshold, which is in agreement with the underlying regression shape. Unlike the A shape, here the **method 2** presents better results (19 out of 20 combinations ended up to a no threshold existence result) than **method 1** (16 out of 20 combinations) in supplying evidence against the threshold existence.

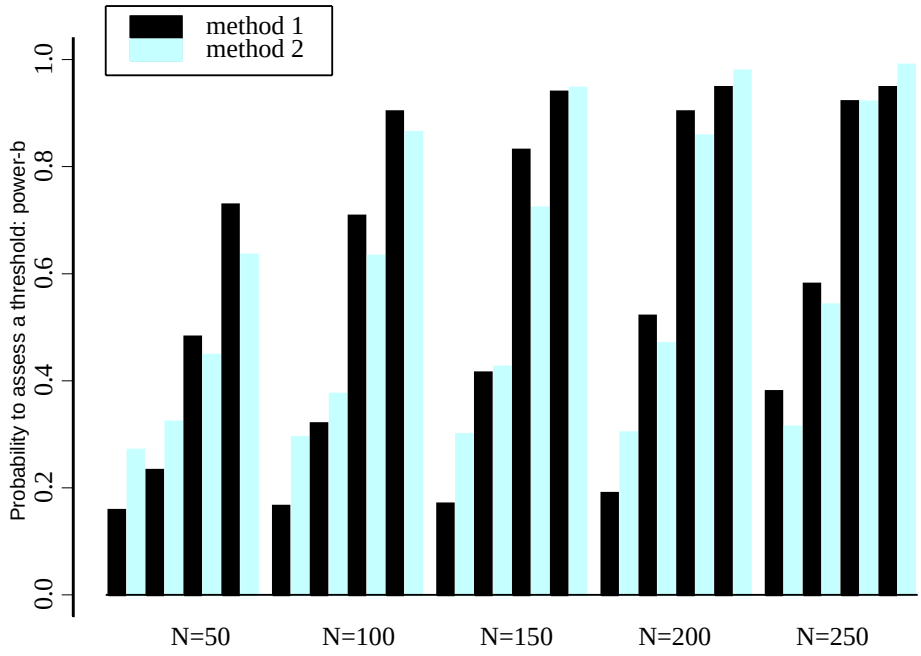

Figure 6.9: Shape C: Comparing **method 1** regression and **method 2** regarding the probability to assess the correct threshold-power$_b$. The bars in each sample size panel correspond to slopes 0.02, 0.05, 0.10 and 0.15.

6.3.4 Comments on simulations

To summarize the results, when a TVL truly exists both procedures will detect it correctly when the increase in the risk is sufficiently high (at least 5%). In the case where a small increase is expected, one should keep in mind that both procedures tend to estimate thresholds higher than the true one. Note that the type I error for **method 2** can be substantially reduced by increasing the sample size.

In case that no TVL exists and the dose-response relationship is not strong (slope

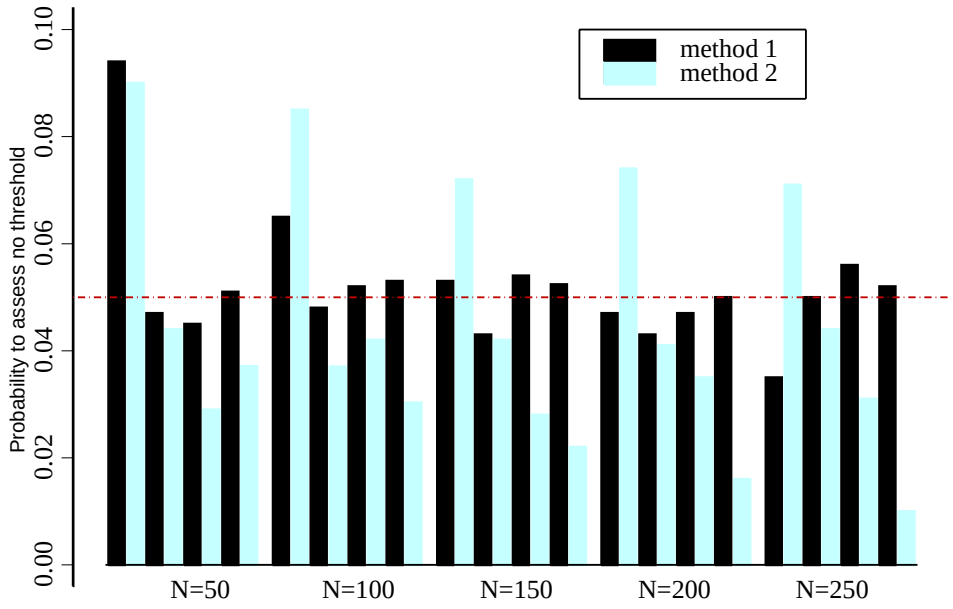

Figure 6.10: Shape C: Comparing **method 1** and **method 2** regarding the probability to assess no threshold when one exists (type I error$_b$). The bars in each sample size panel correspond to slopes 0.02, 0.05, 0.10 and 0.15. The horizontal line is drawn at the nominal level 5%.

less than 5%), both approaches will probably return a threshold. Hence the estimation of TVL does not necessarily justify its existence. In our simulation study the non TVL existence assumption has been linked to the selection of the first dose group as threshold. That is valid only if the dose is zero in the first group. In case that the exposure in the first dose group is not zero but a range of values (i.e. 1-10$\frac{mg}{cm_3}$), assigning a threshold there would simply mean that the true threshold lies somewhere between these values. The "no threshold value existence" conclusion

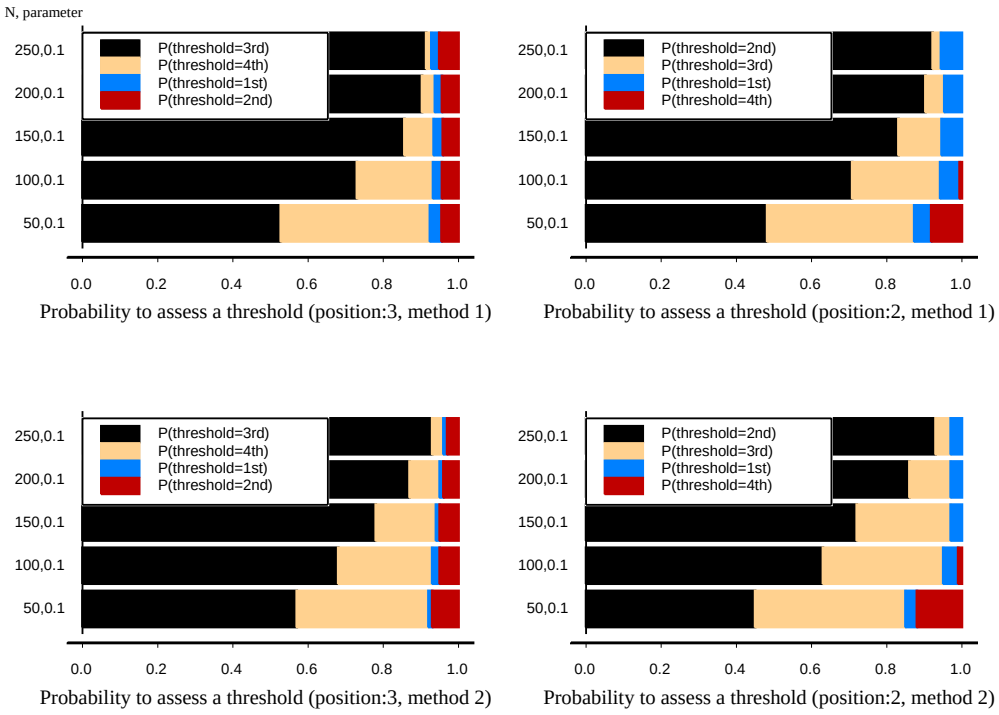

Figure 6.11: Shape C: Probabilities for **method 1** and **method 2** to assess a changepoint as threshold, when its true position is in the 2nd and 3rd level set (regression slope=0.1).

cannot be assessed. A solution to this problem could be to use the dose in continuous form. Categorizing a continuous variable to an ordinal can highly affect the result, since a priori assumptions as the selection of cutpoints and the number of observations per group can produce bias regarding the TVL estimation. For this reason the use of continuous dose is recommended, and accordingly an example is presented in the following section.

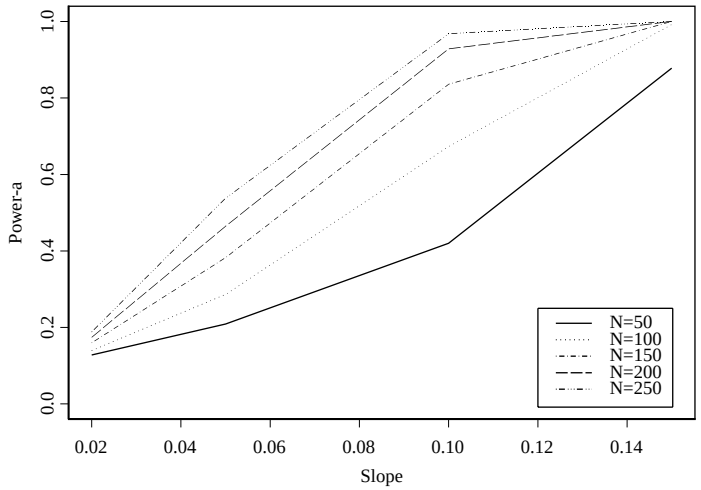

Figure 6.12: Method 1 for shape D corresponding to a segmented line without threshold - The power$_a$.

6.4 Adapting the closed testing procedure for estimating thresholds

An alternative on selecting the set of threshold value "candidate locations" among the isotonic cutpoints, is to applying a closed elimination procedure (**method 3**). We have already discussed how that can be accomplished by applying algorithm 4 (see chapter 4). In this case the main concern of the closed testing procedure is to estimate a minimum concentration above which the risk presents a significant increase and *not* to proceed in a complete re-estimation of the regression line. So, the primary focus is rather to answer the following questions:

1. After the first cutpoint, is there evidence for significant increase in the risk?

2. If not, given a hierarchy of changepoints, which is the lowest one that corresponds to a significant increase in the risk?

Note that the algorithm 4 presents the inconvenient of slightly low power. An option to increase the power is to use level a instead of portions of a at every layer and to use conditional testing to correct the additive increase in family-wise error. Rom et al. [50] proposed to make the one part of the regression line conditional on the other one. Since the beginning of the dose-response is more important for threshold estimation, we want a procedure where testing between higher dose is conditional on rejecting on lower doses. That can be accomplished through conditionality: *every "vertical" hypothesis is conditional on the rejection of the previous hypothesis and every "horizontal" hypothesis is conditional on the retain of its previous hypothesis.* Considering the schema in figure 6.13 and the notation used previews chapter 4 I propose the following algorithm.

Algorithm 9 (*Closed testing for threshold estimation*)

1. *Every hypothesis $H_{l,k}$ and its offsprings are conditional on hypothesis $H_{l',k'}$ with $l' < k' < l < k$, i.e. it is tested only if the hypothesis $H_{l',k'}$ is retain*
2. *Retain every hypothesis implied by any other hypothesis that has not be rejected*
3. *Reject any hypothesis that is tested and rejected at $a = 0.05$*

The procedure starts by testing $H_{1,3}$. If we retain then we test $H_{2,4}$ (rejection : threshold = c_3, no-rejection = no dose-response relationship). If we reject $H_{1,3}$ then we test $H_{1,2}$ (rejection: threshold=c_1 non-rejection: continue by testing $H_{2,3}$). Now by testing $H_{2,3}$ in case of rejection we conclude that the threshold is at c_2 and if we do not reject, we continue by testing $H_{3,4}$. Finally, by rejection we have threshold at c_3 and by retain no-response relationship.

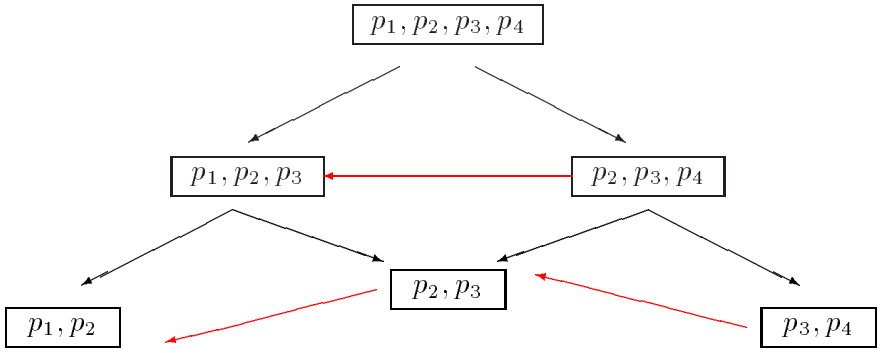

Figure 6.13: Example for closed testing on threshold value estimation.

When the conditioning occurs "horizontally" (red arrows) additional to the vertical restriction (black arrows) the power increases. Consider for example that $H_{1,4}$ is true and each nested hypothesis is rejected at 5% level. Then only with vertical restriction the power for the example 4.1 is $0.95^5 + 0.95^3 \cdot 0.05^2 + 0.95^2 0.05 + 0.95^4 0.05 = 0.862$ whereas for the structure in picture 6.13 the power is $0.95^2 + 0.95^2 \cdot 0.05 = 0.947$. However if this approach is followed, no additional information may be obtained about the shape of the dose-response relationship after the threshold.

6.5 Case study in threshold value estimation

This section goes further with the example presented in section 4.5. The dose-response relationship has been proven, so the procedure goes to the next step: to reject linearity (see section 6.2.3). Parametric bootstrap will be used. The overall response probability is kept constant, and the relationship is assumed to be linear.

Random data (1000) are produced under this assumption and at each replication the change of deviance between the linear and the isotonic model is estimated. The observed change in deviance is D(linear)-D(isotonic)=37.65, which lies outside the range of values estimated by parametric bootstrap (7.47-34.71). Thus, the linearity assumption has to be rejected.

Method 1 was applied to the data. The changes in deviance when adjacent level sets were pooled together are presented in table 6.2. Then, the loss in the fit was compared to 2.71 ($=X^2_{1,90\%}$). The pooling of the first three level sets causes a significant decrease in the likelihood, consequently a minimum value for threshold at concentration $6.97\frac{mg}{m^3}year$ was assessed.

Table 6.2: Results from **method 1**.

Cutpoint	Pooling to cutpoint	Deviance
1st	1.01	983.06
2nd	5.04	983.76
3rd	6.97	984.08
4th	10.05	991.87

Then, **method 2** was applied to the data. The threshold was estimated at $6.97\frac{mg}{m^3}$year. The corresponding background risk was 7.6%. The bootstrap bias corrected and accelerated confidence intervals are presented in table 6.3.

The isotonic level sets have been analyzed with the closed testing procedure (algorithm 9) i.e. **method 3**. The nested hypothesis $H_{1,6}$, $H_{1,5}$, $H_{1,4}$, $H_{1,3}$, are tested and all rejected at level 5%. Hypothesis $H_{1,2}$ results in p-value=0.392, thus we proceed with testing $H_{2,3}$ which is also rejected, the same for $H_{3,4}$. The p-value for $H_{4,5}$ is less than 0.001, thus the estimated threshold is at the fourth isotonic cutpoint i.e. in $10.05mg/m^3years$ cumulative exposure.

Table 6.3: Bootstrap confidence limits for the estimated threshold using **method 2**.

threshold τ	6.97 mg/m^3
$\sum I_{\tau(b)<\tau}$	389
$(z_0,\ a_c)$	(-0.28,0.01)
CI_a	[4.95-10.01]

6.6 Conclusions

The present chapter investigated the use of isotonic regression in threshold limit value problems. Isotonic regression possesses useful features that facilitate the changepoint detection. In total three methods have been proposed in order to assess the TVL.

Method 1 is based on the Likelihood Ratio test, it is straightforward but has its disadvantages. The assumed Chi-square approximation for the distribution of the Likelihood Ratio test while adjacent level sets are pooled, does not always hold. Alternatively one can use bootstrap methods to infer at each pooling, but that would make the method quite cumbersome. The method based on reduced isotonic model (**method 2**) presents a satisfactory efficacy on finding the correct threshold, once a dose-response relationship is proven.

The properties of these two methods have been investigated within a simulation study. In case that the trend is not intensive, the true threshold can also be at the level set right before the estimated i.e. in the previous isotonic cutpoint. The overall elimination procedure could also be considered as a test for trend where the consistency of risk is rejected when at least two level sets have been finally obtained. From this point of view, reduced isotonic regression presents a good control of the trade off between error I (that is controlled to be 5%) and power.

Additionally I proposed as alternative **method 3**. This is based on a modification of the closed testing procedure which is expected to increase the power of the overall testing. The conditional pattern implies that the lower end of the dose range is more important that the higher end, which is true for the threshold value estimation setting.

Modifications of the proposed methodologies are possible. L. Hothorn [31] suggests a procedure based on odds ratios. He argues that when one wants to detect an important increase in the risk, the use of confidence intervals is more accurate than comparing p-values. This idea can lead to a modification of the pooling procedure applied in isotonic level sets and the backward elimination in the reduced isotonic regression. In all three **method 1**, **method 2** and **method 3**, instead to use p-values, the confidence intervals for the level sets odd ratios could indicate a significant increase in the risk.

The presented methods can be incorporated in more sophisticated models, as the generalized additive model. All three algorithms described here can be applied in multivariate modeling, either in isotonic-surfaces models where the level sets correspond to combinations of the predictors, or in additive isotonic models where the partial fit can be the object of reduction. Details and examples are presented in [53] and [63].

In general, when estimating thresholds, I suggest the use of bootstrap methods to evaluate the accuracy of the estimation, by checking their confidence limits. Additionally, modeling the data using smoothing splines or fractional polynomials would be useful in revealing the true shape of the relationship and avoid misinterpretations. However, concluding for or against the existence of a threshold value is not a simple statistical question. The biological plausibility and strong a-priori medical assumptions need to be taken into account. In practice, for agents considered as health hazard a TVL is almost always assumed except for those who are carcinogenic.

7 MONOTONIC REGRESSION IN SURVIVAL ANALYSIS

7.1 Introduction and background

The Cox model is by far the most popular procedure for analyzing survival data. Consider the case where P predictors $\mathbf{X} : X_1, X_2, ..., X_P$, have been identified to affect significantly the survival probability. The Cox model specifies the hazard for an individual i as

$$\lambda(t|\mathbf{X}) = \lambda_0(t)e^{\tilde{\beta}\mathbf{X}}. \tag{7.1}$$

A key assumption of this model is that the ratio of two hazards is independent of time (proportional hazards model or PH model), i.e. the impact of each predictor included in the model does not change during the observation period and therefore the relative risk RR regarding two levels x_i, x_j of an explanatory variable is $exp(\beta(x_i - x_j))$ at any time. However this assumption may not hold for some variables included in

the model. A possible reason is that the coefficient β_i and therefore the RR are functions of time $\beta = \beta(t)$ and $RR = exp(\beta(t)(x_i - x_j))$.

The application of the Cox model requires validation of the proportional hazards assumption. In this direction, several tests have been proposed so far to check the predictors for time-dependency. In case of evidence, the usual PH model needs transformation, in order to include the dynamic structures.

Many graphical approaches have been proposed in order to check for proportionality. Although the judgment is rather subjective, they can be used as a first guide. Consider again a predictor in categories, a first intuitive way is to check the Kaplan-Meier curves for parallelism. If that is true, proportionality is rather likely to be fullfield. The equivalent multivariate approach would be to fit a Cox model stratified for the factor of interest and plot the survival curves for the mean value of the other predictors. The resulting curves should be parallel but also in agreement with the survival curves estimated non-parametrically (for example the Altschuler-Nelson estimates).

Another more sophisticated graphical analysis of PH assumption can be performed by plotting the log minus log survival functions against time for each level of the predictor [1]. If the proportionality assumption holds, the two curves should be parallel. To assess the survival function in each level of the predictor one has to fit again a stratified Cox model. Alternatively one can use the cumulative Schoenfeld residuals. Under the proportional hazard assumption each curve should be a random walk starting and ending at 0 (Brownian bridge). All graphical approaches described above present difficulties of visualizing the actual pattern of time-dependency and to reveal the consequences of the underlying violation of proportional hazards.

Alternatively, one can split the data in subgroups that correspond to pre-selected

[1] That is because: $S(t) = S_0(t)^{e^{\beta x}} \rightarrow S_l(t) = S_k(t)^{RR} \rightarrow ln(S_l(t)) = RRln(S_k(t)) \rightarrow ln(-ln(S_l(t))) = ln(RR) + ln(-ln(S_k(t)))$

time intervals. In each data set a Cox model is fitted and the coefficients obtained are compared to the confidence interval of the overall coefficient. Moreover, in case of violation, the pattern of interval-coefficients can roughly indicate the form of the time dependency. The time-intervals are usually selected to include enough events, but no further cut-off criteria can be established.

The most accurate approach is to apply *time-varying coefficients model* for example [29] where the coefficient $\tilde{\beta}$ is allowed to be a function of time $\tilde{\beta}(t)$. It provides a test for proportional hazards and a modeling alternative in case of violation. As special part of this approach, the Grambsch and Therneau test is defined which is based on scaled Schoenfeld residuals. Regarding this approach, a new proposal will be presented in this chapter: the incorporation of isotonic regression in the Grambsch and Therneau test to improve power

The principal motivation for using isotonic regression in modeling time variation in Cox model, is that it provides a changepoint model regarding time. Therefore, optimal cutpoints can be assessed to split time in intervals within which the effect of the variable of interest remains constant. That is an important task in many clinical studies. Isotonic regression, as already highlighted, provides unbiased estimators for changepoints without any additional requirements.

This chapter focuses on combining the benefits from isotonic regression and the flexibility of the varying-coefficients approach. The first part deals with an isotonic version of the Grambsch and Therneau test [23, 59]. Further, I will present how one can use the isotonic smoother to model the function $\tilde{\beta}(t)$ in a time-varying Cox model. The gain of introducing isotonic regression in testing and modeling PH departures will be outlined and a simulation study will be performed to assess the properties of the approach. Finally I will present an application to a data set containing children with acute lymphoblastic leukemia.

Notation:
D denotes the total number of events and t the random variable for the survival time. By t_j, $j = 1, ..., J$ we denote the unique failure times with $d_j > 0$ individuals failing at t_j and R_j the observations having $t > t_j$.

7.2 The time-varying coefficients Cox Model

As pointed out in the introduction one easily expressed alternative to proportional hazards is provided by applying models with a time-dependent coefficient. That is simply an extension of the Cox model where the time consistency assumption on $\tilde{\beta} = (\beta_p, p = 1, ..., P)$ is relaxed and is allowed to be a function of time $\beta_p(t) = \beta_{0p} + \beta_{1p} f_p(t)$. Model (7.1) takes the following form:

$$\lambda(t) = \lambda_0(t) e^{\tilde{\beta}(t)X} \tag{7.2}$$

where $\tilde{\beta}(t)$ is the vector $(\beta_1(t), \beta_2(t), ..., \beta_P(t))$. If the predictor is a binary variable, $\beta_p(t)$ measures the difference in log(relative risk) between the two groups as a function of time. The advantage of this approach is twofold: On one hand it offers a straightforward way to investigate time-dependent structures, by testing for $\beta_{1p} = 0$. On the other hand, in case of PH rejection, it provides automatically an alternative model that fits adequately the data.

In case that all coefficients in $\tilde{\beta}$ vary ($\tilde{\beta} = \tilde{\beta}(t) = (\beta_1(t), \beta_2(t), ..., \beta_P(t))'$) the usual partial likelihood of the model takes the form:

$$L(\beta_1(t), \beta_2(t), ..., \beta_P(t); t) = \prod_{j=1}^{J} \frac{exp(\sum_{l=1}^{d_j} X_l \tilde{\beta}(t_j))}{[\sum_{s \in R_j} exp(X_s \tilde{\beta}(t_j))]^{d_j}} \tag{7.3}$$

where X_l is the covariate vector corresponding to lth failure at time j.

For one predictor p a function f_p is used so that $\tilde{\beta}_p(t) = \beta_0 + \beta_p f_p(t)$. The adequacy of this approach depends clearly on the choice of function f_p. There are several proposals about how to estimate the appropriate function f_p. Two methods that can be used - smoothing splines and fractional polynomials - are shortly presented in section 7.3.1 together with a new method using isotonic regression. But first, approaches will be presented related to Schoenfeld residuals, that are used to test proportional hazards.

7.3 Detecting PH departures under order restriction

Assume that if there is any PH violation, it follows a monotonic pattern. Starting from a time-varying Cox model (7.2), the Schoenfeld residuals provide a useful tool in detecting time-variation for the predictors of interest. That can be accomplished either graphically, or by applying a specific test as outlined below.

7.3.1 Smoothing Schoenfeld residuals scatterplot

The Schoenfeld residuals are defined at each unique failure times. In absence of ties they are equal to the difference between the observed covariate vector for an event at time $t_j, j = 1, ..., J$ and its expected value.

$$\tilde{r}_j = \tilde{X}_j - E(\tilde{X}_j|R_j) \rightarrow \tilde{r}_j = \tilde{X}_j - \frac{\sum_{l \in R_j} \tilde{X}_l e^{\tilde{\beta} X_l}}{\sum_{l \in R_j} e^{\tilde{\beta} X_l}}. \tag{7.4}$$

In the presence of P covariates, the Schoenfeld residuals $\tilde{r}$ can be presented as a $J \times P$ matrix.

Assume that for each variable p we have one estimated coefficient for each event time i.e. β_{pj}. Grambsch and Therneau [23] showed that if β_p is the coefficient from an ordinary PH Cox model, then

$$E(r^*_{pj}) + \beta_p \approx \beta_{pj}(t) \tag{7.5}$$

where $r^* = V_{\tilde{\beta}}^{-1} r$ are the *scaled Schoenfeld residuals* and $V_{\tilde{\beta}}$ is the variance matrix for the estimated coefficients $\tilde{\beta}$. This suggests to plot $r^*_{pj} + \beta_p$ versus time, to reveal the functional form of time variation. In case that the PH assumption holds, the residuals should form approximately a horizontal line at the constant coefficient β_p from model (7.1). One can used any kind of smoother for this purpose.

A popular choice are *natural cubic splines.* The principal idea is to split the time-axis by selecting an appropriate number of *nodes* and to fit piecewise polynomials. The choice of number of nodes (which determines the degrees of freedom) can affect the result, and no specific functional form is given. *Fractional polynomials* [51, 54] provide an interesting alternative, and result in a functional estimation of the time variation, but again one has to choose a set of exponents and maximal number of components.

The *isotonic smoother* provides an alternative to standard smoothers. It requires a monotonic trend, which is true for many prognostic factors. For example, consider a long-time therapy in which younger people respond better, but its prognostic value decreases with age. Additional considerations as for example the number of nodes need not be taken. The main advantage is that it detects jumps in risk for the time axis. Without any a priori information, the procedure returns some cutpoints, and segments the observational time in homogenous groups. The risk within each group is considered to be constant.

7.3.2 Grambsch and Therneau test and its isotonic version

Next to this graphical approach, Grambsch and Therneau introduced a version of the score test based on the weighted Schoenfeld residuals. Assume that all P predictor variables are time-dependent. The coefficient for the p variable has the time-depended form $\beta_p(t) = \beta_{0p} + \beta_{1p}(f_p(t) - \bar{f}_p)$ where $\bar{f}_p$ is the mean of $f_p(t)$ over time. Then, the PH hypothesis implies that $H_0 : \ \beta_{1p} = 0$.

Using matrix notation the test statistic takes the following form

$$GTtest_{Px1} = \frac{[(\tilde{t} - \bar{t})' r^*]^2}{diag(V_{\tilde{\beta}}) D \sum (t_i - \bar{t})^2} \tag{7.6}$$

where $V_{\tilde{\beta}}$ is the variance-covariance matrix for the estimated coefficients $\tilde{\beta}$. Each one of the resulting values corresponds to a variable and tests for time-dependency. This test is approximately χ^2 distributed with one degree of freedom for each tested coefficient.

This test can be thought of as a generalization of the least-squares statistic for estimating $\beta(t)$ given equation (7.5). Under the assumption of monotonic trend, one can substitute the function $f(t)$ by the isotonic function $is(t)$: if $is(\beta_p(t))$ is a consistent estimator of $\beta_p(t)$ then

$$is(r^*_{pj}) + \beta_p \approx \beta_{pj} \tag{7.7}$$

where $is(r^*_{pj})$ is the residual matrix divided in blocks that correspond to time intervals. Substituting r^* by the isotonic estimation $is(r^*)$ in equation 7.6 results to an *isotonic version of the GT test.*

Note that the idea to use piecewise constant and non overlapping time intervals to estimate $f(t)$ was first proposed by O'Quigley and Pessione [46]. However, as noted

in their paper, the investigator has to choose the partition of the time axis. Although the authors introduce some useful guidelines, the choice of the cutpoints remains rather subjective. Applying isotonic regression this disadvantage is bypassed. In section 7.5 the performance of isotonic transformation in the residuals is assessed and compared to the standard Grambsch and Therneau test.

7.4 Fitting the generalized additive model using isotonic smoothing techniques

Fitting smoothing splines in estimating $\beta(t)$ within the Cox model requires maximization of the *penalized partial likelihood function*. The result is a natural cubic spline, having nodes at each failure time point. The oscillation of the fitted spline increases as the penalty parameter decreases. This parameter need to be pre-specified and defines the degrees of freedom. With fractional polynomials, one has to fit a stratified Cox model where the unique failure time points $t_j, j = 1, ..., k$ determine the strata. At each such strata the corresponding covariate values are attributed and the new observational time is set to $t_{j+1} - t_j$. Using then X and $\tilde{f}(t)X$ as predictors, the stratified Cox model applied in the new data set will provide $\tilde{\beta}(t)$.

With step functions modeling time-varying effects is easier. Once the time-intervals are estimated the varying coefficients model (7.2) shall be estimated. Assuming that PAVA returns m time cutpoints regarding the effect of a variable, the time-varying coefficient for this variable takes the form:

$$\beta(t) = \beta_0 + \alpha_1 I_{t_1}(t) + \alpha_2 I_{t_2}(t) + ... + \alpha_m I_{t_m}(t) \tag{7.8}$$

$$I_{t_j}(t) = \begin{cases} 0 & \text{if } t <= t_j \\ 1 & \text{if } t > t_j \end{cases} .$$

The functional form of $\beta(t)$ has to be introduced in the model in order to estimate $\tilde{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_m)$. Standard likelihood based methods are applied for this purpose. Thereafter the usual Score test or the Likelihood Ratio test with m degrees of freedom can be applied to compare the PH model to the dynamic model, by testing all time-specific coefficients to be zero:

$$H_0 : \alpha_1 = ... = \alpha_m = 0. \tag{7.9}$$

The parameter α_i measures the increase (or decrease) in the risk from time t_{i-1} to time t_i on a logit scale.

It is very often the case that the time axis seems oversegmented. Some of the observed cutpoints do not correspond to an important increase (or decrease) in risk. One has to proceed to *a backward elimination* of the level sets. First the time groups containing few events (less than 10% of the total number of events) are deleted. Once these groups are eliminated, the likelihood ratio test can be applied to test one by one the coefficients $\alpha_i = 0$ in order to define the neighboring level sets that do not differ significantly. The deletion of a coefficient a_i and its time-interval I_{t_i} is equivalent to its union to the previous interval. The elimination proceeds by such time-interval unions, re-fits the Cox model and stops when all α_i are found to be significant. The $(1-\alpha)\%$ confidence band for a time varying predictor is expressed by

$$CI_{1-\alpha} = \tilde{\beta} \pm \sqrt{X^2_{df,1-\alpha/2} diag(ZV_{\tilde{\beta}}Z')} \tag{7.10}$$

where $V_{\tilde{\beta}}$ is the large sample variance-covariance matrix for $\tilde{\beta} = (\beta_0, \alpha_1, \alpha_2, ..., \alpha_m)$.

When more than one covariate is time-varying, the backfitting strategy is applied to fit the model. The general idea is to fit the time-varying coefficients allowing variation at one variable at time while the rest covariates remain time-independent

$$\beta^{it=1}(t) = (\beta_0^{it=1} + \tilde{\alpha}_1^{it=1} f_1(t), \beta_2^{it=1}, ..., \beta_P^{it=1})$$

where $f(t)$ is a step function. The likelihood ratio test will assess the gain in the fit i.e will test ${\alpha_1}^{it=1} = 0$. In case of evidence $f(t)$ is retained. In the next step all coefficients are reestimated, allowing now variation for the first two variables

$$\beta^{it=2}(t) = (\beta_0^{it=2} + \tilde{\alpha}_1^{it=2} f_1(t), \beta_0^{it=2} + \tilde{\alpha}_1^{it=2} f_2(t), ..., \beta_P^{it=2})$$

where only $f_1(t)$ is estimated from the previews step and held constant in step 2. The procedure goes on like that updating in each iteration only the coefficients. Such loops are repeated until a small change in the likelihood is achieved.

7.5 Simulation study

A simulation study was conducted to explore the properties of the new proposal for testing proportional hazards applying the isotonic version of Grambsch and Therneau test (7.6). This section focuses on revealing the advantages of the isotonic GT test against the conventional test. When forming assumptions about the functional form of the regression, I tried to be as consistent as possible with situations frequently observed in clinical studies. The simulations are designed to avoid ties.

Only the case of a simple binary predictor is considered. One proportional and three non proportional hazard models are analyzed. In the baseline group the covariate has been set $X = 0$ and the hazard $\frac{e^{-4}}{1+e^{-4}}$. The treatment group has $X = 1$ and hazard $\frac{e^{-4+\beta(t)}}{1+e^{-4+\beta(t)}}$. Each group contains 100 observations.

To generate the data sets, I proceed separately in each group (treatment or baseline) as follows: starting from time=1 the number of failures is calculated using the hazard function. For the observations remaining at risk, the number censored observations is calculated, as a random binary process. The procedure is repeated for time=2 and stops when no more observations remain at risk. The censoring probability used here was 0.5%. To model dynamic structures that decrease with time the following

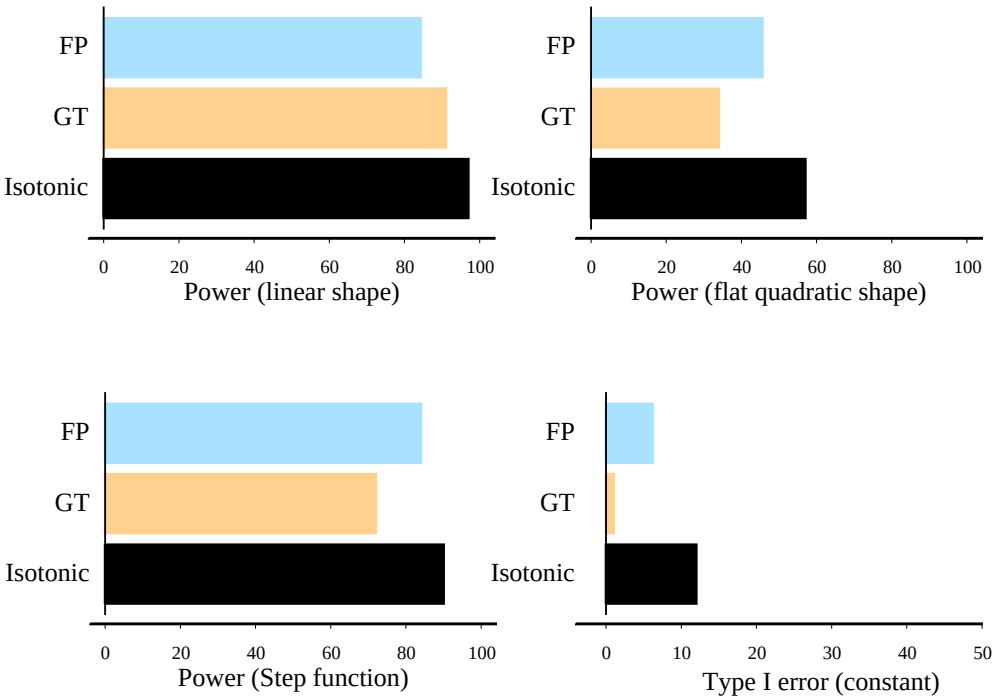

Figure 7.1: Simulations study for survival data. Compare in terms of power (first three figures) and type I error (last figure) the Grambsch and Therneau test (GT test), the fractional polynomials test and the isotonic version of GT test.

scenarios are made:

Linear: a decreasing linear time-dependency where $\beta(t) = -0.02t + 1$.

Quadratic: where $\beta(t) = -0.04t - 0.004t_2 - 1$ representing a decreasing umbrella shape

Step function: having shift at t=24 and $\beta(t) = \begin{cases} 1.5 & t \leq 24 \\ 0 & t > 24 \end{cases}$.

Constant: $\beta(t) = 1$ for estimating the properties in case where the PH assumption is not violated.

Simulations under the first three functions will give information about the power of the compared tests, whereas with the constant function the type I error will be assessed. Three test are compared: a) a test based on fractional polynomials model described in [9] b) the GT test (7.6) assuming linear transformation for time and c) the isotonic version of GT test (7.5). The results are presented in figure 7.1.

The isotonic test presents the best power for all non-constant functions, whereas the conventional GT test gives the lowest power. For every shape the power from fractional polynomial is lower that this from isotonic regression. One would expect that this advantage of the isotonic test is eliminated in case of a non-monotonic function. The more flexible approach as the fractional polynomials should present a better performance in case of the flat quadratic function. This is not the case, as outlined in figure 7.1: isotonic regression gives higher power for this shape as it gives for a step function. However, the price one has to pay for the increasing power in the isotonic test is a higher type I error.

7.6 Case Study in time-varying Cox model

The data set used to illustrate the above approaches contains 141 observations from children having acute leukemia (ALL). The endpoint was overall survival time. The probability to die within a period of 7 days to about 10 years follow up, has been found to be dependent upon the following binary variables:

- Remission after the first induction (REMI, 1: yes)
- ALL relapse after the first Chemotherapy (RELP, 1: yes)
- The size of massive spleen below the rib (MSPS, 1: > 1 cm)
- White blood cell count (WBC, 1: $> 60.6\ 10^9/L$)

Survival time is measured in years. The main of the study was to estimate if there is a time variation in the effect of MSPS, and in case of evidence to describe this variation. The sample is characterized by a high event rate (122/141), and the value of deviance in absence of any predictor is estimated to be 1037.59.

The Cox PH model with forward LR selection has been applied and table 7.1 shows the estimated coefficients. Time variation in the predictive value of MSPS has been tested applying Grambsch and Therneau test, Kaplan-Meier curves (figure 7.6) and smoothing the Schoenfeld residuals using splines, fractional polynomials and isotonic regression (figure 7.3).

Table 7.1: Acute lymphoblastic leukemia study: The PH Cox model. The deviance is 957.08 with 5 degrees fo freedom.

Variables	Coefficients	SE	p-value
RELP	0.507	0.219	0.021
REMI	-0.991	0.387	0.010
MSPS	0.549	0.203	0.007
WBC	0.785	0.232	0.000
CONTS	-0.974	0.362	0.007

These different methods are more or less in agreement: there is a dynamic effect for MSPS. The Grambsch and Therneau test results in a test value 3.930 and the corresponding p-value is 0.047. Fitting the varying coefficients model using splines, the constant predictor lies out of the confidence bands for more than 10% of the total number of events (figure not shown). There is a decreasing positive prognostic value for MSPS. Children that do not have massive spleen have better prognosis that decreases progressively, and after about four years the direction of the prognosis changes. This conclusion is quite strange and against any biological plausibility. However a possible explanation could be the following: perhaps many children get

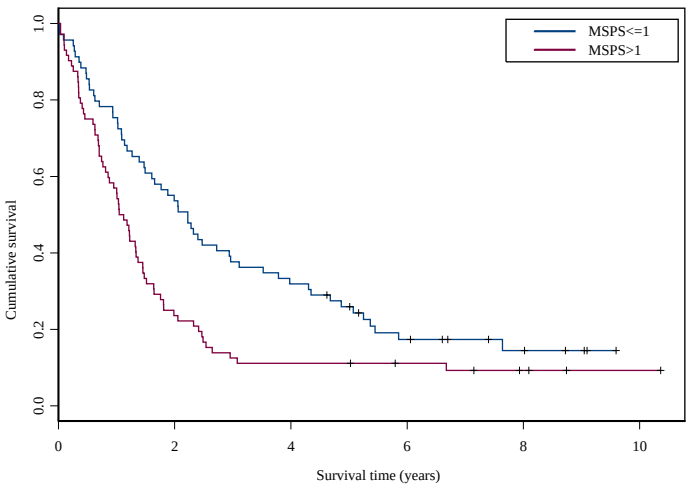

Figure 7.2: Kaplan-Meier cumulative survival curves for MSPS.

a very intensive chemotherapy that is effective against the tumor but is also too burdensome. So, it may cause a preliminary death to many children. But once a child overcomes that crucial period and does not relapse, it has the best chances to survive.

By isotonizing the Schoenfeld residuals (figure 7.3) the appropriate time-cutpoints are revealed. The confidence intervals correspond to fractional polynomials. However some of the resulting steps contain very few events and therefore do not offer a lot of information while increasing the degrees of freedom. Each group is restricted to contain at least 10% of the total number of events. After elimination of those groups, model 7.8 can be written for the resulting time-cutpoints:

$$\beta(t) = \beta_0 + \alpha_1 \cdot I_{1.98}(t) + \alpha_3 \cdot I_{3.52}(t). \tag{7.11}$$

The time-stratified Cox model can now be fitted again to estimate whether some

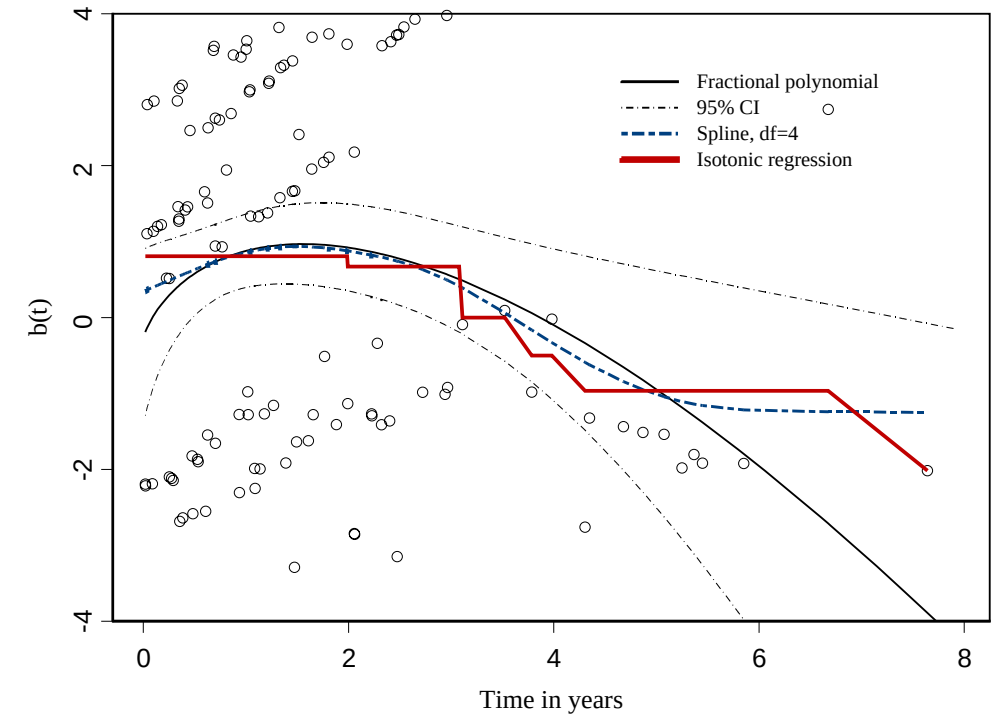

Figure 7.3: Smoothing the scaled Schoenfeld residuals for MSPS.

of the $I_t(t)$ variables are non significant predictors and to delete them. Recall that any coefficient α that is found to be non significant corresponds in a union of the above defined time-level sets (table 7.2). Note that p-value correction has to be considered because of the multiple comparisons i.e. $a = 1 - \sqrt[c]{0.95}$, c the number of time-segments.

Both time-interval variables $I_{1.98}$, $I_{3.52}$ are significant. The fitted function with the corresponding confidence bands are presented in figure 7.4. The dynamic form $\beta(t)$ for MSPS is:

$$\beta_{MSPS}(t) = 1.57 - 1.18 \cdot I_{1.98}(t) - 2.85 \cdot I_{3.52}(t). \tag{7.12}$$

Table 7.2: Elimination of the time level sets for MSPS dynamic coefficient.

Coefficient	Deviance	p-value
β_0	957.08	0.0000
α_1	928.59	0.0000
α_2	915.64	0.0013

The final achieved deviance have been estimated 915.64, that yields an overall LR test for PH of 41.44 ($p < 0.001$). Finally the model containing all the significant predictors and their time dependent effects is:

$$h(t) = h_0(t)e^{\beta(t,x)}$$

where

$$\beta(t,x) = 0.490 \cdot RELP - 1.105 \cdot REMI +$$
$$+[1.568 - 1.184 \cdot I_{1.98}(t) - 2.851 \cdot I_{3.52}(t)] \cdot MSPS + 0.235 \cdot WBC - 0.604 \cdot CONTS$$

7.7 Extensions

One can imagine implementations of isotonic regression in several approaches regarding survival settings. John O'Quigley [46] for example introduced a test for proportional hazards based on the model: $\lambda(t) = \lambda_0(t)exp[(\tilde{\beta} + \Psi\tilde{\theta})'X]$ The matrix $\Psi = diag(\tilde{\psi}_1, \tilde{\psi}_2, ..., \tilde{\psi}_P)$ is a score matrix determined by the user. Obviously if $\tilde{\theta} = 0$ the proportional hazards model is recovered. The model is fitted using the stratified Likelihood, where arbitrary time cutpoints define the strata, and a sort of score test is applied to test for $\tilde{\theta} = 0$. Isotonic regression can be easily introduced into this context and improve the performance of this approach.

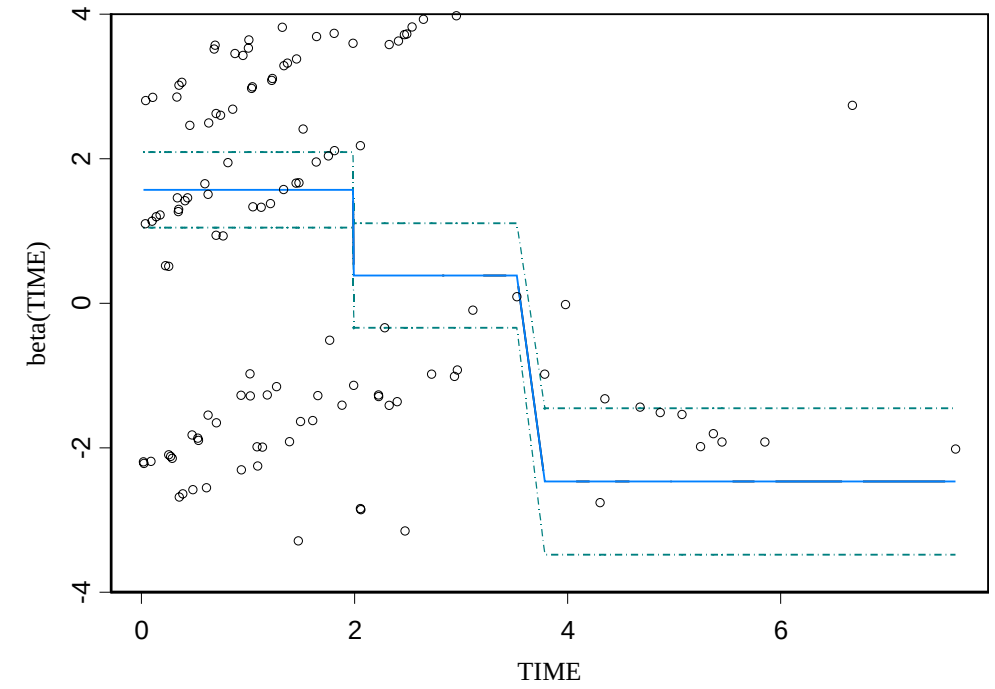

Figure 7.4: Isotonic fit for time-dependent coefficient for MSPS.

Another assumption undertaken by the Cox model is that each variable enters the model linearly, assumption that may also be violated. This case entails that both coefficient and RR depend on the variable ($\beta = \beta(X), RR = exp(\beta(x_i) - \beta(x_j))$). The adequacy of the linear form of a predictor in the Cox model can be visualized by smoothing the martingale residuals plotted against the predictor. If the shape seems not linear, the predictor has to be transformed. An approach similar to this used for modeling time variation can be applied to model properly non linear predictors.

An alternative approach that uses step functions in modeling dynamic structures is accomplished with CART [67]. The main advantage provided is that the time-

cutpoints are not prespecified, but the pruning parameter has to be calculated through cross validation. The PAVA algorithm can modify the splitting criteria, to include monotonicity restrains if so required.

8 SUMMARY

8.1 Summary

Categorizing continuous variables arises as an important task in statistical analysis, especially in analyzing dose-response relationships. Creating meaningful groups of the predictor variables regarding the outcome variable is desirable in many settings, especially if the form of the relationship is unknown. However it is not always obvious how many groups should be build and where the cutpoints should be placed. Usually more than one explanatory variable has to be included in the analysis, and therefore one has to apply an appropriate statistical model. For this purpose we need a simple approach to model the data without many requirements. Another important issue in statistical analysis and especially in toxicology studies is proving a dose response relationship: increasing response probability with increasing predictor variable. This theses deals with cases where categorization of numerical or categorical predictor variables results as an effect of the dose-response relationship.

Isotonic regression is an alternative proposal when one wishes to establish a dose-response relationship, categorize continuous variables and estimate threshold values. The only assumption for this approach is the monotonicity in the response variable.

The isotonic regression summarizes the description of n observations to l categories (level sets or solution blocks) by automatically splitting the predictor in constant risk groups. The result is always a step function, and therefore the isotonic regression can be used to fit a changepoint model. The Pooled Adjacent Violators Algorithm (PAVA) is used to fit the data.

In relation to model fitting and testing, some problems arise when the response is binary, and in the present work the difficulties are highlighted and some proposals to solve them are given. Regarding isotonic regression and binary response, the isotonic test for trend, the reduced isotonic model, multidimensional isotonic models and methods to assess threshold limit values are discussed.

The isotonic framework provides a reliable test for trend which unlike other widely used tests (the Cochran-Armitage test for example) is independent of any monotonic transformation of the dose variable and does not assume a linear shape. However the proposed large sample approximation (a weighted chi-square distribution) does not hold when the overall response probability is less than 5% and thus exact methods are proposed in order to assess the correct p-value. In a simulation study it has been shown that the isotonic likelihood ratio test is more powerful than the Cochran-Armitage test, the Wilcoxon test and the Iso-chi-squared test.

The model resulting from PAVA can become more parsimonious if the level sets which correspond to a non significant change for the response variable are eliminated. This model is called reduced isotonic regression. That can be accomplished by two means: a sequence of Fisher tests for the adjacent 2×2 tables or the application of a variation of a "closed testing" procedure. The correction for multiple comparisons is made for the first method by an a-priori estimation of the overall significance level in a permutation procedure. In the second method the control for the expense of the type I error is effected by the closure principal. To select between full isotonic and reduced model, a procedure based on parametric bootstrap is proposed. A

simulation study proved that when the maximal coefficient of determination $\bar{R}^2_{max}$ for the analyzed data set is at least 50% and the data can be represented by a step function, the reduced monotonic regression controls successfully the trade off between model complexity and goodness of fit.

When more than one predictor is to be taken into account an additive isotonic model can be applied. Alternatively, an isotonic-surfaces model is proposed. This can be estimated by an iterative version of the Pooled Adjacent Violators Algorithm. The result is a sequence of surfaces which is monotonic in every dimension. This approach models interaction and categorizes the predictors in "multivariate" groups by combining them regarding restrictions to the outcome variable. This approach is very useful since, unlike the additive model, it can be easily combined with the reducing procedures to give a simple and interpretable model. However, for practical reasons a maximum of three predictors can be taken into account.

A special aspect in analyzing dose-response relationships for a compound known to have harmful health effects, is to estimate a threshold limit value (TVL). On this regard a "hockey stick" threshold model is usually used. As alternative the use of a step function model by fitting the data using isotonic regression is proposed. A set of candidate threshold values is returned, and some threshold value estimation procedures are studied here. One of them starts from the isotonic model and applies the likelihood ratio test to detect the threshold value (method 1). Method 2 is based on the reduced isotonic regression. The performance of these two approaches is outlined in a simulation study under different scenarios and their properties are explored with categorical predictors. It is concluded that these methods possess a satisfactory power to reject the constant risk assumption, when a dose-response relationship exists as well as to estimate the actual threshold. Some limitations regarding the sample size and the force of trend are also discussed. A third method has also been presented. This modifies the closed testing procedure for the special case of thresholds, by setting one end of the regression line conditional to the other.

All three threshold value estimation methods can be combined with the isotonic-surfaces model to provide thresholds, taking into account interactions between the predictor variables.

The use of isotonic regression and its reduced version can also be extended to other settings. The capability of isotonic regression to be implemented in several models is outlined by describing how isotonic regression can model and test time-varying effects in Cox regression. The monotonic variation in the impact of a predictor included in the model during an observational period can be represented by a step function. An estimation of the time-dependent effect in the extended Cox model is presented based on isotonic regression framework. Smoothing the Schoenfeld residuals plotted against time applying PAVA, can reveal the changepoints without any a priori information about their location. The corresponding step function is then introduced in the model. The power of the Grambsch and Therneau test (which tests for time-variation in the effect of the predictors) can be improved if the isotonic transformation for the Schoenfeld residuals is used. Although this test appears to increase the type I error, its power is higher compared to conventional Grambsch and Therneau test and tests based on fractional polynomials.

In summary it arises that isotonic framework is characterized by simplicity and stability. The main drawback underlying its application is the lack of asymptotic support in testing. This can make the use of isotonic models cumbersome since exact or bootstrap methods need to be used.

8.2 Zusammenfassung

Die Kategorisierung von stetigen Merkmalen erweist sich als eine sehr wichtige Aufgabe innerhalb statistischer Analysen, ganz besonders in der Analyse von Dosis-Wirkungs-Beziehungen. Es ist in vielen Situationen wünschenswert, sinnvolle Gruppen innerhalb der Prädiktorvariablen zu finden und zu bilden. Dennoch bleibt oft die Frage, wieviele Gruppen gebildet werden sollen und wo genau die jeweiligen Grenzwerte liegen sollen. Wird mehr als eine erklärende Variable in die Analyse eingeschlossen, muss ein passendes statistisches Modell gefunden und angewendet werden. Wünschenswert wäre ein möglichtst einfacher Ansatz zur Modellierung der Daten, der wenige Voraussetzungen erfordert.

Ein wichtiges Problem in der statistischen Analyse, besonders in toxikologischen Studien, ist der Nachweis von Dosis-Wirkungs Beziehung, d.h. wenn mit einem Ansteigen der erklärenden Variablen auch eine Steigung der Wahrscheinlichkeit für das Auftreten der Zielgrösse einhergeht. Diese Doktorarbeit behandelt Situationen, bei denen die Kategorisierung von stetigen oder kategorialen Variablen als Ergebnis der Analyse von Dosis-Wirkungs-Beziehung (DWZ) einhergeht.

Isotone Regression liefert einen alternativen Ansatz, um eine Dosis-Wirkungs-Beziehung nachzuweisen, stetige Merkmale zu kategorisieren und Grenzwerte zu schätzen. Die einzige Voraussetzung bei diesem Ansatz ist die Monotonie in der Zielgrösse. Die isotone Regression fasst n verschiedene Beobachtungen in l verschiedene Blöcke zusammen, indem sie die Prädiktoren in Gruppen mit jeweils konstantem Risiko einteilt. Da das Resultat eine Treppenfunktion ist, kann die isotone Regression benutzt werden, um Schwellenwerte zu erkennen. Der Pool Adjacent Violators Algorithmus (PAVA) setzt diesen nicht-parametrischen Ansatz um.

Bei binärer Zielgrösse entstehen hier Probleme bezüglich der Modellschätzung und der Modelltests. Ein Hauptaugenmerk dieser Arbeit liegt auf der genauen Unter-

suchung dieser Probleme und bietet teilweise Lösungsvorschläge an. Bezüglich der Isotonen Regression mit binärer Zielgrösse werden mehrere Gebiete genauer diskutiert: das reduzierte isotone Modell, das multidimensionale isotone Modell und Methoden zur Bewertung von Schwellenwerten.

Der isotone Ansatz liefert auch einen Trendtest, der, im Gegensatz zu anderen Trendtests (wie z.B. der Cochran-Armitage Test), unbeeinflusst von monotonen Transformationen der Dosisvariable ist und auch keinen linearen Zusammenhang voraussetzt. Die vorgeschlagene asymptotische Verteilung (eine gewichtete Chi-Quadrat Verteilung) liegt nicht vor, wenn die Wahrscheinlichkeit für das Auftreten der Zielgrösse unter 5% sinkt. Hier sind exakte Methoden erforderlich, die einen genauen P-Wert bestimmen. In einer Simulationsstudie konnte gezeigt werden, dass dieser isotone Likelihood-Quotienten-Test eine grössere Power besitzt als der Cochran-Armitage-Test, der Wilcoxon Test und der Iso-Chi-Quadrat-Test.

Das isotone Modell kann noch vereinfacht werden, indem die Blöcke, die einen nichtsignifikanten Einfluss haben, zusammengefasst werden. Hierzu wurden zwei verschiedene Methoden verglichen: einer Sequenz von exakten Fisher-Tests für die benachbarten Blöcke sowie eine Variante eines "closed testing" Prozesses. Die Korrektur für multiple Vergleiche des P-Wertes wird bei der ersten Methode durch eine apriori Schätzung des Gesamtsignifikanzniveaus mittels eines Permutationsverfahrens erreicht. Bei der zweiten Methode ist die Kontrolle des Fehlers erster Art durch das Einschliessungsverfahren beeinflusst. Um letztendlich zwischen dem vollen Modell und seinem reduzierten Äquivalent zu entscheiden, wurde ein parametrisches Bootstrap-Verfahren vorgeschlagen. In einer Simulationsstudie zeigte sich, wenn der maximale Koeffizient $\bar{R}^2_{max}$ für die Daten mindestens 50% betragen soll und die Daten durch eine Treppenfunktion dargestellt werden können, dann stellt die reduzierte isotone Regression einen guten Kompromiss zwischen hoher Modellkomplexität und Güte dar.

Wurde mehr als eine Prädiktorvariable berücksichtigt, dann kann ein additives Modell verwendet werden. Alternativ hierzu wurde ein "isotone-Fläche"-Modell vorgeschlagen. Dieses kann mittels einer iterativen Version des PAVA geschätzt werden und resultiert in einer Sequenz von Flächen, die in jeder Dimension monoton sind. Es werden hierbei Interaktionen modelliert und die Prädiktoren in multidimensionale Gruppen bezüglich bestimmter Einschränkungen der Zielgrösse unterteilt. Dieser Ansatz ist sehr elegant, da er, im Gegensatz zum additiven Modell, leicht mit dem Reduzierungsverfahren kombiniert werden kann, und so einfache und leicht interpretierbare Modelle liefert. Aus praktischen Gründen können hierbei jedoch nur bis zu maximal drei Prädiktorvariablen in das Modell genommen werden.

Die Schätzung von Schwellenwerten für Stoffe, die sich bekanntermassen negativ auf die Gesundheit auswirken, ist von grösster Bedeutung in der Epidemiologie. In diesem Zusammenhang wird normalerweise ein "hockey stick"-Schwellenwertmodell angewandt. Alternativ wurde ein Modell vorgeschlagen, das auf dem Resultat einer isotonen Regression, also einer Treppenfunktion, basiert. Es gilt aus einer Reihe von Schwellenwerten einen Wert auszuwählen. Verschiedene Schätzer wurden untersucht. Eine Methode setzt beim isotonen Modell an und führt einen Likelihood-Quotienten-Test durch. Die zweite Methode basiert auf der reduzierten isotonen Regression. Die Leistung der beiden Algorithmen wurde in einer Simulationsstudie kurz dargestellt. Die Eigenschaften wurden hierzu in verschiedenen Situationen mit kategorialen Einflussgrössen untersucht. Falls eine Dosis-Wirkungs-Beziehung vorliegt, erweisen sich diese zwei Methoden als ausreichend mächtig, um die Hypothese "das Risiko ändert sich nicht" zu verwerfen. Sie sind zufriedenstellend bezüglich ihrer Fähigkeit, den Schwellenwert zu schätzen. Einige Einschränkungen, entstehend aus der Stichprobengrösse und dem Einfluss des Trends, wurden ebenso diskutiert. Als dritte Methode wurde eine Modifikation der "closed testing" Verfahren vorgeschlagen. Dabei stellt sie ein Ende der Regressionslinie in Abhängigkeit zum anderen Ende dar. Alle drei Schwellenwertschätzer können mit dem "isotone-

Fläche"-Modell kombiniert werden, unter Berücksichtigung von Interaktion zwischen den Einflussgrössen.

Die Implementierung der isotonen Regression in verschiedene Modelle wird exemplarisch hervorgehoben in einer Anwendung der isotonen Regression im Cox-Modell mit zeitveränderlichen Effekten. Die monotone Variation des Einflusses eines Prädiktors über eine bestimmte Zeitperiode kann durch eine Treppenfunktion dargestellt werden. Eine Schätzung der zeitabhängigen Effekte im erweiterten Cox-Modell, basierend auf isotoner Regression, wurde beschrieben. Werden geglättete Schoenfeld-Residuen gegen die Zeit in einem Diagramm eingetragen, unter Zuhilfenahme von PAVA, können auch ohne a-priori Informationen über ihre Lage, Grenzwerte gefunden werden. Die Power des Grambsch-Therneau Tests zu Untersuchung der Veränderung des Einflusses eines Prädiktors über die Zeit, kann verbessert werden, wenn die Schoenfeld-Residuen mittels PAVA transformiert werden. Obwohl dieser Test scheinbar den Fehler erster Art erhöht, ist seine Power höher im Vergleich zu herkömmlichen Grambsch-Therneau-Test sowie zu Tests, die auf fraktionalen Polynomen basieren.

Abschliessend bleibt zu sagen, dass sich meiner Meinung nach die Analyse mittels isotoner Methoden durch Einfachheit und Stabilität auszeichnet. Ihr Hauptnachteil liegt in dem Mangel an asymptotischen Hilfestellungen beim Testen. Dies kann die Verwendung von isotonen Modellen erschweren, da dann exakte oder bootstrap Methoden verwendet werden müssen.

A APPENDIX: Software implementation in S+

This chapter provides help on the isotonic library `isotonic.S.library`. This library contains original functions in S+ language, except for `ccaddir.cov` programmed by Morton-Jones for additive isotonic models. This library is available in my personal web-page.

Chapter 2: MONOTONIC REGRESSION

The data set `Munich` contains the variables `CBR, ZEIT1,GESAMT, RAUCH`: event, time since first exposure, total dust concentration and smoking habits. The data set `Mu.r` is the subset for smokers. The basic generic function is `isotone.simple.order.fun` having arguments the observed proportions, the weights and the trend option (increasing or decreasing).

```
> attach(Mu.r)
> y_table(ZEIT1,CBR)[,2]
> w_table(ZEIT1)
> isot.TIME_isotone.simple.order.fun(y/w,w,trend="I")
> isot.TIME_rep(isot.TIME,w)
```

The isotonic plot in figure 1.1 has been made applying:

```
> plot(ZEIT1,isot.TIME,type="l")
```

Chapter 3: TESTS FOR TREND IN BINARY RESPONSE

Analysis of the para-aramid data (section3.4)

The function `CA.test` calculates the Cochran-Armitage test and returns the test statistic and the p-value (two sided) according to the chi-square distribution with correction for skewness.

```
>ni<-c(137, 133, 132, 137, 92)
>pi<-c(1,1,1,4,4)/ni
>CA<-CA.test(pi,ni, score=1:5) # the index as score
>CA$gamma
[1] 0.02309635 # the coefficient of skewness
```

The function `isoR.fun` assesses the isotonic likelihood ratio test R. Select `trend="D"` for decreasing trend.

```
> isoR.fun(pi,ni,trend="I")
[1] 6.471335
```

Equivalently the iso-chi-squared test:

```
> gautamiso.test(pi,ni)
$statistic:
[1] 7.083767
$rejectH0:
[1] 1
```

The value `$rejectH0` is 1 for rejecting H_0 and it is assessed according to approximations due to Gautam (see table at the end of Appendix).
A p-value estimated by 10 000 simulations can be calculated for every tests statistic applying the function `exact.p.value.fun`:

```
> exact.p.value.fun(pi,ni,CA.test,1:5)
[1] 0.0182
```

The function `poly3.fun` calculates the Poly-3 test. On analyzing the example in section 3.5.3:

```
> attach(EMIAT)
> Poly3(Y=ACM, time=follow, X=LVEF6, score=c(6:1))
$statistic:
[1] 2.237923
$p.value:
[1] 0.01261305
```

Chapter 4: REDUCED MONOTONIC REGRESSION

Analysis of MAK study using cumulative exposure (section 4.5)

The reduced isotonic regression using Fisher's test (see section 4.2.1) can be obtained by `uniisoRED.fun`. This function returns the isotonic and the reduced isotonic estimators, estimation for ε^*, a test for trend based on 10 000 permutations and a graph. Regarding the example analyzed in this chapter the function

```
> cumRES <- uniisoRED.fun(response = CBR, predictor = cumex,
data = Mu.r, reapplic = 10, test = F, graph = F)
```

gives the output:

```
Univariable Isotonic Regression
Likelihood functions
-2log(Likelihood) for H0: 1058.17193971207
-2log(Likelihood) for Isotonic regression: 983.057208460027
-2log(Likelihood) for reduced regression: 988.327637917929
sls
 0.0238
p-value for H0: 0.00009999
```

whereas `cumRES$out` is a database containing the estimated values. The function `unicompareRI.fun(reduced.object, sls, reapplic)` is used to compare the reduced model to its isotonic, having as arguments, the output of the function `uniisoRED.fun`, the ε^* (`sls`) and the number of bootstrap samples used (`reapplic`). The elimination based on closed procedure has been made on the use of `closed.test` function

```
> out <- as.data.frame(cumRES$out)
> attach(out)
> pi <- sort(unique(isotonic.pi))
> ni <- table(isotonic.pi)
> out <- as.data.frame(cumRES$out)
> attach(out)
> pi <- sort(unique(isotonic.pi))
> ni <- table(isotonic.pi)
> reduced.pi.closed <- closedtest.fun(pi, ni)
> reduced.pi.closed
       pfin  nfin
 0.19327542  618
```

```
0.35497835  231
0.56337606   71
```

Chapter 5: MULTIDIMENSIONAL MONOTONIC MODELS

Analysis of MAK study using CBR and ZEIT1 (section 5.7)

The variables `qgesamt` and `qzeit` correspond to grouped variables total dust and time since first exposure. Two dimensional isotonic estimates can be curried out by applying the function:

```
> d2Mu.r<-d2.isotRED.fun(response=CBR, predictor1=qgesamt, predictor2
=qzeit, plot = T, Likelihood = T, reduced = T, w, dataset)
```

The function automatically assesses the correct ε^* or it can be defined with the argument `sls`, or by applying the function `SLS.fun`. The result contains the isotonic and the reduced fitted values `fitted.iso, fitted.red`. Global and partial significance in a two dimensional model can be assessed by the function:

```
> test<-d2.isotest.fun(d2isotREDoutput=d2Mu.r, response=CBR,
predictor=qgesamt, predictor2=qzeit, reapplic=5000, H0simulations = T,
plot = T, CIbands = T, conditional = F, CIbandsH1= F, dataset)
> test
[1] 0.00019996
```

To compare the reduced model to the full isotonic the function

```
> compareRI.fun(d2.isotREDoutput=d2Mu.r, reapplic=5000)
```

can be applied.

The additive isotonic model is computed by the function `ccaddir.cov` (programmed by Morton-Jones)

```
> AIMmodel <- ccaddir.cov (X=cbind(GESAMT,ZEIT), w=rep(1,920), Y=CBR,
Z=RAUCH)
```

I added the function

```
> AIMtest.fun(predictor=GESAMT, response=CBR, predictor2=ZEIT1,
 linterm=RAUCH, Mcdata=5000)
```

which assesses the *conditional* significance for one predictor included in the model.

Chapter 6: THRESHOLD VALUE ESTIMATION

Searching for threshold in MAK study (section 6.5)

The approach for estimating thresholds referred to as **method 1** is implemented in `uniisoTHRES.fun`

```
> method1<-uniisoTHRES.fun(response=CBR, predictor=cumex, data=Mu.r)
> method1$threshold
[1] 5.04
```

The function `AIMthres.fun` assesses a threshold value using **method 1** in the partial fitted function for the first variable included in the model. For example:

```
> AIMthresh.fun(AIM=AIMmodel, data=Mu.r, response=CBR,
threshold.variable = 1, lpred=RAUCH)
```

where the arguments are: `ccaddir.cov` function output, the data set, the response

variable, the position of the variable for which the threshold should be estimated, and eventually the linear predictor included in the semiparametric model.

The function `BCCI.fun` estimates bootstrap corrected and accelerated confidence intervals.

```
> BCCI.fun  <- function(xdata, boot = 1000, theta.function = theta, ...,
measure = "threshold locations", a = 0.95, graph = F)
```

The user need to specify the number of bootstrap samples (`boot`) and the data set (`xdata`) which has in the first column the response variable and then the predictor.

Chapter 7: MONOTONIC REGRESSION IN SURVIVAL ANALYSIS

Analysis of Acute Leukemia Study (section 7.6)

The Cox model allowing isotonic time variation for MSPS has been fitted by:

```
> Cox.isomodel <- cox.isph(data, Trend=c(NA,NA,"D",NA,NA))
```

The data have to be in the following order: survival time, status, covariates.

```
> Cox.isomodel$sres  #returns the Schoenfeld residuals for all
covariates and the weighted isotonic estimation of Schoenfeld residuals
(only for time-varying variables)
> Cox.isomodel$newdata   # the starting data base where at the end is
added time contrast for the time-varying variables
> Cox.isomodel$time.breaks   # returns the time cutpoints
```

The function `my.cox.zph` assesses the isotonic test for time variation based on Schoenfeld residuals (equation 7.6). This function has been used in the simulation

study in section 7.5.

Table A.1: Approximate critical values for significance level α=0.2, 0.1, 0.5, 0.01 for the iso-chi-squared statistic W. (derived by S. Gautam [21])

		α			
		.2	.1	.05	.01
	3	2.50	3.67	4.91	7.96
	4	2.99	4.27	5.59	8.61
	5	3.32	4.67	6.06	9.17
k	6	3.57	4.99	6.38	9.62
	7	3.80	5.26	6.70	9.96
	8	4.02	5.48	6.97	10.23
	9	4.17	5.69	7.17	10.60
	10	4.33	5.85	7.35	10.87

Bibliography

[1] Armitage P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics*, **11**: 375-386

[2] Bacchetti P. (1989) Additive isotonic models. *Journal of American Statistical Association*, **84**: 289-294

[3] Barlow RE., Bartholomew DJ., Bremner JM., Brunk HD. (1972) *Statistical Inference Under Order Restrictions.* Wiley, New York.

[4] Bailer J., Portier C. (1988) Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. *Biometrics*, **44**: 417-431

[5] Bauer P. (1991) Multiple Testing in Clinical Trials. *Statistics in Medicine*, **10**: 871-890

[6] Becher H. (2002) Analysis of continuous covariables in epidemiological studies: Dose-response modeling and confounder adjustment. *Biometrical Journal*, **44**: 684-699

[7] Begun J., Gabriel K. (1981) Closure of the Newman-Keuls Multiple Comparisons Procedure. *Journal of the American Statistical Association*, **76**: 241-245

[8] Bender R., Lange S. (2001) Adjusting for multiple testing - when and how? *Journal of Clinical Epidemiology*, **54**: 343-349

[9] Berger U., Gerein P., Ulm K., Schfer J. (2000) On the use of fractional polynomials in dynamic cox models. *Discussion Paper No. 207* Sonderforschungsbereich 386 Ludwig-Maximilians-University, Munich.

[10] Breslow N., Day N. (1987) The design and analysis of cohort studies. *International Agency for Research on Cancer*, scientific publication no **82**. Lyon, France

[11] Carpenter J., Bithell J. (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, **19**: 1141-1164

[12] Chuang-Stein C., Agresti A. (1997) Tutorial in biostatistics: A review of tests for detecting amonotone dose-response relationship with ordinal response data. *Statistics in Medicine*, **26**: 2599-2618

[13] Collings B., Margolin B. (1981) Analyses for binomial data, with application to the fluctuation test for mutagenicity. *Biometrics*, **37**: 775-794

[14] Cox C. (1987) Threshold dose-response models in toxicology. *Biometrics*, **43**: 511-523

[15] DFG (2001) *List of MAK and BAT Values 2001.* Wiley-VCH, Weinheim.

[16] Efron B., Tibshirani R. (1993) *An introduction to the Bootstrap.* Chapman & Hall, London.

[17] Faraggi D., Simon R. (1996) A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Statistics in Medicine*, **15**: 2203-2213

[18] Fleiss, J. L. (1981) *Statistical Methods for Rates and Proportions.* Second Edition J. Wiley, New York.

[19] Friedman JH., Tibshirani R. (1984) The monotone smoothing of scatterplots. *Technometrics*, **26**: 243-250

[20] Galloway S., Clark G. (1996) Practical p-value adjustment for optimally selected cutpoints. *Statistics in Medicine*, **15**: 103-112

[21] Gautam S., Sampson A., Singh H. (2001) Iso-chi-squared testing of 2 x k ordered tables. *The Canadian Journal of Statistics*, **29**: 609-619

[22] Gebhard F. (1970) An algorithm for monotone regression with one or more independent variables. *Biometrika*, **57**: 263-271

[23] Grambsch P., Therneau T. (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**: 515-526

[24] Graubart B., Korn E. (1987) Choice of column scores for testing independence in ordering 2 x K contingency tables. *Biometrics*, **37**: 471-476

[25] Greenland S., Michels K., Robins J, et al. (1999) Presenting statistical uncertainty in trends and dose-response relations. *American Journal of Epidemiology*, **149**: 1077-1086

[26] Greenland S. (1995) Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology*, **6**: 356-365

[27] Greim H. (1999) *Occupational toxicants. Critical data evaluation for MAK values and classification of carcinogens.* Commission for the investigation of health hazards of chemical compounds in work area. Vol 12. Willey-VCH, Weinheim.

[28] Harrell FE., Lee KL., Mark DB. (1996) Tutorial in Biostatistics, multivariable prognostic models: issues in developing models, evaluation assumptions and

adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**: 361-387

[29] Hastie T, Tibshirani R. (1990) *Generalized additive models.* Chapman & Hall, London.

[30] Holm S. (1979) Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, **92**: 299-306.

[31] Hothorn L. (1999) Trend tests in epidemiology: P-values or confidence intervals? *Biometrical Journal*, **41**: 817-825

[32] IARC Monographs (1997) *Silica, Some Silicates, Coal Dust and para-Aramid Fibrils.* World Health Organization International Agency for Research on Cancer, report no 68.

[33] Jones RH., Molitoris BA. (1984) A statistical method for determining the breakpoint of two lines. *Annals of Biochemistry*, **141**: 287-290

[34] Koziol J. (1991) On maximally selected chi-square statistics. *Biometrics*, **47**: 1557-1561

[35] Kuechenhoff H., Ulm K. (1997) Comparison of statistical methods for assessing threshold limiting values in occupational epidemiology. *Statistics in Medicine*, **12**: 249-264

[36] Lausen B., Scumacher M. (1992) Maximally selected rank statistics. *Biometrics*, **48**: 73-85

[37] Leuraud K., Benichou J. (2001) A comparison of several methods to test for the existence of a monotonic dose-response relationship in clinical and epidemiological studies. *Statistics in Medicine*, **20**: 3335-3351

[38] Maclure M., Greenland S. (1992) Tests for trend and dose response: misinterpretations and alternatives. *American Journal of Epidemiology*, **135**: 96-104

[39] Mancuso J., Ahn H., Chen J. (2001) Order-restricted dose-related trend tests. *Statistics in Medicine*, **20**: 2305-2318

[40] Marcus R., Reritz E., Gabriel R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**: 655-660

[41] McGullagh P., Nelder J.A. (1983) *Generalized linear models.* Chapman & Hall, London.

[42] McLaughlin K., Lipworth L., Marano E., Tarone R. (2001) A critical evaluation of the scientific basis of the MAK commission's new general threshold limit values for dust. *Int. Arch. Occ. Env. Health*, **74**: 303-314

[43] Miller R., Siegmund D. (1982) Maximally selected Chi square Statistics. *Biometrics*, **38**: 1011-1016

[44] Morton-Jones T., Diggle P., Parker L. (2000) Additive isotonic regression models in epidemiology. *Statistics in Medicine*, **19**: 849-860

[45] Nagelkerke N. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, **78**: 691-692

[46] O'Quigley J., Pessione F. (1989) Score tests for homogeneity of regression effect in the proportional hazards model. *Biometrics*, **45**: 135-144

[47] Pastor R., Guallar E. (1998) Use of two-segmented logistic regression to estimate change-points in epidemiologic studies. *Journal of American Statistical Association*, **148**: 631-642

[48] Peddada S., Prescott K., Conaway M. (2001) Tests for order restrictions in binary response. *Biometrics*, **57**: 1219-1227

[49] Robertson T., Wright FT. and Dykstra RL. (1988) *Order Restricted Statistical Inference.* Wiley, New York.

[50] Rom D., Costello R., Connell L. (1994) On Closed Test Procedures for Dose-Response Analysis. *Statistics in Medicine*, **13**: 1583-1596

[51] Royston P., Altman DG. (1994) Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling (with discussion). *Applied Statistics*, **43**: 429-467

[52] Salanti G., Ulm K. (2001) Modelling under order restrictions. *Discussion Paper No. 265* Sonderforschungsbereich 386 Ludwig-Maximilians-University, Munich.

[53] Salanti G., Ulm K. (2001) Multidimensional isotonic regression and estimation of the Threshold Value. *Discussion Paper No. 234* Sonderforschungsbereich 386 Ludwig-Maximilians-University, Munich.

[54] Sauerbrei W., Royston P. (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors using fractional polynomials. *Journal of the Royal Statistical Society*, Series A **162**: 71-94

[55] Schell MJ., Singh B. (1997) The reduced monotonic regression method. *Journal of American Statistical Association*, **92**: 128-135

[56] Silvapulle M. (1991) On the estimation of threshold values (Correspondence). *Biometrics*, **47**: 1629

[57] Tarone R. (1986) Correcting tests for trend in proportions for skewness. *Communications in Statistics: Theory and Methods*, **15(2)**: 317-328

[58] Tarone R., Gart J. (1980) On the robustness of combined tests for trends in proportions. *Journal of American Statistical Association*, **75**: 110-116

[59] Therneau T., Grambsch P. (2001) *Modelling survival data.* Springer, New York.

[60] Ulm K., Dannegger F., Becker U. (1998) Tests for trends in binary response. *Discussion paper No. 115* Sonderforschungsbereich 386 Ludwig-Maximilians-University, Munich.

[61] Ulm K. (1991) A statistical method for assessing a threshold in epidemiological studies. *Statistics in Medicine*, **10**: 341-349

[62] Ulm K. (1999) Nonparametric Analysis of Dose-Response Relationships. *Annals. NY Acad. of Sciences*, **895**: 223-231

[63] Ulm K., Salanti G. (2001) Estimation of the general threshold limit values for dust. *Int. Arch. Occ. Env. Health* **76**(3): 233-40

[64] Ulm K. (1989) On the estimation of threshold values (Correspondence). *Biometrics*, **45**: 1324-1326

[65] Westfall P. (1997) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**: 65-70.

[66] Worsley J. (1982) An improved Bonferroni inequality and applications. *Biometrika*, **69**: 297-302

[67] Xu R., Adak S. (2002) Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics*, **58**(2): 305-315

[68] Zhang H., Singer B. (1999) *Recursive partitioning in the health sciences.* Springer, New York.

List of Tables

1.1 The MAK study data (sample from Munich). 15

3.1 Notation used for the various test-statistics. 27

3.2 The coefficient for skewness γ if $K = 5$ and $n_i = 50$. 36

3.3 Simulations under H_0 (constant risk). Critical values for comparing $K = 5$ dose groups with $n_i = 50$ observations in each dose group, and response probability 10% and 4%. In first column the result from the theoretical distribution is depicted and in the second column the critical value estimated from 10 000 simulations. For the W iso-chi-squared test, the theoretical value corresponds to the approximation derived by Gautam. 37

3.4 Simulations under H_0 assumption (constant risk). Isotonic Likelihood Ratio (R) test: Critical values for comparing $K = 8$ dose groups and sample size 400 (50 in each dose group) when the response rate is 5%, 10% and 25%. The estimation has been accomplished using 10 000 simulations. 38

3.5 Simulations under H_1 assumption (increasing risk). Comparison of the power under various situations considered among the *CA* test, the Isotonic Likelihood Ratio test R, the *Wilcoxon* and the iso-chi-squared W test. Five groups with 50 observations per group are assumed. The critical values used for R and W are estimated through permutations. 41
3.6 Data from the para-aramid study (IARC 1997). Tumor: adenoma, bronchido-alveolar without keratinising squamous-cell carcinoma. . . 42
3.7 Analysing the data from the para-aramid study: results. 43
3.8 Notation used for survival adjustment in 2xK tables. 45
3.9 Data from the EMIAT study. The response is all causes death. 47
3.10 EMIAT: The influence of left ventricular ejection fraction in mortality. Results from isotonic regression and poly-3 test. 47

4.1 Estimation of ε^* based on 5000 simulations for different sample size and response rate. The overall significance level was 5%. 55
4.2 Deviancies and degrees of freedom of models applied in data set from Munich (smokers). 66

5.1 Extract of the original data matrix, example of no convergence (original data). 75
5.2 Extract of the isotonic matrix, example of no convergence (isotonic estimation). 75
5.3 The deviance of isotonic and reduced models. 83
5.4 Several criteria to compare isotonic-surface, reduced-surface and additive isotonic model. 88
5.5 The final classification tree in numbers. 89

6.1 Simulation's study for threshold value estimation. 103
6.2 Results from **method 1**. 115
6.3 Bootstrap confidence limits for the estimated threshold using **method 2**.116

7.1 Acute lymphoblastic leukemia study: The PH Cox model. The deviance is 957.08 with 5 degrees fo freedom. 130

7.2 Elimination of the time level sets for MSPS dynamic coefficient. . . . 133

A.1 Approximate critical values for significance level α=0.2, 0.1, 0.5, 0.01 for the iso-chi-squared statistic W. (derived by S. Gautam [21]) 151

List of Figures

1.1 Example from MAK study. Isotonic regression together with smoothing spline. 16

3.1 Theoretical (estimation from formula) and empirical (estimation from permutations) cumulative distribution for the Isotonic Likelihood Ratio test R (response probability 4%, $n_i = 50$ observations in each dose and $K = 5$ dose groups). 38

4.1 Example for closed testing. 57

4.2 Results from simulation study: comparing monotonic regression and reduced monotonic regression regarding the relative loss in the fit as function of the sample size. 63

4.3 Full isotonic and reduced regression using a sequence of Fisher's test for the sample from Munich (smokers). The 95% confidence bands corresponds to the reduced isotonic regression. 65

4.4 Reducing the isotonic model for Munich (smokers) using two different approaches. 67

5.1 The two-dimensional isotonic model (L=40 blocks). 84

5.2 The width of the confidence surfaces simulated under the isotonic estimates. 85

5.3 The two-dimensional reduced isotonic model (L=3 blocks). 86

5.4 The additive isotonic model (L=12 blocks). 87

6.1 Change in the deviance (equation 6.2) while pooling (slope=0.02, 5 dose groups, sample size 250). As Di-Dj is denoted the change in the deviance between the model pooling information until the j-th group and the model pooling information until the i-th group. 99

6.2 The first pooling (D2-D1) distribution assuming different slopes (5 dose groups, sample size 250). 100

6.3 Shapes for dose-response studied in simulation study. 102

6.4 **Method 1** for shape A corresponding to constant event rate - The type I error$_a$. 104

6.5 **Method 1** for shape B corresponding to linear increase - The power$_a$. 105

6.6 **Method 2** for shape C corresponding to a segmented line with threshold - The power$_a$. 106

6.7 Shape C: The probability to assess the correct threshold (power$_b$) using **method 1**. The true threshold location is in the second dose-group. 107

6.8 Shape C: The probability to assess the correct threshold (power$_b$) using **method 2**. The true threshold location is in the second dose-group. 108

6.9 Shape C: Comparing **method 1** regression and **method 2** regarding the probability to assess the correct threshold-power$_b$. The bars in each sample size panel correspond to slopes 0.02, 0.05, 0.10 and 0.15. 109

6.10 Shape C: Comparing **method 1** and **method 2** regarding the probability to assess no threshold when one exists (type I error$_b$). The bars in each sample size panel correspond to slopes 0.02, 0.05, 0.10 and 0.15. The horizontal line is drawn at the nominal level 5%. . . . 110
6.11 Shape C: Probabilities for **method 1** and **method 2** to assess a changepoint as threshold, when its true position is in the 2nd and 3rd level set (regression slope=0.1). 111
6.12 **Method 1** for shape D corresponding to a segmented line without threshold - The power$_a$. 112
6.13 Example for closed testing on threshold value estimation. 114

7.1 Simulations study for survival data. Compare in terms of power (first three figures) and type I error (last figure) the Grambsch and Therneau test (GT test), the fractional polynomials test and the isotonic version of GT test. 128
7.2 Kaplan-Meier cumulative survival curves for MSPS. 131
7.3 Smoothing the scaled Schoenfeld residuals for MSPS. 132
7.4 Isotonic fit for time-dependent coefficient for MSPS. 134

List of Algorithms

Algorithm 1: *The Pool Adjacent Violators Algorithm* 22

Algorithm 2: *Univariate permutation test* 33

Algorithm 3: *Backward Elimination's procedure* 53

Algorithm 4: *Closed elimination procedure* 59

Algorithm 5: *Compare reduced to isotonic model* 61

Algorithm 6: *Backfitting for additive isotonic models* 70

Algorithm 7: *The Iterative Algorithm for Partial Order* 73

Algorithm 8: *Conditional permutations for partial significance* 79

Algorithm 9: *Closed testing for threshold estimation* 113

Curiculum vitae

Personal data

Name:	Georgia Salanti
Born on:	16 December 1976
in:	Athens, Greece
Address:	IMSE, Klinikum Rechts der Isar, Ismaninger Strasse 22,
	81675 Munich - Germany
Telephone:	++49 (0) 89 41404355
e-mail:	georgia.salanti@imse.med.tu-muenchen.de

Qualifications

April 2003 — Ph.D in Statistics Dr. rer. nat.
Thesis : *"The isotonic regression framework: modelling and testing under order restrictions"* (in English)
Supervisors: Prof. Kurt Ulm, Prof. Ludwig Fahrmeir
Statistical department, Ludwig-Maximillians University of Munich.

March 2000 — D.E.S. (Hons) Diploma of Specialised Studies
"Statistical, epidemiological and operational methods in medicine and public health"
Thesis :*"Assessing prognostic factors in patients with coronal heart disease"* (in French)
School of Public Health, U.L.B University of Brussels.

June 1999 — B.Sc in Mathematics, option "Applied Mathematics"
Department of Mathematics, National University of Athens.

Education

July 2000-April 2003	Ph.D student in Ludwig-Maximillians University of Munich, Department of Statistics, Germany.
September 1999-March 2000	D.E.S postgraduate student at the University of Brussels, School of Public Health, Belgium.
September 1994-June 1999	Graduate student at the National University of Athens, Department of Mathematics, Greece.
(January 1998-June 1998	Elective student at the National University of Ireland, Galway, Department of Mathematics, Ireland.)
(October 1996-January 1997	Elective student at the Julius-Maximilians University Würzburg, Department of Mathematics, Germany.)

Professional experience

Since July 2000	• Research associate at the Institute for Medical Statistics and Epidemiology, University Hospital "Klinikum Rechts der Isar" Medical School of the Technical University of Munich. • Statistical consulting for basic research and clinical studies. • Teaching assistance - supervising of diploma theses for students of the Ludwig-Maximillians University of Munich.
Since July 2000	Project: "Threshold value estimation methods", German Research Foundation (DFG) Grant UL 94/11-1.
Since January	Project: " Non-parametric and semi-parametric estimation" DFG's Special Research Area SFB 386 "Statistical Analysis of discrete Structures" Grant by the Special Research Area (SFB).

Language skills

English	Fluent
German	Fluent
French	Fluent
Greek	Native

Computing skills

Statistical Software	S-Plus (advanced programming level), SPSS, SAS, EPI-INFO, Statgraphics, GAUSS
Programming languages	Turbo Pascal
Operating Systems	Win NT/2000/XP, Unix
Other	LaTeX, HTML

Membership

International Society of Clinical Biostatistics
Hellenic Statistical Institute